

**Introduction to IBM SPSS Text
Analytics for IBM SPSS Modeler (v16)
Student Guide
Course Code: 0A105**

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Introduction to IBM SPSS Text Analytics for IBM SPSS Modeler (v16)

0A105

ERC: 1.0

Published December 2014

All files and material for this course, 0A105 Introduction to IBM SPSS Text Analytics for IBM SPSS Modeler (v16), are IBM copyright property covered by the following copyright notice.

© Copyright IBM Corp. 2010, 2014

US Government Users Restricted Rights – Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM corp.

IBM, the IBM logo, ibm.com and SPSS are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at “Copyright and trademark information” at www.ibm.com/legal/copytrade.shtml.

Adobe and the Adobe logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Microsoft, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

P-2

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Contents

PREFACE.....	P-1
CONTENTS.....	P-3
COURSE OVERVIEW	P-11
DOCUMENT CONVENTIONS	P-13
WORKSHOPS	P-14
ADDITIONAL TRAINING RESOURCES.....	P-15
IBM PRODUCT HELP.....	P-16
INTRODUCTION TO TEXT MINING	1-1
OBJECTIVES	1-3
TEXT MINING AND DATA MINING	1-5
TEXT MINING APPLICATIONS	1-6
A STRATEGY FOR DATA MINING: CRISP-DM	1-8
IDENTIFYING THE STAGES IN CRISP-DM.....	1-10
STAGE 1: BUSINESS UNDERSTANDING	1-11
STAGE 2: DATA UNDERSTANDING	1-12
STAGE 3: DATA PREPARATION.....	1-13
STAGE 4: MODELING.....	1-16
STAGE 5: EVALUATION	1-18
STAGE 6: DEPLOYMENT	1-19
APPLY YOUR KNOWLEDGE.....	1-20
SUMMARY	1-23
WORKSHOP 1: MAKING PREPARATIONS FOR A TEXT MINING PROJECT	1-24
A TEXT MINING OVERVIEW.....	2-1
OBJECTIVES	2-3
TEXT MINING NODES.....	2-4
TEXT MINING MODELING NODE.....	2-6
INTERACTIVE WORKBENCH	2-7
EXTRACTED RESULTS PANE	2-8
DATA PANE.....	2-9
CATEGORIES PANE.....	2-10
VISUALIZATION PANE	2-11
RESOURCE EDITOR.....	2-12
TYPICAL TEXT MINING SESSION.....	2-14

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

P-3

DEMO 1: A TYPICAL TEXT MINING SESSION	2-16
APPLY YOUR KNOWLEDGE.....	2-31
SUMMARY	2-33
WORKSHOP 1: TEXT MINING CUSTOMER OPINIONS ABOUT PORTABLE MUSIC PLAYERS.....	2-34
READING TEXT DATA.....	3-1
OBJECTIVES	3-3
FILE LIST NODE	3-4
USING THE FILE LIST NODE IN TEXT MINING.....	3-5
DEMO 1: USING THE FILE LIST NODE TO READ TEXT FROM MULTIPLE FILES	3-7
FILE VIEWER NODE.....	3-13
DEMO 2: USING THE FILE VIEWER NODE TO VIEW DOCUMENTS IN MODELER.....	3-14
WEB FEED NODE.....	3-18
WEB FEED NODE - RSS FORMAT	3-19
WEB FEED NODE - HTML FORMAT	3-20
DEMO 3: READING TEXT FROM WEB FEEDS	3-21
APPLY YOUR KNOWLEDGE.....	3-27
SUMMARY	3-29
WORKSHOP 1: TEXT MINING DATA FROM AN RSS FEED	3-30
LINGUISTIC ANALYSIS AND TEXT MINING	4-1
OBJECTIVES	4-3
LINGUISTIC ANALYSIS	4-4
PARTS OF SPEECH	4-6
EXTRACTOR COMPONENT WORKFLOW	4-9
TEXT PREPROCESSING	4-10
IDENTIFICATION OF CANDIDATE TERMS.....	4-11
IDENTIFICATION OF EQUIVALENCE CLASSES.....	4-13
FORCING / EXCLUDING	4-15
ASSIGNING OF TYPES	4-17
CATEGORIZING EXTRACTED CONCEPTS	4-18
USING TEMPLATES AND LIBRARIES	4-20

TEXT ANALYSIS PACKAGES.....	4-22
LINGUISTIC RESOURCE RELATIONSHIPS	4-23
APPLY YOUR KNOWLEDGE.....	4-24
SUMMARY	4-27
CREATING A TEXT MINING CONCEPT MODEL	5-1
OBJECTIVES	5-3
TEXT MINING CONCEPT MODEL.....	5-4
CREATING A CONCEPT MODEL	5-5
SPECIFYING MODEL OPTIONS	5-6
SELECTING A RESOURCE TEMPLATE / TAP	5-7
SELECTING EXPERT TAB OPTIONS	5-8
TEXT MINING NUGGET: CONCEPT MODEL	5-11
UNDERLYING TERMS IN CONCEPT MODELS	5-13
DEMO 1: CREATING A TEXT MINING CONCEPT MODEL	5-14
COMPARING DIFFERENT TEMPLATES	5-19
SELECTING CONCEPTS FOR SCORING.....	5-21
DEMO 2: SELECTING CONCEPTS FOR SCORING	5-22
SCORING MODEL DATA	5-25
SPECIFYING THE SCORING MODE	5-26
DEMO 3: SCORING THE DATA	5-28
RELATING CONCEPTS TO OTHER DATA	5-33
DEMO 4: EXAMINING THE RELATIONSHIP BETWEEN CONCEPTS AND OTHER CUSTOMER DATA	5-34
APPLY YOUR KNOWLEDGE.....	5-41
SUMMARY	5-44
WORKSHOP 1: CREATING A TEXT MINING CONCEPT MODEL.....	5-45
REVIEWING TYPES AND CONCEPTS IN THE INTERACTIVE WORKBENCH	6-1
OBJECTIVES	6-3
INTERACTIVE WORKBENCH VIEWS	6-4
CATEGORIES AND CONCEPTS VIEW	6-7
REVIEWING EXTRACTED CONCEPTS	6-8
FILTER CONCEPTS DIALOG	6-9
UPDATING THE TEXT MINING NODE.....	6-10

DEMO 1: FILTERING EXTRACTED RESULTS.....	6-12
REVIEWING EXTRACTED TYPES.....	6-21
DEMO 2: REVIEWING EXTRACTED TYPES	6-23
USING AN UPDATED MODELING NODE.....	6-29
DEMO 3: USING AN UPDATED MODELING NODE	6-30
APPLY YOUR KNOWLEDGE.....	6-34
SUMMARY	6-36
WORKSHOP 1: REVIEWING EXTRACTED RESULTS IN THE INTERACTIVE WORKBENCH.....	6-37
EDITING LINGUISTIC RESOURCES	7-1
OBJECTIVES	7-3
USING RESOURCE TEMPLATES.....	7-4
USING LIBRARIES.....	7-6
USING LIBRARY DICTIONARIES	7-8
USING TYPE DICTIONARIES	7-10
USING SUBSTITUTION DICTIONARIES	7-13
USING EXCLUSION DICTIONARIES	7-16
MODIFYING THE DICTIONARIES.....	7-17
EXTRACTING UNEXTRACTED TEXT	7-19
PREPARING FOR LINGUISTIC EDITING.....	7-20
SAMPLING TEXT DATA	7-22
SETTING TEXT MINING GOALS	7-23
TYPES VERSUS SYNONYMS	7-25
DEMO 1: MODIFYING THE DICTIONARIES	7-27
APPLY YOUR KNOWLEDGE.....	7-43
SUMMARY	7-45
WORKSHOP 1: EDITING DICTIONARIES	7-46
FINE TUNING RESOURCES.....	8-1
OBJECTIVES	8-3
FUZZY GROUPING	8-4
NON-LINGUISTIC ENTITIES.....	8-5
NORMALIZING NON-LINGUISTIC ENTITIES	8-8
CONFIGURATION	8-9
LANGUAGE HANDLING	8-11

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

EXTRACTION PATTERNS	8-12
FORCED DEFINITIONS	8-15
ABBREVIATIONS	8-17
DEMO 1: EDITING ADVANCED RESOURCES	8-18
APPLY YOUR KNOWLEDGE.....	8-25
SUMMARY	8-27
WORKSHOP 1: EDITING ADVANCED RESOURCES.....	8-28
PERFORMING TEXT LINK ANALYSIS.....	9-1
OBJECTIVES	9-3
TEXT LINK ANALYSIS PANES.....	9-5
TYPE PATTERNS PANE	9-6
CONCEPT PATTERNS PANE.....	9-7
VISUALIZATION PANE.....	9-8
TLA GRAPH LAYOUTS.....	9-9
DATA PANE.....	9-10
DEMO 1: EXPLORING TEXT LINK PATTERNS IN THE INTERACTIVE WORKBENCH.....	9-11
USING THE TEXT LINK RULES EDITOR	9-20
SETTING UP TEXT LINK RULE VALUES	9-21
RULE OUTPUT TABLE	9-23
WHEN TO CREATE OR EDIT RULES	9-25
DEMO 2: CREATING TEXT LINK RULES.....	9-27
CONVERTING TLA PATTERNS TO CATEGORIES	9-37
DEMO 3: CONVERTING TLA PATTERNS INTO CATEGORIES	9-38
TEXT LINK ANALYSIS NODE.....	9-44
TEXT LINK ANALYSIS NODE OUTPUT.....	9-45
TEXT LINK ANALYSIS NODE - TIPS.....	9-47
DEMO 4: USING THE TEXT LINK ANALYSIS NODE.....	9-48
APPLY YOUR KNOWLEDGE.....	9-58
SUMMARY	9-60
WORKSHOP 1: TEXT LINK ANALYSIS.....	9-61

CLUSTERING CONCEPTS.....	10-1
OBJECTIVES	10-3
CLUSTERS OF CONCEPTS.....	10-4
BUILDING CLUSTERS	10-5
CLUSTERS VIEW.....	10-6
CLUSTER ANALYSIS SETTINGS	10-7
SETTING THE CLUSTER LINK VALUE	10-9
INITIAL CLUSTERING RESULTS	10-11
CLUSTER WEB GRAPHS	10-13
CLUSTER CONCEPTS INTO CATEGORIES	10-15
DEMO 1: CLUSTERING CONCEPTS	10-16
APPLY YOUR KNOWLEDGE.....	10-26
SUMMARY	10-28
WORKSHOP 1: CLUSTERING CONCEPTS	10-29
CATEGORIZATION TECHNIQUES.....	11-1
OBJECTIVES	11-3
STRATEGIES FOR CREATING CATEGORIES	11-5
TEXT ANALYSIS PACKAGE (TAP).....	11-6
DEMO 1: USING A TEXT ANALYSIS PACKAGE TO CATEGORIZE DATA	11-7
IMPORTING PREDEFINED CATEGORIES	11-11
DEMO 2: IMPORTING PREDEFINED CATEGORIES FROM A MICROSOFT EXCEL FILE	11-14
AUTOMATED CLASSIFICATION.....	11-19
AUTOMATED CLASSIFICATION METHODS	11-20
LINGUISTIC CATEGORIZATION TECHNIQUES	11-22
ADDITIONAL CATEGORIZATION OPTIONS.....	11-25
DEMO 3: AUTOMATED CLASSIFICATION	11-28
SUMMARY	11-40
WORKSHOP 1: IMPORTING PREDEFINED CATEGORIES	11-41
WORKSHOP 2: USING AUTOCLASSIFICATION TECHNIQUES TO CATEGORIZE DATA	11-45

CREATING CATEGORIES.....	12-1
OBJECTIVES	12-3
WHICH TECHNIQUE SHOULD YOU USE?.....	12-4
USING AUTOMATED CLASSIFICATION	12-5
DEMO 1: CATEGORIZING ASTROSERVE CALL CENTER DATA	12-7
EXTENDING CATEGORIES.....	12-23
CREATING RULES.....	12-25
DEMO 2: FINE TUNING CATEGORIES.....	12-27
CREATING A FINAL SET OF CATEGORIES	12-32
CREATING A TEXT ANALYSIS PACKAGE.....	12-33
DEMO 3: CREATING A FINAL SET OF CATEGORIES AND TEXT ANALYSIS PACKAGE	12-35
APPLY YOUR KNOWLEDGE.....	12-41
SUMMARY	12-43
WORKSHOP 1: CREATING CATEGORIES	12-44
MANAGING LINGUISTIC RESOURCES	13-1
OBJECTIVES	13-3
TEMPLATE EDITOR.....	13-5
CREATING A RESOURCE TEMPLATE.....	13-6
LOCAL AND PUBLIC LIBRARIES	13-7
PUBLISHING LIBRARIES	13-9
EDITING FORCED TERMS	13-10
PUBLISHING LIBRARIES	13-11
SHARING LIBRARIES	13-13
MANAGING RESOURCE TEMPLATES	13-14
BACKING UP RESOURCES	13-15
DEMO 1: MANAGING LINGUISTIC RESOURCES	13-16
APPLY YOUR KNOWLEDGE.....	13-33
SUMMARY	13-35
WORKSHOP 1: MANAGING LINGUISTIC RESOURCES.....	13-36

USING TEXT MINING MODELS	14-1
OBJECTIVES	14-3
EXPLORING A TEXT MINING MODEL	14-4
DEVELOPING A MODEL BY COMBINING CATEGORIES AND CUSTOMER DATA.....	14-9
SCORING NEW DATA	14-19
APPLY YOUR KNOWLEDGE.....	14-25
SUMMARY	14-27
WORKSHOP 1: USING TEXT MINING MODELS	14-28
THE PROCESS OF TEXT MINING.....	A-1
OBJECTIVES	A-3
SUMMARY	A-8

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

P-10

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Course Overview

Course Overview

Introduction to IBM SPSS Text Analytics for IBM SPSS Modeler (v16) teaches users how to analyze text data using IBM SPSS Modeler Text Analytics. Students will see the complete set of steps involved in working with text data, from reading the text data to creating the final categories for additional analysis. After the final model has been created, there is an example of how to apply the model to perform Churn analysis. Topics include how to automatically and manually create and modify categories, how to edit synonym, type, and exclude dictionaries, and how to perform Text Link Analysis and Cluster Analysis with text data. Also included are examples of how to create resource templates and Text Analysis packages to share work with other projects and other users.

Important Note: These course materials were created using IBM SPSS Modeler Text Analytics v16.0. You may not be able to exactly replicate the results you see in the screenshots contained in the manual if you are using a different version of the software. This is because the extraction engine is modified with each release.

Recommended Duration: 2 days

Recommended Modality: Instructor led

Intended Audience

IBM SPSS Modeler Analysts who want to become familiar with performing Text Analytics in IBM SPSS Modeler.

Specifically, this is an introductory course for:

- anyone who is new to IBM SPSS Modeler
- anyone considering purchasing IBM SPSS Modeler
- anyone interested in Data Mining
- anyone interested in Text Mining

Topics Covered

- Topics covered in this course include:
- Introduction to Text Mining
- An Overview of Text Mining and IBM SPSS Modeler
- Reading Text Data
- Linguistic Analysis and Text Mining
- Creating a Text Mining Concept Model
- Reviewing Types and Concepts in the Interactive Workbench
- Editing Linguistic Resources
- Fine Tuning Resources
- Performing Text Link Analysis
- Clustering Concepts
- Categorization Techniques
- Creating Categories
- Managing Linguistic Resources
- Using Text Mining Models
- The Process of Text Mining

Course Prerequisites

Participants should have:

- General computer literacy

Document Conventions

Conventions used in this guide follow Microsoft Windows application standards, where applicable. As well, the following conventions are observed:

Bold

Bold style is used in demo and workshop step-by-step solutions to indicate either:

- actionable items

(Point to **Sort**, and then click **Ascending.**)

- text to type or keys to press

(Type **Sales Report**, and then press **Enter.**)

- UI elements that are the focus of attention

(In the **Format** pane, click **Data**)

Italic

Used to reference book titles.

CAPITALIZATION

All file names, table names, column names, and folder names appear in this guide exactly as they appear in the application.

To keep capitalization consistent with this guide, type text exactly as shown.

Workshops

Workshop Format

Workshops are designed to allow you to work according to your own pace. Content contained in a workshop is not fully scripted out to provide an additional challenge. Refer back to demonstrations if you need assistance with a particular task. The workshops are structured as follows:

The Business Question Section

This section presents a business-type question followed by a series of tasks. These tasks provide additional information to help guide you through the workshop. Within each task, there may be numbered questions relating to the task. Complete the tasks by using the skills you learned in the module. If you need more assistance, you can refer to the Task and Results section for more detailed instruction.

The Task and Results Section

This section provides a task based set of instructions that presents the question as a series of numbered tasks to be accomplished. The information in the tasks expands on the business case, providing more details on how to accomplish a task. Screen captures are also provided at the end of some tasks and at the end of the workshop to show the expected results.

Additional Training Resources

Bookmark [Business Analytics Product Training](#)

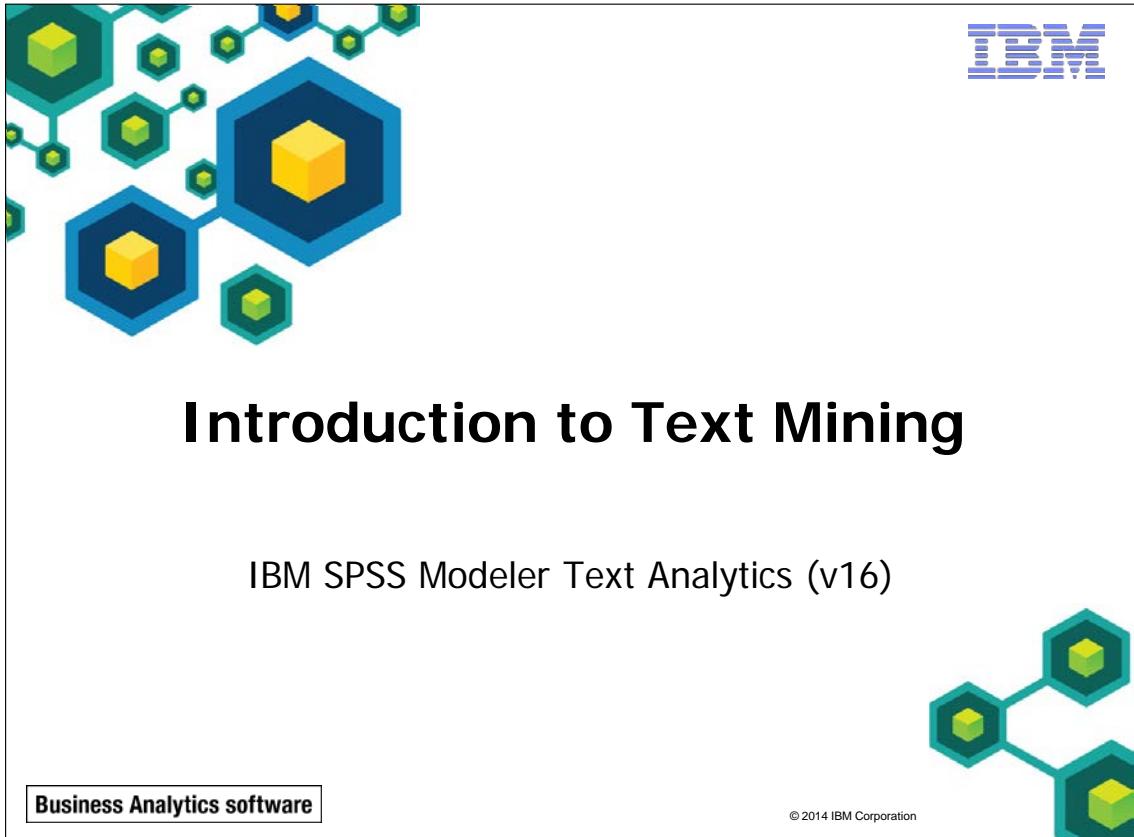
<http://www-01.ibm.com/software/analytics/training-and-certification/> for details on:

- instructor-led training in a classroom or online
- self-paced training that fits your needs and schedule
- comprehensive curricula and training paths that help you identify the courses that are right for you
- IBM Business Analytics Certification program
- other resources that will enhance your success with IBM Business Analytics Software

IBM Product Help

Help type	When to use	Location
Task-oriented	You are working in the product and you need specific task-oriented help.	<i>IBM Product - Help link</i>
Books for Printing (.pdf)	<p>You want to use search engines to find information. You can then print out selected pages, a section, or the whole book.</p> <p>Use Step-by-Step online books (.pdf) if you want to know how to complete a task but prefer to read about it in a book.</p> <p>The Step-by-Step online books contain the same information as the online help, but the method of presentation is different.</p>	Start/Programs/ <i>IBM Product/Documentation</i>
IBM on the Web	<p>You want to access any of the following:</p> <ul style="list-style-type: none"> • Training and Certification Web site • Online support • IBM Web site 	<ul style="list-style-type: none"> • http://www-01.ibm.com/software/analytics/training-and-certification/ • http://www-947.ibm.com/support/entry/portal/Overview/Software • http://www.ibm.com

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Objectives

- At the end of this module, you should be able to:
 - describe text mining and its relationship to data mining
 - explain CRISP-DM methodology as it applies to text mining
 - describe the steps in a text-mining project

© 2014 IBM Corporation

An organization's information database includes an increasing amount of data in unstructured and semi-structured formats from customer e-mails, call center logs, incident descriptions, suspicious transaction reports, customer open-ended survey responses, news feeds, Web forms, and so on. Using this information in a thorough and systematic way is increasingly necessary to understand customer behavior and attitudes and develop successful predictive models.

There are, though, no standard rules for writing text so that a computer can understand it. Language and meaning vary for every piece of text depending on the purpose, the recording medium and who is actually recording it. The only way to accurately include unstructured data in a data-mining project is to understand the language and the context within which the text was created.

Traditional approaches to extracting concepts from unstructured information are non-linguistic including, keyword or Boolean searches, inverted indexes, statistical or probabilistic algorithms, concept agents, neural networks, and pattern recognition. These methods are based on the comparison of character strings in both queries and text and therefore do not provide an understanding of concepts. For example, in the text string "reproduction of documents" the term "reproduction" should be associated with synonyms such as "copy" or "duplication". Otherwise equivalent phrases such as "copying of documents" will be overlooked. A non-linguistic-based system attempting to perform this kind of synonymy will likely include the synonyms procreation and propagation, which would retrieve irrelevant information. Text context is crucial.

Understanding human language is based on linguistics, commonly referred to as Natural Language Processing (NLP). A system that incorporates NLP can intelligently extract terms, including compound phrases, and also permit classification of terms into related groups, such as products, organizations, or people. Linguistic systems are knowledge sensitive—the more information contained in the linguistic resources (dictionaries), the higher the quality of results. Modification of the dictionary content, such as synonym definitions, can simplify the resulting information and focus attention on the most relevant concepts.

For example, consider the sentence:

Innovative solutions from SPSS(R) Inc. enable your organization to both uncover concepts hidden in text and use them to predict future conditions, behavior, and trends.

An IBM SPSS Modeler Text Mining node extracts the concepts "innovative solutions", "spss inc", "organization", "concepts", "text", "future conditions", "behavior", and "trends" from this sentence. These concepts describe the basic meaning of the sentence, as follows:

Innovative solutions from SPSS Inc. enable your organization to both uncover concepts hidden in text and use them to predict future conditions, behavior, and trends.

Text Mining and Data Mining

- Data Mining algorithms require structured data as input.
- Before data mining tools and algorithms can be used to find patterns or create models from text data, the information must be structured.
- Text Mining is used to structure the text data into categories that it can be used as input in data mining.

© 2014 IBM Corporation



Text mining is the process of extracting knowledge and information from natural language texts. Text mining proceeds in two stages.

- Stage 1: Key concepts/terms are extracted from the text that represents the essence of information the text contains.
- Stage 2: These concepts/terms are grouped into categories that represent the higher-level ideas contained in the text.

As a simple example of extraction and then categorization, in Stage 1 the terms "disk drive," "CPU," and "CD-ROM" could be extracted from text. In Stage 2, these three terms would be automatically grouped into a category that might be labeled "computer hardware".

Data-mining models can then answer such questions as:

- What behaviors do the text categories predict?
- How do the text categories combine with other information to improve predictability?
- Which text categories are associated with specific segments of customers?

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Text Mining Applications

- Text mining applications can be classified into two broad classes:
 - text understanding / summarization
 - modeling with text

© 2014 IBM Corporation



In general, any organization that routinely needs to review large volumes of documents to identify key elements for further exploration can benefit from IBM SPSS Modeler Text Analytics. Text mining applications can be classified into two broad classes:

Text understanding/summarization: The large volumes of text that are collected today are too extensive to be analyzed by humans. This is especially true in such areas as call center data or with databases of scientific articles. One goal of text mining can be to extract meaningful information from text so that users can understand the key concepts contained therein. Thus, you might want to know the percentage of customers that complain about slow repairs, or delayed orders, and how that varies over time and by product or type of customer.

Modeling with text: More commonly, text mining can be one phase of developing a model to predict customer behavior, for example, churning or product purchase. The concepts or categories extracted from the text are used as input or target fields, along with other information, to develop a predictive model. Extracted concepts or categories can also be used in cluster and association models.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Of course, both of these can be the focus of a text-mining project. In fact, the extraction of key concepts from text for incorporation in a model provides you, by definition, with textual understanding. However, the emphasis when constructing a model is often not the text itself, but instead how the text can improve a model's performance. Some specific applications of text mining include:

- CRM (Customer Relationship Management): Information is extracted from text data from all customer touch points, such as e-mails, transactions, call center communications, and surveys. This information is then used to predict customer churn and build up-sell and cross-sell models.
- Blog and Web mining: Text mining can extract information from free-form comments on websites and blogs about various topics of interest and provide insight into trends and themes in such commentary.
- Fraud detection: Potential fraud is discovered by looking for patterns and anomalies in large amounts of text data in healthcare, insurance, and government.
- Scientific and medical research: Extraction from text materials such as patent reports, journal articles, published research results, and other publications. Models then seek to identify associations that were previously unknown (such as a symptom associated with a particular drug), presenting avenues for further exploration. Text mining can minimize the time spent in the drug discovery process, and it can be an aid in genomics research.
- Security/Intelligence: Review large volumes of text to look for patterns and links between organizations and individuals to anticipate and deter terrorist threats and criminal behavior.
- Investment research: Systematic review of daily analyst reports, news articles, and company press releases to identify key strategy points or market shifts. Trend analysis of such information reveals emerging issues or opportunities for a firm or industry over time.
- Market research: Search and monitor published documents, press releases, and Web sites for measures of market penetration including competitive analysis. Text mining allows for the application of quantitative analytical methods to qualitative data found in open-ended survey questions, focus groups, and interviews.

A Strategy for Data Mining: CRISP-DM

- A data-mining project can become complicated quickly.
- A model is needed that guides you through the critical issues.
- Recommendation: use the Cross-Industry Standard Process for Data Mining (CRISP-DM)

© 2014 IBM Corporation

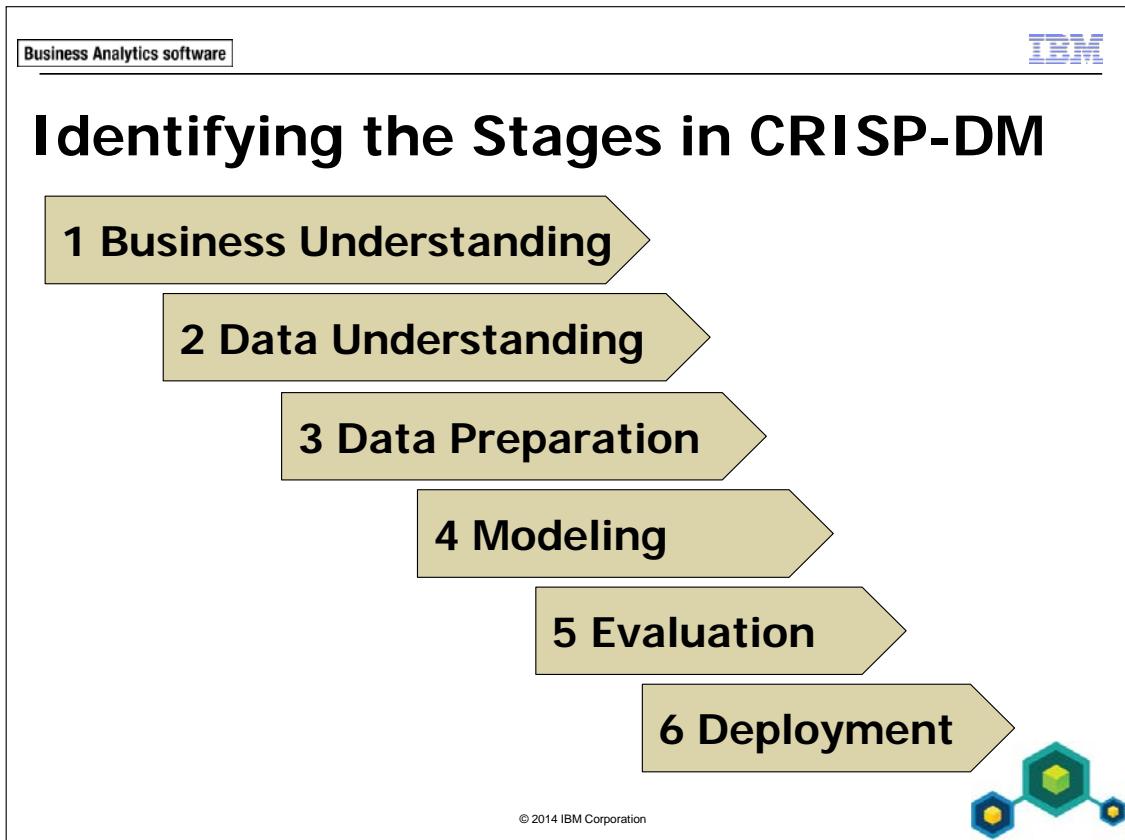


Data mining is much more effective if done in a planned, systematic way. Even with powerful data-mining tools such as Modeler, the majority of the work in data mining requires the careful eye of a knowledgeable business analyst to keep the process on track. This is just as true when incorporating text mining into a data-mining project. To guide your planning, it is important to answer the following questions when beginning a project:

- What substantive problem do you want to solve?
- What data sources are available, and what portions of the data are relevant to the current problem?
- What kind of preprocessing and data cleaning do you need to do before you start mining the data?
- What data-mining technique(s) will you use?
- How will you evaluate the results of the text mining/data mining analysis?
- How will you get the most out of the information you obtained from text mining/data mining?

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

To stay on track, it helps to have an explicitly defined process model for data mining. The process model guides users through the critical issues outlined above and makes sure that the important points are addressed. It serves as a data-mining road map so that users will not lose their way as you dig into the complexities of the data.



The data-mining process model recommended for use with Modeler is the Cross-Industry Standard Process for Data Mining (CRISP-DM). As you can tell from the name, this model is designed as a general model that can be applied to a wide variety of industries and business problems. For additional details on CRISP-DM, refer to Help\CRISP-DM Help in Modeler's main help menu.

The general CRISP-DM process model includes six phases that address the main issues in data mining. The six phases fit together in a cyclical process. These six phases cover the full data-mining process, including how to incorporate data mining into your larger business practices. These phases are listed in the diagram. The diagram illustrates the iterative nature of a data-mining project.

Business Analytics software

IBM

Stage 1: Business Understanding

Task	Sub task 1	Sub task 2	Sub task 3
Determine business objectives	Background	Business objectives	Business success criteria
Assess situation	Inventory of resources	Risks and contingencies	Terminology
Determine data-mining objectives	Data-mining success criteria		
Produce project plan	Write a project plan	Initial assessment of tools and techniques	

© 2014 IBM Corporation



Business understanding is perhaps the most important phase of data mining. It includes determining business objectives, assessing the situation, determining data-mining goals, and producing a project plan.

For text mining, business understanding includes determining what text data are available, and crucially, thinking about what information can be gained from the text data that is not in the structured data. Although this second question cannot be answered fully until the text mining is completed, it must be considered early in the project because the effort to mine text data can be extensive.

Thus, if you have already developed successful predictive models using only structured data, the addition of text mining may or may not improve performance. Conversely, if these models are not attaining necessary levels of accuracy, that is a good motivation and reason for including text data as another set of model inputs. If you have not developed any models yet for a particular outcome, then using text data along with structured data is usually a good strategy from the beginning.

Stage 2: Data Understanding

Task	Sub task 1
Collect initial data	Initial data-collection report
Describe data	Data-description report
Explore data	Data-exploration report
Verify data quality	Data-quality report

© 2014 IBM Corporation



Data provides the "raw materials" of data mining. The data understanding phase addresses the need to understand what your data resources are and the characteristics of those resources. It includes collecting initial data, describing data, exploring data, and verifying data quality.

For text mining, detailed understanding may be needed because of a specialized vocabulary used by customers or those who created the documents you are mining. The creation of appropriate dictionary resources (see next section) to handle specialized terms may be a lengthy effort, and users may need to draw on the expertise of those with relevant business and specialized knowledge.

At this stage you should also be thinking about what concepts and categories of text, and patterns of these concepts, you expect to discover in the text. This is especially important for projects whose goal is text understanding and summarization. The dictionary resources can be adjusted to search for expected concepts, and group together terms that you want to place under one category. The more that can be planned in advance, the more efficient the project will proceed.

Stage 3: Data Preparation

Task	Sub task 1	Sub task 2
Select data	Rational for inclusion and exclusion	
Clean data	Data-cleaning report	
Construct data	Derived attributes	
Format data and combine datasets	Set the unit of analysis	Integrate data

© 2014 IBM Corporation



After cataloging the data resources, you will need to prepare the data for data mining. The data preparation phase includes selecting, cleaning, constructing, integrating, and formatting data. These tasks will likely be performed multiple times, and not in any prescribed order. These tasks can be very time consuming, but are critical for the success of the data-mining project. For example, all text mining projects are adversely affected by problems of missing text data and errors in the text data. Consider each of these problems in turn.

1. Missing text data: Some records will not have text entries, although this will be more common when analyzing open-ended survey data where some customers do not provide answers to every question. For call center data there is text available for every case by definition, that is, a record was created only when there was something to record from a phone call or e-mail.

2. Errors in text data: Errors are usually more numerous in some types of text data. Call center data, survey data, or anything created directly by humans with minimal editing will include both spelling and grammatical errors. This type of text data will also commonly include abbreviations or different formats for the same information, such as dates. Modeler's Text Mining modeling node in Modeler provides you with various extraction methods that can fix spelling, or at least overcome variants in spelling, and also correct standard grammatical problems. Thus, there is no need to correct text before it is used in text mining. This can save quite a bit of time compared to using the same amount of structured data.

Missing text data is never estimated or imputed, as is commonly done with numeric structured data. This makes data preparation simpler, but it also implies that when the extracted categories are added to the structured data, there can be missing data for those customers who did not have a text response. Often, models are developed, such as decision trees, which incorporate the missing data directly into the model, that is, as a separate category. Using the Derive Node, blanks can also be created as flag fields to represent missing text as a category. Users will need to think about whether that is appropriate for the particular application.

Errors in the data may require the creation and editing of various text dictionaries to successfully extract the text and create meaningful categories. Dictionary resources include synonyms, words to be excluded from extraction, types that group together multiple terms, and other more specialized tuning. These tasks have no counterpart in structured data mining, and you can expect to spend a significant amount of time in editing dictionary resources. The higher the accuracy you desire, the more editing that will be necessary.

As exemplified by the iterative nature of the CRISP-DM process model, you may modify dictionary resources, develop a text-mining model, and then decide that further modifications are required based on the model results, thus moving back and forth between modeling and data preparation.

Another step in data preparation may be language translation from several languages into a common language to be used for text mining. Language translation is available in Modeler with the Translate node. Depending on the level of accuracy you desire for translation, you may spend some time reviewing and adjusting the translation process. If you have text data in common European languages, including French, German, and Spanish, Modeler can work directly in that language for text extraction. This allows users to create text-mining models in the native language.

Note: You must be able to connect to SDL's Software as a Service (SaaS) to be able to use the Translate node.

Stage 4: Modeling

Task	Sub task 1	Sub task 2
Select modeling techniques	Modeling assumptions	
Generate test design	Test design	
Build model	Set model parameters	Model descriptions
Assess model	Model assessment	Revise model parameters

© 2014 IBM Corporation



This phase involves selecting modeling techniques, generating test model designs, and building and assessing models. Similar to building standard statistical models, developing a model is an iterative process and you should expect to try several models and modeling techniques, before finding a best model.

The modeling phase when using text has two distinct stages, although they blend together in any real-life project.

- Firstly, you develop a text-mining model, whose output will be a set of categories that can be used to classify or score the text for each record. The number of categories constructed from the unstructured data could be very large because the amount of text is usually large and varied. Therefore, you may need to reduce the number of categories to a smaller set of key categories, perhaps to a few key sentiments. Depending on the modeling technique(s) you choose, this may be more or less critical. For example, decision trees can handle any number of inputs, but neural networks perform better with a more limited number of predictors, as is also true for the classical statistics-based techniques of regression or discriminant analysis.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

- Secondly, adding the categories to the structured data and developing a model per the usual procedures. The model must be validated on data not used to create the model.

If a model is not performing acceptably, you can also return to the text-mining model to modify the dictionary resources and/or change how text is categorized. Then results from the modified text-mining model can be input into the data-mining model to see whether performance improves.

Technically, you can also validate a text mining model itself, but if you use large enough data files for modeling, and the text data are diverse, experience suggests that data validation is much less critical for text mining itself. The objective of a text-mining model is to summarize the data, and that can be done quite successfully without validation. Additionally, it is difficult to develop validation criteria for text mining since you are not predicting anything, just categorizing text. In fact, if there are new concepts in the validation text, they may not be extracted, but users will not have any method of easily detecting them.

Stage 5: Evaluation

Task	Sub task 1	Sub task 2
Evaluate results	Assessment of data-mining results with respect to business success criteria	Approve models
Review process	Review of process	
Determine next steps	List of possible actions	Decision

© 2014 IBM Corporation



Once you have chosen the models from a technical perspective, you need to evaluate them. Since text data is included, one component of evaluation can be to compare results with and without text data (this would require building a second model without text data). This will show why text mining makes a difference in modeling and what the crucial information is from the text information.

Another aspect of evaluation is deciding whether a model meets the business criteria for success that were defined in the Business Understanding and Data Understanding phases of CRISP-DM. A technically acceptable model may not be useful within the business environment.

As with data-mining models in general, models that include text data provide opportunities for process improvement within the business. A model using call center data may show that dissatisfaction with the response time of service reps leads to more customer turnover. A model predicting likely customer churn that includes dissatisfaction with response time will be useful, but the relationship itself between response time and churn can be even more important. Such a finding provides direction to the organization on where to focus quality improvement efforts that are likely to provide a significant payoff. Models should always be examined with process improvement in mind.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Stage 6: Deployment

Task	Sub task 1	Sub task 2
Plan deployment	Deployment plan	
Maintenance	Maintenance plan	
Produce final report	Final report	Final presentation
Review project	Documentation	

© 2014 IBM Corporation



Most critical is deployment of the model to make predictions or create scores against new data. This might be relatively simple if done within Modeler, or more complex if the model is to be applied directly against an existing database. At the earliest phases of the CRISP-DM process, the availability of text data used in modeling should have been established. If the developed model is to be used for real-time scoring, then the text data must be similarly available in real time.

A plan should be developed to monitor the model's predictions and success in order to verify that the model still holds true over time. This might comprise automated analyses, which produce warnings if certain events occur, for example when the gap between the predicted and observed values exceeds a specified amount. When a model begins to underperform, there can be a variety of causes, but users will now need to also investigate whether the text data or structured data are the more likely source of reduced accuracy.

Apply Your Knowledge

Purpose:

Test your knowledge of the material covered in this module.

Question 1: True or False: The most import phase of data mining is business understanding.

- A. True
- B. False

Question 2: Fill in the blank. There are ____ stages in the CRISP-DM process model.

Question 3: True or False: When analyzing text, categories are the end result.

- A. True
- B. False

Question 4: True or False: Data Understanding is the first phase of the CRISP-DM process.

- A. True
- B. False

Question 5: True or False: Data Preparation is the most time consuming phase of the CRISP-DM process.

- A. True
- B. False

Question 6: True or False: Text mining proceeds in two stages: extraction and categorization.

- A. True
- B. False

- Question 7: True or False: Text Mining is purely knowledge discovery. It is unnecessary to be a subject matter expert to successfully perform Text Mining.
- A. True
 - B. False
- Question 8: True or False: Text mining in IBM SPSS Text Analytics for Modeler does not require the creation and editing of various text dictionaries to successfully extract the text and create meaningful categories. The resources that ship with the software should be sufficient for most text mining projects.
- A. True
 - B. False
- Question 9: True or False: Data mining techniques can successfully be used to find patterns and create models from both unstructured and structured text data.
- A. True
 - B. False
- Question 10: True or False: Data preparation for text mining is less time consuming than data preparation for structured data.
- A. True
 - B. False

Apply Your Knowledge - Solutions

- Answer 1: B. False. Business understanding is the most important part of a text mining project. Text mining projects are likely to fail unless the analyst has domain knowledge about the subject matter.
- Answer 2: There are 6 stages in the CRISP-DM methodology.
- Answer 3: A. True. Categories are usually preferred over concepts because they capture key themes in the data as opposed to concepts which are far more specific
- Answer 4: B. False. Business Understanding is the first phase
- Answer 5: A. True. Usually data preparation can take between 50 to 80% of the time in a data mining project
- Answer 6: A. True. The extraction process captures the concepts. Usually, though not always, most analysts prefer to combine these together into themes or categories. For example, the overall theme would be Pricing if customers complained about the cost of items. By reducing the amount of detail, analysts are better able to key themes in what customers are saying as opposed to specifics. On the other hand, you are losing detail. Pricing may only be a concern for a few items, not all of them.
- Answer 7: B. False. While there is certainly a knowledge discovery component to every text mining project, it is usually helpful to have some advance knowledge of what concepts and categories you expect to find in the text.
- Answer 8: B. False. In most cases, in order to successfully text mine the data, it will be necessary to build a customized library that is specific to the data. For example, if specialized vocabulary is used by customers or those who created the documents, the creation of appropriate dictionary resources to handle specialized terms will be necessary.
- Answer 9: B. False. In order to use data mining algorithms to find patterns and create models for text data, this type of data must be transformed into structured data first
- Answer 10: A. True. Far more data is contained in a single text field than in most structured fields. A single document or response can contain multiple misspellings or grammatical errors, while a structured field may only contain one or two errors at most.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Summary

- At the end of this module, you should be able to:
 - describe text mining and its relationship to data mining
 - explain CRISP-DM methodology as it applies to text mining
 - describe the steps in a text-mining project

© 2014 IBM Corporation

Business Analytics software



Workshop 1

Making Preparations for a Text Mining Project



© 2014 IBM Corporation

There are no files for this workshop.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

1-24

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Workshop 1: Making Preparations for a Text Mining Project

Suppose you are working for Astroserve, a major telecommunications company, and were assigned to identify the key topics customers called to complain about when they contact the company call center. It is your job to extract this information from the high volume of calls the company receives each month. Ultimately, Astroserve would like to use the data to predict which customers are likely to churn.

Prior to beginning the assignment, answer the following questions:

1. Given the nature of call center data, would you expect to find that the number of calls with negative comments would exceed those with positive comments?
2. Your goal is to identify factors that might lead to customer churn. What concepts and categories in the text might you expect that might help you predict which customers who are most likely to want to switch to another telecommunications company?
3. Does the company use a lot of specialized terminology? For example, does it use a lot of acronyms? Are there specific products that the company sells? If so, will need to add Types to the dictionary resources to force the extraction of these specialized terms. What is some of the terminology that is specific to the telecommunications industry? Which ones might be necessary to help predict churn?
4. Does the text you will be analyzing contain a lot of misspellings, abbreviations, and acronyms that will need to be accounted in the Synonym and/or Type dictionaries to successfully extract the text and create meaningful categories?

Workshop 1: Tasks and Results

In this section you will find some possible answers to these workshop questions.

1. Probably Negative. Customers generally contact call centers with a problem or complaint, not a compliment. If the results were not consistent with expectations, you would definitely need to do a detailed review of the concepts included for each type.
2. In order to answer this question, you need to have some idea about the types of things customers call to complain about. Otherwise, you will not be able to easily identify the key concepts you need to focus on among the hundreds or even thousands of concepts that are extracted. This requires that you are already familiar with the types of calls that are coming in as well as the types of complaints the company has received in the past. Here are some possibilities. Can you think of any others?
 - Dropped calls
 - Poor reception
 - Poor service
 - Equipment does not work properly
 - Billing issues
 - Positive feedback about competitors
3. There are several possibilities and you would need to be a subject matter expert to come up with an exhaustive list. Types will need to be added to dictionaries to force the extraction of these items. Here are just a few of them. Can you think of any others?
 - List of Astroserve's competitors customers may express opinions about during the call
 - Specific model names and numbers for the mobile and land-line phones customers may complain about during the call

Names of specific call plans customers may refer to in the call (International, Unlimited, 200 Minutes, etc.)
4. Given the fact that it is call content was probably hand entered by the person who answered the call while they were talking on the phone with the customer, no doubt there will be misspellings. It can probably be also expected that the person taking the call would use abbreviations whenever possible.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE



A Text Mining Overview

IBM SPSS Modeler Text Analytics (v16)



Business Analytics software

© 2014 IBM Corporation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Objectives

- At the end of this module, you should be able to:
 - provide an overview of text mining in IBM SPSS Modeler
 - describe the nodes that were specifically developed for text mining
 - complete a typical text-mining modeling session

© 2014 IBM Corporation

Text mining in Modeler is accomplished via several nodes that are specialized for handling text. This module will begin by briefly presenting each node to familiarize you with them and the text-mining environment in Modeler. Then a simple text-mining example will be run that will allow you to examine how the Text Mining node operates, how concepts are extracted from text, how the interactive workbench environment is setup, and how these concepts are grouped into higher-level categories. This example will provide a practical foundation for the remainder of the course. A lot of details will not be provided, as that type of information will be provided in subsequent modules.

Text Mining Nodes

- Text Mining
- Web Feed
- Translate
- File List
- Text Link Analysis
- File Viewer

© 2014 IBM Corporation



IBM SPSS Text Analytics is an add-on option that can be used to analyze Text can be stored in a database, Statistics file, or a spreadsheet and imported into Modeler for processing, just as with structured data. IBM SPSS Text Analytics for Modeler includes six nodes that read or process data. These nodes are contained in their own Text Mining palette.

Two of the nodes are source nodes because they allow you to import text data from multiple documents or the Web as opposed to within a single text file.

- File List: generates a list of document names as input to the text mining process. This node is used when the text resides in external documents rather than one or more fields in a database or other file. The node outputs a single field with one record for each document or folder listed.
- Web Feed: reads in text from Web feeds, such as blogs or news feeds in RSS or HTML formats. The node outputs one or more fields for each record found in the feeds

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

The remaining nodes are used to process the data:

- **Text Mining:** This node uses linguistic methods to extract key concepts from the text, enables you to create categories with these concepts and other data, and offers the ability to identify relationships and associations between concepts based on known patterns referred to as text link analysis. The node can be used to explore the text data, or to produce either a concept model or category model, and then use that generated model information in subsequent modeling.
- **Translate:** This node is used to translate text, from either fields or documents, from several supported languages, such as Arabic, Chinese, and Russian, into English for modeling. This makes it possible to mine text in a language even if analysts are unable to read that language. The same functionality can be invoked from the text modeling nodes, but use of a separate Translate node makes it possible to cache and reuse a translation in multiple nodes.
- **Text Link Analysis:** This node extracts concepts and also identifies relationships between concepts based on known patterns within the text. Pattern extraction can be used to discover relationships between your concepts, as well as any opinions or qualifiers attached to these concepts. Text Link analysis can also be performed from the Text Mining node.
- **File Viewer:** This node enables you to view text contained in documents from within Modeler. The File Viewer node provides you with direct access to the original text data in documents, and would typically be used with the File List node. The node helps you to better understand the results from the text extraction process by providing you access to the source text from which concepts were extracted, since it is otherwise inaccessible in the stream.

Text Mining Modeling Node

- Used to extract key concepts from the text and create categories with these concepts and other data
- Can be used to explore the text data contents or to produce either a concept model nugget or category model nugget
- Accepts text data from a Web Feed node, File List node, or any of the standard source nodes

© 2014 IBM Corporation



The Text Mining node uses linguistic and frequency techniques to extract key concepts from the text and create categories from those concepts. The node can be used to explore the text data contents or to produce either a concept model nugget or category model nugget. When you execute this modeling node, an internal linguistic extraction engine extracts and organizes the concepts, patterns, and/or categories using natural language processing methods.

This node is located on the IBM SPSS Modeler Text Analytics tab of nodes palette.

You can execute the Text Mining node and automatically produce a concept or category model nugget using the Generate directly option. Alternatively, you can use a more exploratory approach using the Build interactively mode in which not only can you extract concepts, create categories, and refine your linguistic resources, but also perform text link analysis and explore clusters.

This module will focus on the Interactive Workbench mode.

Business Analytics software

IBM

Interactive Workbench

- Four Different Views:
 - Extracted Results pane
 - Categories pane
 - Data Pane
 - Visualization pane

© 2014 IBM Corporation



The interactive workbench has four views for different types of analysis or for editing the dictionary resources. The default view is Categories and Concepts, and in this view there are four panes:

- Extracted Results pane: Located in the lower left corner, this pane is the area in which you perform an extraction and where the results can be edited and refined. The extracted concepts are listed, along with their frequency in the text.
- Categories pane: Located in the upper left corner, this area presents a table of the categories that have been created along with their frequency in the text. The categories can be managed and edited from this pane.
- Data pane: Located in the lower right corner, this pane is initially blank and is used to present the text data corresponding to selections in the other panes. The text is not displayed automatically but will be displayed when the Display button in the Extracted Results or Categories pane is clicked.
- Visualization pane: Located in the upper right corner, this pane provides various graphical representations of categorization (and will provide other types of visualization for other views).

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Extracted Results Pane

- Located in the lower left corner
- Displays extracted concept and their types
- Each concept can be a single word such as "complaint" or compound words such as "phone service"
- Ordered by the number of documents in which they appear

© 2014 IBM Corporation



This area presents the extraction results. When you run an extraction, the extraction engine reads through the text data and identifies the relevant concepts. When concepts are extracted, they are assigned a Type to help group similar concepts. They are color coded according to their type. There are several built-in types delivered with Text Mining for Modeler, such as Location, Product, Person, Positive (qualifiers), and Negative (qualifiers).

Concepts are ordered by their document frequency which is the number of documents/records in which they appear. You also have the number of times a concept appears in the entire set of documents or records, which is referred to as global count.

You can examine the set of underlying terms for a concept by pointing the mouse at the concept name. Doing so will display a tooltip showing the concept name and up to several lines of terms that are grouped under that concept. These underlying terms include the synonyms defined in the linguistic resources (regardless of whether they were found in the text or not) as well as the any extracted plural/singular terms, and so on.

Business Analytics software

IBM

Data Pane

- Located in the lower right corner
- Displays the documents or records corresponding to a selection in another pane (for example, the Extracted Results or Category pane)
- Click the Display button after making a selection to populate the Data pane with the corresponding text

© 2014 IBM Corporation



The Data pane displays one row per document or record corresponding to a selection in another pane. For example, if you select a concept in the Extracted Results pane, Modeler will return the records where the concept was found. The word or words that were selected are highlighted in yellow in the Data pane. If the text data is relatively short in length, the text field displays most or all of the text data. But the call center records can be quite lengthy, and then the text field column shows a short portion of the text and there is a Text Preview pane to the right that displays more of the text. This may not be visible immediately. To see it, you may need to maximize the interactive workbench window and move the pane divider between the two columns so both are visible.

In the text data, all words in color have been extracted, so you can see that a large portion of the text in this record was extracted. All words in black were not extracted. The words that are not extracted are often connectors (for example, "that"), verbs (such as, "advised" or "purchased"), or pronouns. These words are used during the extraction to make sense of the text but are not terms in themselves. Some words that could be verbs, such as "repair" may have been extracted, but in the context of the text, "repair" may be a noun, not a verb. This is a result of the natural language processing used by Modeler.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

2-9

Categories Pane

- Located in the upper left corner
- Represent higher level themes in the text data
- Categories are built by using automated techniques such as semantic networks and concept inclusion, or by creating them manually
- Category names are based on key concept(s) extracted from the set of text

© 2014 IBM Corporation



This area presents a table in which you can manage any categories you build. After extracting the concepts and types from your text data, you can begin building categories by using techniques such as semantic networks and concept inclusion, or by creating them manually. They can be thought of as higher-level concepts that represent higher-level ideas and information in the text. They can also represent all those terms that use certain words such as the terms using "mobile". Categories can also represent a single concept depending on the choice of method and settings. Some concepts are important enough, or unusual enough, that they represent a distinct category. Category names are based on the key concept(s) extracted from the set of text.

When you select a row in the pane, you can then display information about corresponding documents/records or descriptors in the Data and Visualization panes.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Business Analytics software

IBM

Visualization Pane

- Located in the upper right corner
- Offers three perspectives on the commonalities in document or record categorization:
 - Category Bar Chart
 - Category Web Graph
 - Category Web Table
- Shows overlap between categories



© 2014 IBM Corporation

The Visualization pane in the Categories and Concepts view offers the following graphs and charts:

- Category Bar Chart. A table and bar chart present the overlap between the documents/records corresponding to your selection and the associated categories. The bar chart also presents ratios of the documents/records in categories to the total number of documents/records.
- Category Web Graph. This graph presents the document/record overlap for the categories to which the documents/records belong according to the selection in the other panes.
- Category Web Table. This table presents the same information as the Category Web tab but in a table format. The table contains three columns that can be sorted by clicking the column headers.

When refining categories, you can use this pane to review your category definitions to uncover categories that are too similar (for example, they share more than 75% of their documents or records) or too distinct. If two categories are too similar, you can combine the two categories. Alternatively, you might decide to refine the category definitions by removing certain descriptors from one category or the other.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

2-11

Business Analytics software

IBM

Resource Editor

- Available from the pull-down menu in the upper right corner of the Interactive Workbench
- Allows you to edit and refine the linguistic resources to tune the dictionaries
- Displays four panes:
 - Type pane
 - Synonyms pane
 - Excluded pane
 - Libraries pane

© 2014 IBM Corporation



The Resource Editor displays four panes.

- The Library pane in the top left shows libraries you loaded at the start of your session. There is always a Local library in an interactive workbench session, by default, although it is empty when you begin.
- The Type pane is located in the upper center section of the window and displays the types and associated terms. For example, Great Britain and Los Angeles are associated with the Location type and are in the Core library. Notice that all text is in lower case. In the case of the Location type, it may seem odd that only a few locations are listed. The reason is that the vast majority of supplied type information is in compiled resources that are not visible to the user. Thus Modeler is able to successfully recognize thousands of geographical locations from the compiled Location type resources. You only need to add other locations if they are not recognized in the extraction process.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

- The Synonym pane at the bottom shows synonyms that are applied to handle words with the same meaning, plural forms, and spelling variants. The Target term will be displayed and used in the extracted results, and any of the Synonyms found will be replaced by the target. Invariably, you will make changes to the synonyms to tune the results for a specific text.
- The Excluded pane on the right hand side lists words that are not to be extracted, usually because they are not meaningful or they add clutter to the results. In the Astroserve call center data, the terms "cust" and "customer" fall into this category, and you may want to consider excluding them from extraction.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

2-13

Typical Text Mining Session

- Importing the text data
- Extracting concepts
- Categorization
- Model generation
- Editing resources

© 2014 IBM Corporation



Conducting a text mining analysis comprises these essential steps:

- Import: Text data are read into Modeler, either in a field, from documents, or from a Web feed.
- Extraction: An extractor engine automatically locates and collects the key terms from the text data. It also collects these terms into higher-level groups. Types are collections of similar terms, such as organizations, products, or positive statements. Patterns are combinations of terms and types that represent qualifiers, such as positive comments about an organization.
- Categorization: Using linguistic methods, co-occurrence rules, or a standard term frequency approach, categories are created from the extracted results. The categories represent higher-level concepts that capture the chief ideas and key information in the text data.
- Model Generation: Once categories are created, a text mining model node can be generated that allows text data to be scored so that text data can be turned into structured information, typically flag fields indicating whether a concept is expressed in the text in a record, document, etc.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

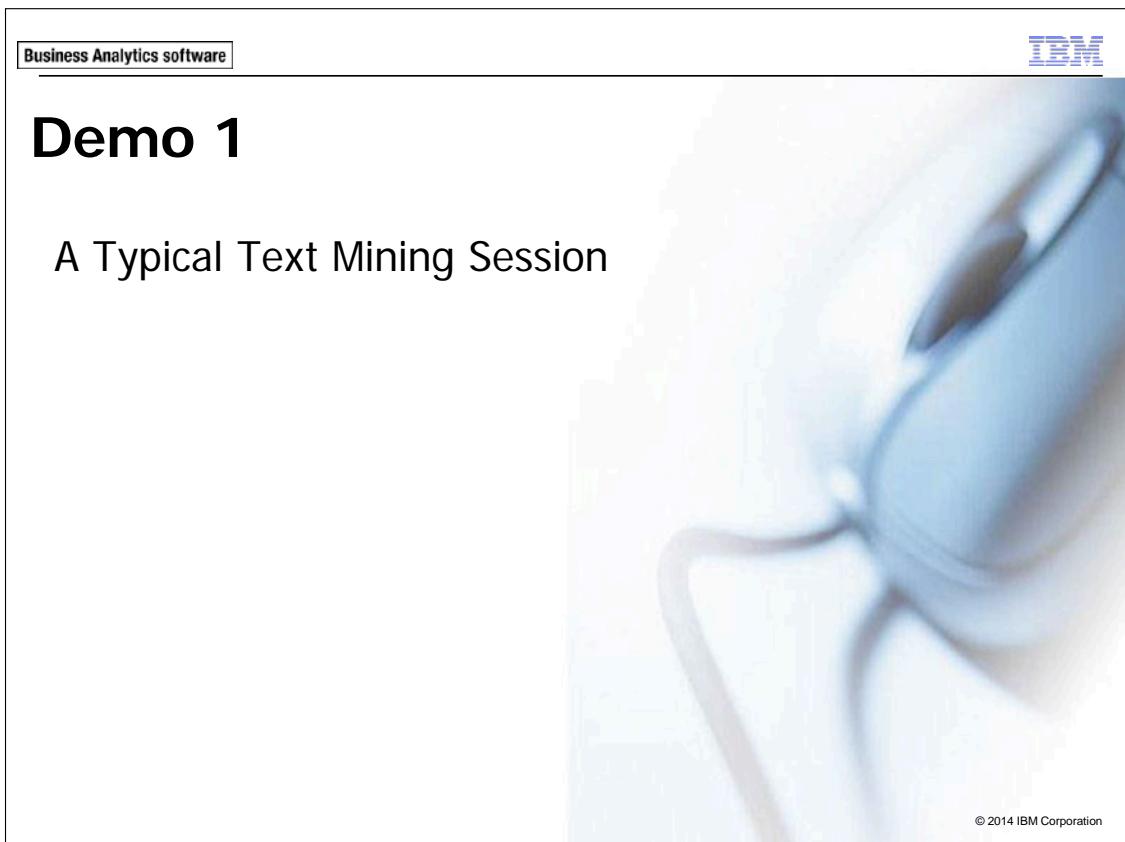
- Editing Resources: Although the steps above are sufficient to conduct text mining, they are usually not adequate. Invariably, you will edit the dictionary resources supplied with Modeler, adding information and making modifications to ensure that the appropriate terms are extracted and categories created. This editing is done using the interactive workbench. Editing is an iterative process, and it can occur at any point after data have been imported.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

2-15



Before you begin with the demo, ensure that:

- You have started Modeler.
- You have set Modeler's working folder. (In Modeler, from the File menu, click Set Directory. When you are working in an IBM operated environment, browse to the folder C:\Train\0A105 and then click Set to select it. When you are not working in an IBM operated environment, ask your trainer for the folder name.)
- Set available memory in Modeler to 1MB of memory. To do so, from the Tools menu, point to Options, click System Options and type 1024 in the Maximum memory (MB) box.

This demo uses the following datasets coming from a (fictitious) telecommunications firm:

- C:\Train\0A105\Astroserve0304.txt - a tab delimited file storing call center data for March and April

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

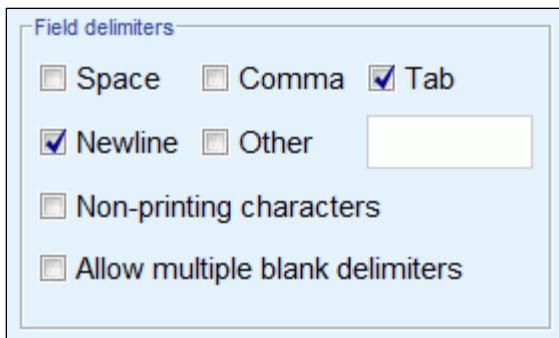
Demo 1: A Typical Text Mining Session

Purpose:

You have been assigned by Astroserve to analyze customer calls collected their call center. Your initial goal is to just extract the key terms in order to get a preliminary look at the nature of their complaints. Ultimately, you would like to use your findings to predict which customers are likely to churn.

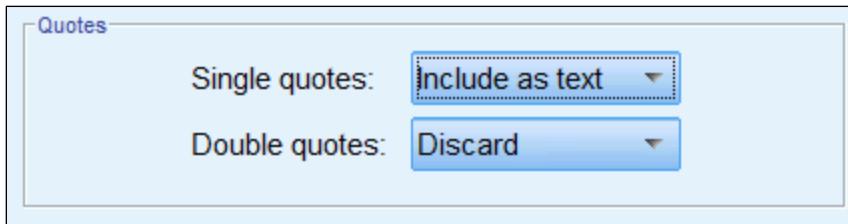
Task 1. Taking an initial look at the Astroserve call center data.

1. Add a **Var. File** node from the **Sources** palette to the stream canvas.
2. Edit the **Var. File** node.
3. Select the file **Astroserve0304.txt** from the **C:\Train\0A105** folder.
4. Under **Field delimiters**, select **Tab** and deselect the **Comma** check box.



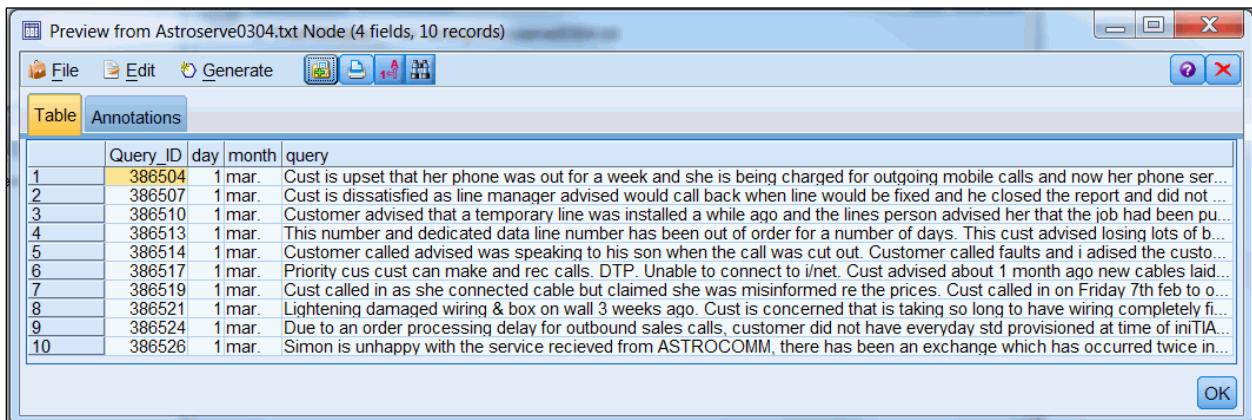
How quotes are handled can be critical when text data is read. Also, since text data often contains commas, normally a comma is not used as the delimiter character.

5. Under **Quotes**, beside **Single quotes**, select **Include as text** and beside **Double quotes**, select **Discard**.



6. Click **Preview**  to preview the data.

The results appear as follows:



	Query_ID	day	month	query
1	386504	1	mar.	Cust is upset that her phone was out for a week and she is being charged for outgoing mobile calls and now her phone ser...
2	386507	1	mar.	Cust is dissatisfied as line manager advised would call back when line would be fixed and he closed the report and did not...
3	386510	1	mar.	Customer advised that a temporary line was installed a while ago and the lines person advised her that the job had been pu...
4	386513	1	mar.	This number and dedicated data line number has been out of order for a number of days. This cust advised losing lots of b...
5	386514	1	mar.	Customer called advised was speaking to his son when the call was cut out. Customer called faults and i advised the custo...
6	386517	1	mar.	Priority cus cust can make and rec calls. DTP. Unable to connect to i/net. Cust advised about 1 month ago new cables laid...
7	386519	1	mar.	Cust called in as she connected cable but claimed she was misinformed re the prices. Cust called in on Friday 7th feb to o...
8	386521	1	mar.	Lightening damaged wiring & box on wall 3 weeks ago. Cust is concerned that is taking so long to have wiring completely fi...
9	386524	1	mar.	Due to an order processing delay for outbound sales calls, customer did not have everyday std provisioned at time of initIA...
10	386526	1	mar.	Simon is unhappy with the service received from ASTROCOMM, there has been an exchange which has occurred twice in...

There is an ID field for each customer call (Query_ID), day and month fields, and the text field itself (query). The entries are of varied length, and can be lengthy. Note that there are spelling errors in the text ("lightening" and "recieved"), abbreviations ("cust" and "i/net"), different date formats, and special terms specific to this organization or industry ("ntu" and "DTP"). Sometimes the linguistic resources will automatically handle these situations, but often some editing of the dictionary resources will be necessary.

7. Close the **Preview** window.
8. Click **OK**.

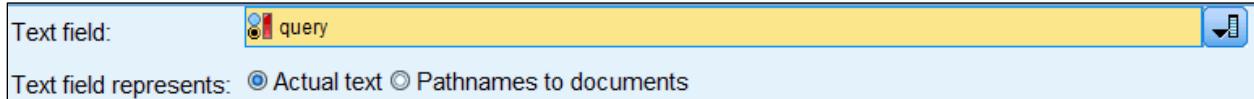
Task 2. Performing the text analysis.

1. Click the **IBM SPSS Text Analytics** tab and then add a **Text Mining** node to the stream.

You may have noticed a delay when adding the Text Mining node to the stream. The first time you add a Text Mining node to the stream in a Modeler session, the software has to install resources to be available to the node when run. This makes the node "heavy" and requires more loading time.

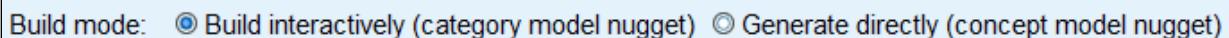
2. Connect the **Text Mining** node to the **Var. File** node.
3. Edit the **Text Mining** node.

4. Beside **Text field**, click the **Field Chooser** button , select **query**, and then ensure that **Actual text** is selected beside **Text field represents**.



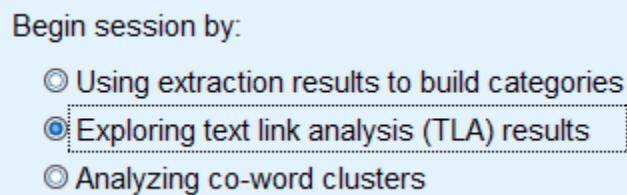
Actual text is checked in the Text field Represents area because you are reading from a text field in the current file. If you were instead reading from documents, you would select the Pathnames to documents option.

5. Click the **Model** tab.
 6. Beside **Build mode**, ensure that the **Build interactively (category model nugget)** is option selected.

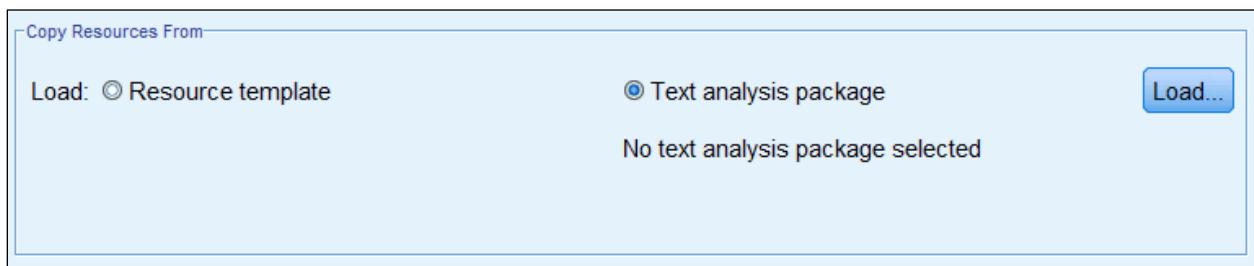


The default is Build interactively (category model nugget), which will open a special interactive environment in which you can perform exploratory analysis. Alternatively, the Generate directly (concept model nugget) selection will build a model automatically, using the node settings.

7. Under **Begin session by**, select **Exploring text link analysis (TLA) results** to produce more results in the interactive workbench.

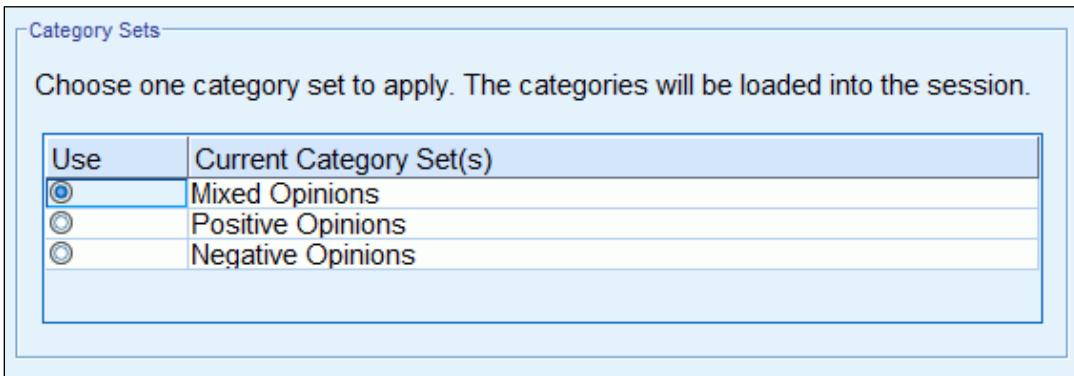


8. Under **Copy Resources From**, select **Text analysis package** and then click **Load**.



9. Select **Customer_Satisfaction.tap** from the list of **Text Analysis Packages**. You are selecting this TAP (Text Analysis Package) because it contains linguistic resources and categories that are designed to capture customer opinions, which is the purpose with the Astroserve data. IBM SPSS Text Analytics for Modeler offers several pre-built TAP files that are fine tuned for specific types of surveys.

10. Under **Category Sets**, ensure that **Mixed Opinions** is selected.



11. Click the **Load** button.
12. Click the **Expert** tab.
13. Select **Accommodate spelling for a minimum root character limit of**.

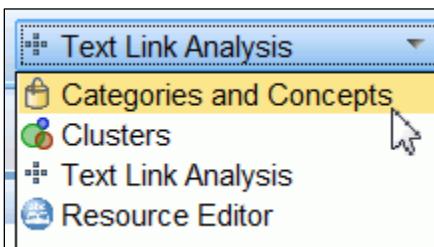
By default the option to try to fix spelling errors is off but it will be used for this text because text entry was rather haphazard and done on the fly by customer service representatives.

14. Click **Run**.

When the extraction is complete, the extracted results are displayed in the interactive workbench.

Task 3. Examining the results in the Interactive Workbench.

1. In the Interactive Workbench, ensure that **Categories and Concepts** is selected in the list at the top right corner.



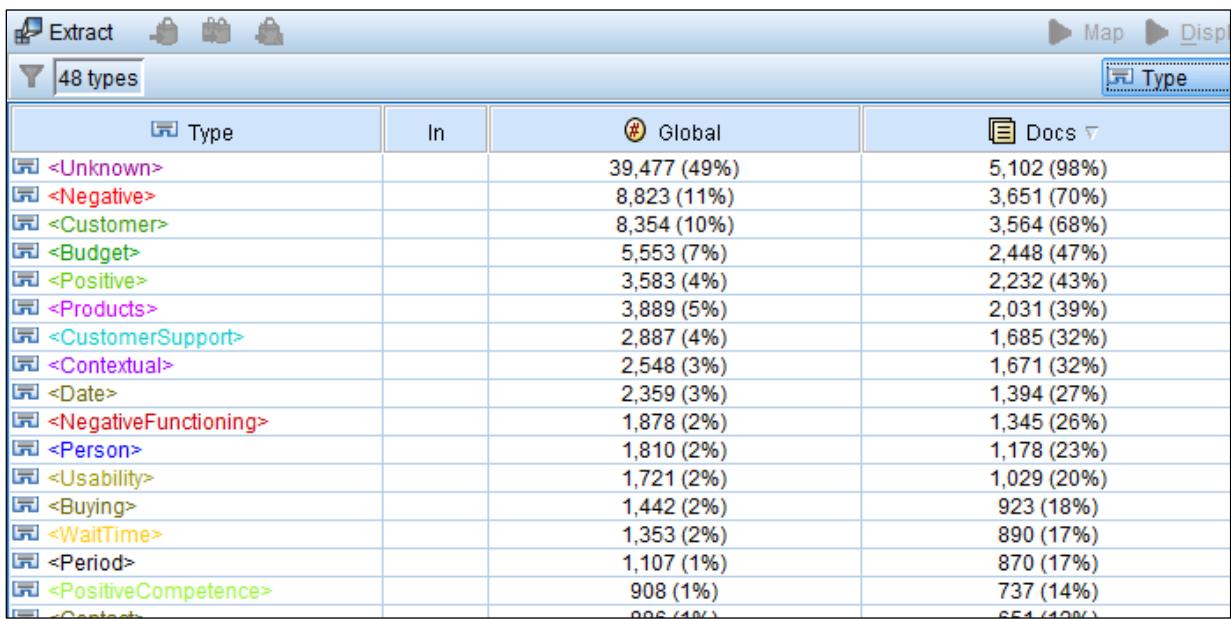
The extracted concepts will appear in the Extract results pane in the lower left corner of the interactive workbench.

The screenshot shows the 'Interactive Workbench - query' window. The 'Extract' pane is active, displaying a table of 5,000 concepts. The columns are labeled 'Concept', 'In', 'Global', 'Docs', and 'Type'. The 'Docs' column shows the number of documents each concept appears in, such as 'customer' appearing in 8,354 documents (10% of the total). The 'Type' column indicates the category of the concept, such as '<Customer>' or '<Negative>'. A message in the pane states: 'To show bars or graphs after building categories, make a selection in another pane and click Display.' Below the table, it says: 'To populate data pane, make a selection in another pane and click Display to see corresponding results.' The top navigation bar shows 'Categories and Concepts' is selected.

Concept	In	Global	Docs	Type
customer	fx	8,354 (10%)	3,564 (68%)	<Customer>
complaint	fx	1,565 (2%)	1,244 (24%)	<Negative>
service	fx	1,722 (2%)	1,225 (23%)	<Unknown>
astrocomm	fx	1,521 (2%)	1,070 (20%)	<Unknown>
fault	fx	1,458 (2%)	946 (18%)	<Negative>
problem	fx	1,348 (2%)	941 (18%)	<Negative>
calls	fx	1,064 (1%)	760 (15%)	<Unknown>
technical support	fx	1,154 (1%)	726 (14%)	<Customer Support>
line	fx	975 (1%)	643 (12%)	<Wait Time>
connection	fx	933 (1%)	632 (12%)	<Usability>

The default display in this pane shows the extracted concepts. These are single words, such as "complaint" or compound words, such as "phone service". All words are displayed in lower case. The concepts are ordered by their document frequency (Docs column), which is the number of documents/records in which they appear. The concept "service" appears in 1,225 records, which is 23% of the total number of customer calls. The most frequent concept is "customer", which is expected for call center data from customers. The first interesting concepts that may be important for text mining that are visible are probably "service", "complaint", and "fault".

2. In the upper right corner of the **Extraction Results** pane, click the **View selection** button, and then select **Type**.

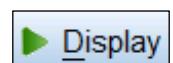


The screenshot shows the Extraction Results pane with a table titled "Type". The table has four columns: "Type", "In", "Global", and "Docs". The "Type" column lists various text types, some of which are colored (e.g., <Unknown> is blue, <Negative> is red). The "In" column shows the count of occurrences, and the "Global" and "Docs" columns show the count and percentage of documents containing each type. The table includes rows for <Unknown>, <Negative>, <Customer>, <Budget>, <Positive>, <Products>, <CustomerSupport>, <Contextual>, <Date>, <NegativeFunctioning>, <Person>, <Usability>, <Buying>, <WaitTime>, <Period>, <PositiveCompetence>, and <Contact>.

Type	In	Global	Docs
<Unknown>		39,477 (49%)	5,102 (98%)
<Negative>		8,823 (11%)	3,651 (70%)
<Customer>		8,354 (10%)	3,564 (68%)
<Budget>		5,553 (7%)	2,448 (47%)
<Positive>		3,583 (4%)	2,232 (43%)
<Products>		3,889 (5%)	2,031 (39%)
<CustomerSupport>		2,887 (4%)	1,685 (32%)
<Contextual>		2,548 (3%)	1,671 (32%)
<Date>		2,359 (3%)	1,394 (27%)
<NegativeFunctioning>		1,878 (2%)	1,345 (26%)
<Person>		1,810 (2%)	1,178 (23%)
<Usability>		1,721 (2%)	1,029 (20%)
<Buying>		1,442 (2%)	923 (18%)
<WaitTime>		1,353 (2%)	890 (17%)
<Period>		1,107 (1%)	870 (17%)
<PositiveCompetence>		908 (1%)	737 (14%)
<Contact>		906 (1%)	651 (12%)

The display is similar to that for concepts, with frequencies by both documents and globally. By far the most frequent type is Unknown, and you can expect this to be true the first time you extract text data using the default linguistic resources. Even so, Modeler has found 2,359 occurrences of dates, 1,810 occurrences of persons, and 5,553 occurrences of budget related items in the text, among other types.

3. Click the **Negative** type to select it, and then click the **Display** button.



4. Click the second text record to select it.

The screenshot shows the IBM SPSS Text Miner interface. The Data pane on the left displays three rows of extracted text records. Row 1 contains text about a customer being upset about a phone service issue. Row 2 contains text about a customer being dissatisfied with a line manager's response. Row 3 contains text about a customer being advised by a temporary person. The Text Preview pane on the right shows a detailed view of the second record, highlighting the word "dissatisfied" in yellow. The interface includes various toolbar icons at the top and a status bar at the bottom.

	query (1000 - Max)	Categories	Text Preview
1	Cust is upset that her phone was out for a week and she is being charged for outgoing mobile calls and now her phone service is out again.		Cust is dissatisfied as line manager advised would call back when line would be fixed and he closed the report and did not call cust.
2	Cust is dissatisfied as line manager advised would call back when line would be fixed and he closed the report and did not call cust.		
3	Customer advised that a temporary line was installed a while ago and the lines person advised her that the job had been put through as urgent but as yet no one has attended. Customer advised that the...		

The Data pane displays one row per document or record corresponding to a selection in another pane (in this instance, the Extracted Results pane). The word or words that were extracted and placed in the Negative type are highlighted in yellow. For the second text entry, "dissatisfied" was extracted from this call center entry.

5. Click the **View selection** button, and then select **Concept**.
6. Click the **Concept** column header until you see this with the arrow pointing down.

7. Scroll down until the concepts related to "mobile" are shown.

Concept	In	Global	Docs	Type
mobile service		8 (0%)	8 (0%)	<Unknown>
mobile service		58 (0%)	55 (1%)	<Unknown>
mobile repayment		15 (0%)	12 (0%)	<Unknown>
mobile reception		3 (0%)	3 (0%)	<Characteristics>
mobile rates		21 (0%)	19 (0%)	<Budget>
mobile plan		9 (0%)	9 (0%)	<Unknown>
mobile phone rates		2 (0%)	2 (0%)	<Products>
mobile phone plan		2 (0%)	2 (0%)	<Products>
mobile phone		2 (0%)	2 (0%)	<Products>
mobile phone calls		5 (0%)	5 (0%)	<Products>
mobile phone bill		3 (0%)	3 (0%)	<Products>
mobile phone		3 (0%)	3 (0%)	<Products>
mobile phone		69 (0%)	64 (1%)	<Products>
mobile number		2 (0%)	2 (0%)	<Unknown>

There are many separate concepts that begin with the word mobile, such as mobile rates or mobile plan. In the extraction process, although these terms are related, as they all refer to "mobile" (cell phone) use or "service", they are kept separate. This is to allow maximum flexibility, but also because text extraction is, in reality, a two-step process. First the program must find the meaningful information in the text. Second, the program groups related concepts, which is the process of categorization.

8. In the top right corner of the Interactive Workbench, click the list and select **Resource Editor**.
9. Click **Opinions Library (English)** in the library tree corner.

10. Scroll down until you see the word "Negative" listed in the **Type** column.

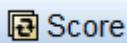
Term =	Match	Inflect	Type	Library
works well	Entire (no compounds)	<input type="checkbox"/>	PositiveFunctioning	Opinions Library (Eng)
works when needed	Entire (no compounds)	<input type="checkbox"/>	PositiveFunctioning	Opinions Library (Eng)
works where ever i go	Entire (no compounds)	<input type="checkbox"/>	PositiveFunctioning	Opinions Library (Eng)
works wherever i go	Entire (no compounds)	<input type="checkbox"/>	PositiveFunctioning	Opinions Library (Eng)
a bit less than expected	Entire (no compounds)	<input type="checkbox"/>	Negative	Opinions Library (Eng)
a little doubt	Entire (no compounds)	<input type="checkbox"/>	Negative	Opinions Library (Eng)
a little tired of	Entire (no compounds)	<input type="checkbox"/>	Negative	Opinions Library (Eng)
abashed	Entire (no compounds)	<input type="checkbox"/>	Negative	Opinions Library (Eng)
abhor	Entire (no compounds)	<input type="checkbox"/>	Negative	Opinions Library (Eng)
abnormal	Entire (no compounds)	<input type="checkbox"/>	Negative	Opinions Library (Eng)
abnormality	Entire (no compounds)	<input type="checkbox"/>	Negative	Opinions Library (Eng)
abominable	Entire (no compounds)	<input type="checkbox"/>	Negative	Opinions Library (Eng)
abrasive	Entire (no compounds)	<input type="checkbox"/>	Negative	Opinions Library (Eng)
abrupt	Entire (no compounds)	<input type="checkbox"/>	Negative	Opinions Library (Eng)

The Resource Editor window displays the libraries contained within the Customer Satisfaction Opinions Text Analysis Package (TAP). Notice that the Opinions Library (English) contains several negative terms. Extracted terms that match any of these terms will be typed as Negative.

11. In the top right corner of the Interactive Workbench, click the list and select **Categories and Concepts**.

The categories created from the Call Center Data are shown in the Categories pane in the upper left corner. Notice that the Docs column now has an icon with two arrows in each cell for each category rather than a document count. This is because the data must be scored to determine which records contain which categories.



12. Click the **Score** button  to calculate the number of records for each category listed.



Category	Description	Docs
All Documents	-	5220
Uncategorized	-	1166
No concepts extracted	-	2
Contx: Company: Public Image-Reputation	3	156
Contx: Pricing and Billing	5	256
Contx: Quality	1	11
Contx: Service	2	111
Neg: Company: Public Image-Reputation	13	44
Neg: General Dissatisfaction	12	308
Neg: Plan to Change-Not Recommended	3	4
Neg: Pricing and Billing	9	526
Neg: Product: Availability-Variety-Size	12	115

The results show that the category "Neg: General Dissatisfaction" captured the comments from 308 customers who indicated that they are generally dissatisfied with Astroserve. Also, you see that 526 customers are upset with pricing and billing. The category labeled "Uncategorized" lists how many responses are uncategorized (here 1166).

13. Double-click **Neg: General Dissatisfaction** so you can get more detail about the category.

Descriptors	Docs	Type
<i>fx</i> [waste of time]	4	Rule
<i>fx</i> [too many problems]	2	Rule
<i>fx</i> [not satisfied + .]	191	Rule
<i>fx</i> [needs improvement + .]	0	Rule
<i>fx</i> [irritating + .]	1	Rule
<i>fx</i> [frustrating + .]	17	Rule
<i>fx</i> [doesn't meet expectation]	14	Rule
<i>fx</i> [dislike + .]	17	Rule
<i>fx</i> [disappointing + .]	18	Rule
<i>fx</i> [better + .]	11	Rule
<i>fx</i> [<Organization> & <Negative> & !(slow)]	27	Rule
<i>fx</i> [(<Product> <Products>) & (dislike not satisfied)]	20	Rule

The symbol **fx** by each component of this category indicates that each of them is a rule. Rules themselves consist of two or more concepts, types, or patterns. For example, there is a rule linking Product(s) and negative sentiment:

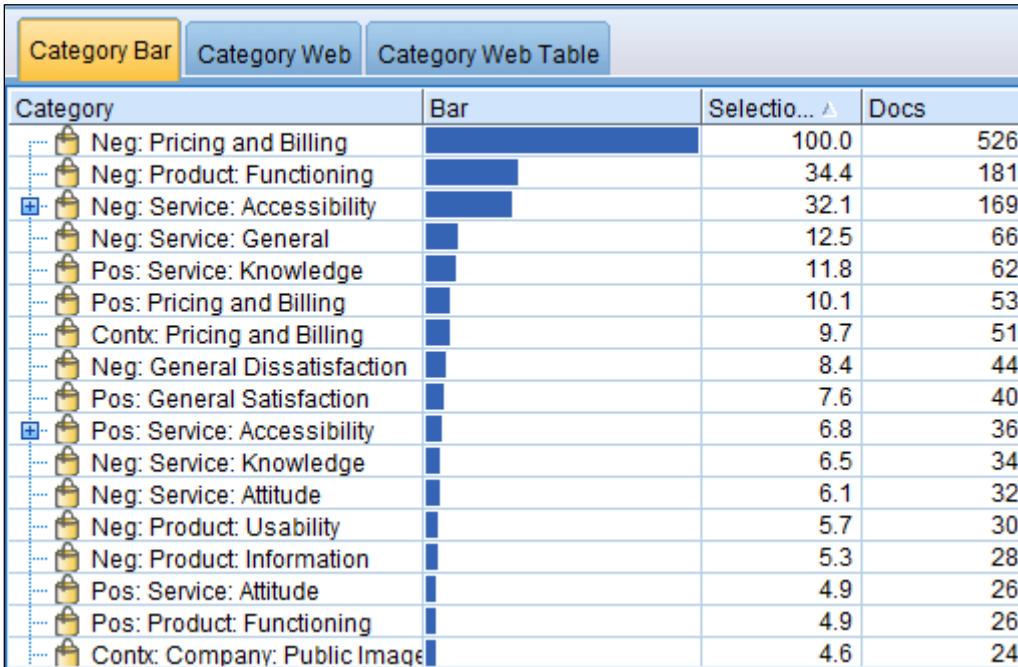
[(<Product> | <Products>) & (dislike | not satisfied)]

This rule places responses in this category if the respondent mentioned a Product and had a negative comment about the product.

14. Close the Category Definitions window.
15. Click the **Neg: Pricing and Billing** category in the categories list.
16. Click the **Category Bar** tab in the Visualization pane (if necessary).
17. In the Categories pane, click **Display**.

After a few moments a bar chart appears in the Visualization pane. You may need to maximize the size of the graph so that the Docs column is visible.

18. Sort Selection % in descending order.

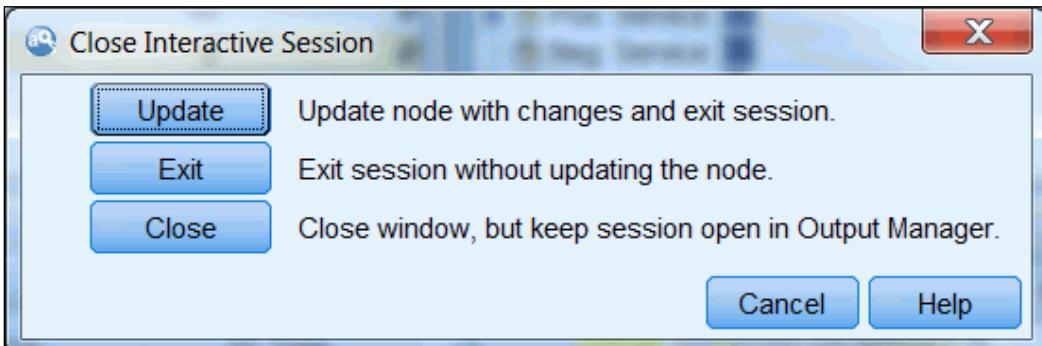


This chart lists the overlap between categories for the 526 records that contain terms in the "Neg:Pricing and Billing" category. Notice that "Neg: Product: Functioning" occurs in 181 of these 526 records, or 34.4%. This type of information can be very helpful in either understanding a category or thinking about other categories that might be created by grouping two, or more, categories that occur together frequently.

19. From the **Generate** menu, click **Generate Model**.

20. From the **File** menu, click **Close** to close the Interactive Workbench.

Modeler provides an option to save the interactive session so you can return to it in the current state, exit the session without saving, or close the window but keep the session available.



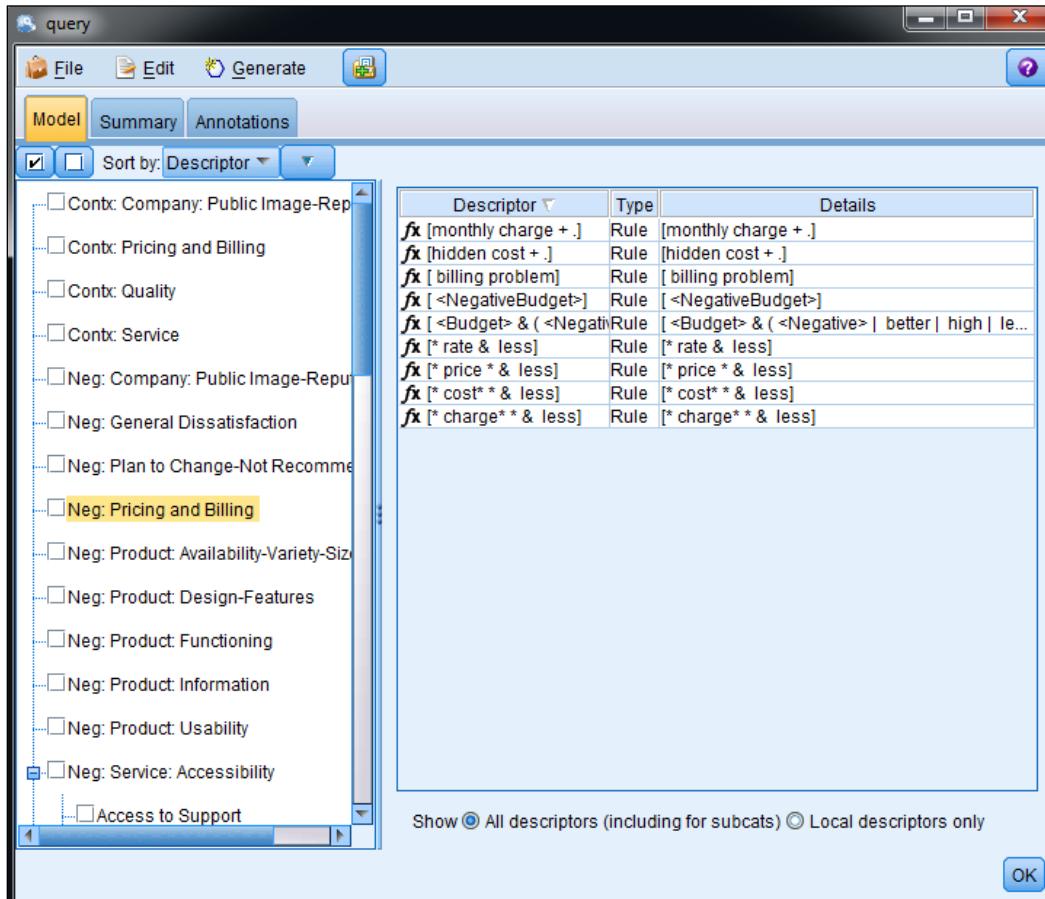
For this example, simply exit.

21. Click **Exit** to return to the Modeler environment.

Task 4. Examining the generated model nugget.

1. In the **Models** managed area, right-click the generated model nugget named **query**, and then click **Browse**.
2. Click the **Neg: Pricing and Billing** category.

Notice each descriptor in this category, the descriptor type, and another column listing details about other items that were added under a specific concept.



3. Click **OK** to close the model nugget.
4. Add the model nugget named **query** downstream from the **Var. File** node.
5. Connect the **data source** node to the **generated model**.
6. Edit the **model nugget**.

- Click **Preview** and scroll to the right in the output past the query field to see the results.

The screenshot shows a 'Preview from query Node (55 fields, 10 records)' window. The table has four columns labeled 'Category_Context_Company_Public_Image-Reputation', 'Category_Context_Pricing and Billing', 'Category_Context_Quality', and 'Category_Context_Service_Level'. The data is as follows:

	Category_Context_Company_Public_Image-Reputation	Category_Context_Pricing and Billing	Category_Context_Quality	Category_Context_Service_Level
1	F	F	F	F
2	F	F	F	F
3	F	F	F	F
4	F	F	F	F
5	F	T	F	F
6	F	F	F	F
7	F	T	F	F
8	F	F	F	F
9	F	F	F	F
10	F	F	F	F

New flag fields have been created, having values of T or F, for each category. A record is coded T (true) if that category is contained in the text for that record. A record is coded F (false) if the category is not in the text.

These new fields can now be used to generate reports, further investigate the relationship between the categories, study the relationship between other fields, such as demographic information, and the categories, or develop models that use the category fields as inputs. You could even use a category field as an outcome and attempt to predict what factors lead to particular comments in customer calls.

- Close the **Preview** window.
- From the **File** menu, click **Exit** to close the Modeler session, and then click **Exit** again, when prompted.

Results:

You extracted the key terms in order to get a preliminary look at the nature of complaints. You used your findings to predict which customers are likely to churn.

Apply Your Knowledge

Purpose:

Text your knowledge of the material covered in this module.

Question 1: True or False: The File List is not used to read text in the form of Web feeds.

- A. True
- B. False

Question 2: True or False: The Text Link Analysis node can be used to explore the text data or to produce either a concept model or category model, and then use that generated model information in subsequent modeling.

- A. True
- B. False

Question 3: Which of the following nodes is used to view the contents of documents from modeler?

- A. Text Mining
- B. Web Feed
- C. Text Link Analysis
- D. File Viewer

Question 4: True or False: Model generation is a required step in any text mining project.

- A. True
- B. False

Question 5: True or False: The Translate node in Text Analytics can be used to translate non-English languages into English and vice versa.

- A. True
- B. False

Apply Your Knowledge - Solutions

- Answer 1: A. True. The Web Feed node should be used to read Web feeds.
- Answer 2: B. False: The Text Link Analysis node is used to perform text link analysis. A Text Mining node should be used to produce a concept or category model.
- Answer 3: D. The File Viewer node.
- Answer 4: B. False. Many text mining projects focus on summarizing the content in the form of reports and/or graphs without doing any modeling.
- Answer 5: B. False. You can only translate non-English languages into English but not the other way around.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

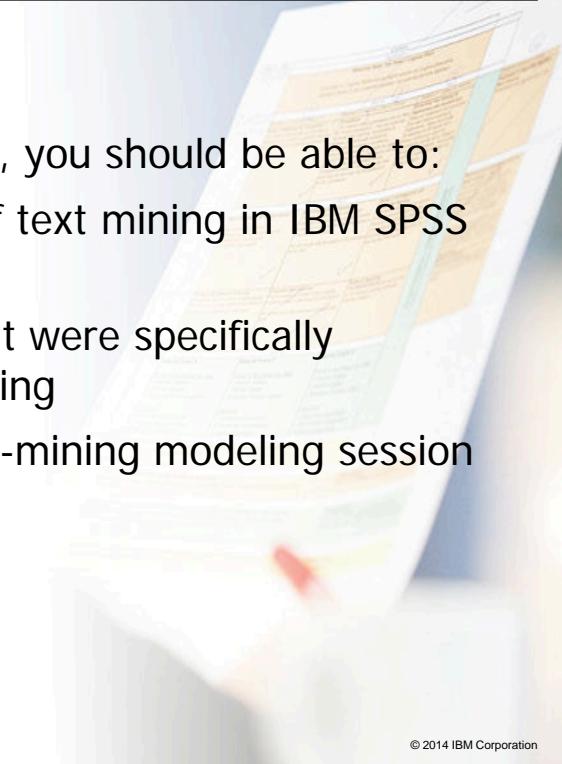
Business Analytics software

IBM

Summary

- At the end of this module, you should be able to:
 - provide an overview of text mining in IBM SPSS Modeler
 - describe the nodes that were specifically developed for text mining
 - complete a typical text-mining modeling session

© 2014 IBM Corporation



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

2-33

Business Analytics software

IBM

Workshop 1

Text Mining Customer Opinions About Portable Music Players



© 2014 IBM Corporation

The following file will be used in this workshop:

- C:\Train\0A105\Music_Survey.sav - a Statistics file containing customer likes and dislikes about portable music players.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Workshop 1: Text Mining Customer Opinions About Portable Music Players

Before actually text mining the data, you should familiarize yourself with the text data contained in music_survey.sav in the Interactive Workbench.

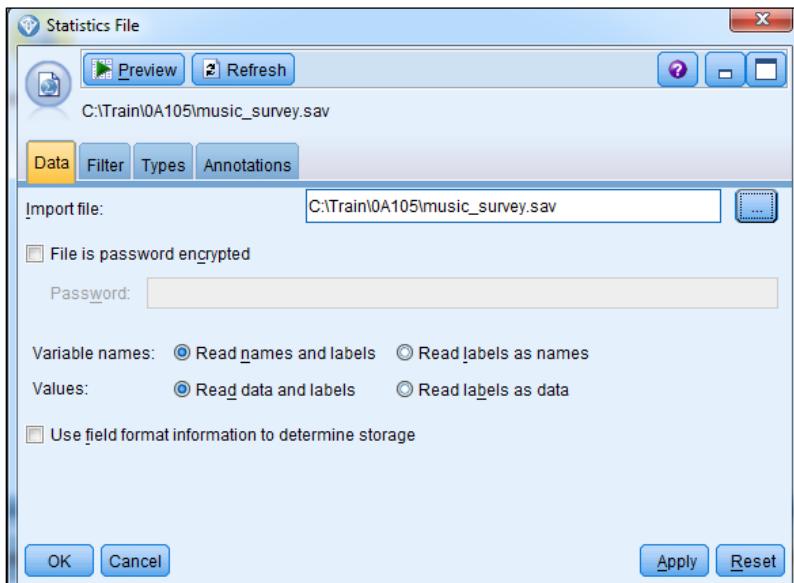
- Begin by starting a new stream in Modeler and import the data from music_survey.sav (An IBM SPSS Statistics file).
- Add a Text Mining node downstream from the Statistics import node, and import the following field, "What do you like most about this portable music player".
- Use the Product Satisfaction TAP, and try using the Positive Opinions category set, since only the positive opinions are stored in this field.
- Create text link analysis patterns.
- Run the Text Mining node.
- Select the Categories and Concepts view to get familiar with what people were saying about portable music players.
- Score the categories.

For more information about where to work and the workshop results, refer to the Tasks and Results section that follows. If you need more information to complete a task, refer to earlier demos for detailed steps.

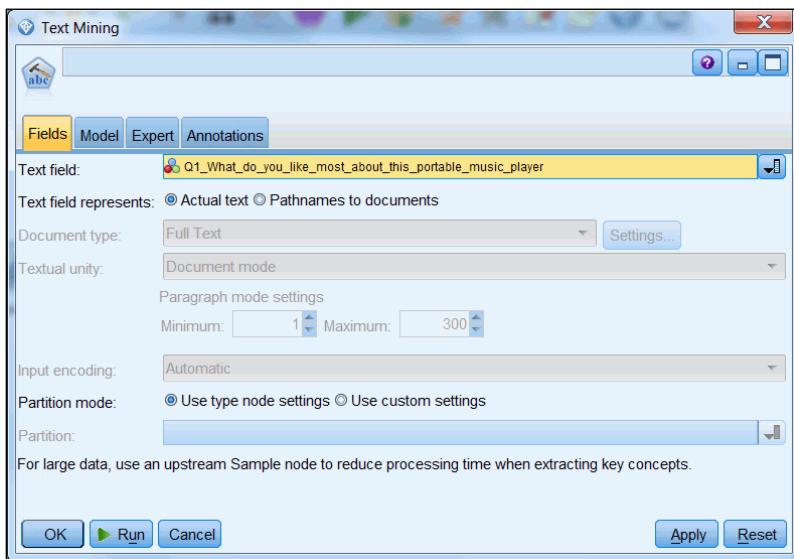
Workshop 1: Tasks and Results

In this workshop, you will use music_survey.sav to familiarize yourself about customer likes and dislikes about portable music players.

- Begin by starting a new stream (**File \ New Stream**) in Modeler, dragging a **Statistics File** from the **Sources** tab, and importing the data from **music_survey.sav**.

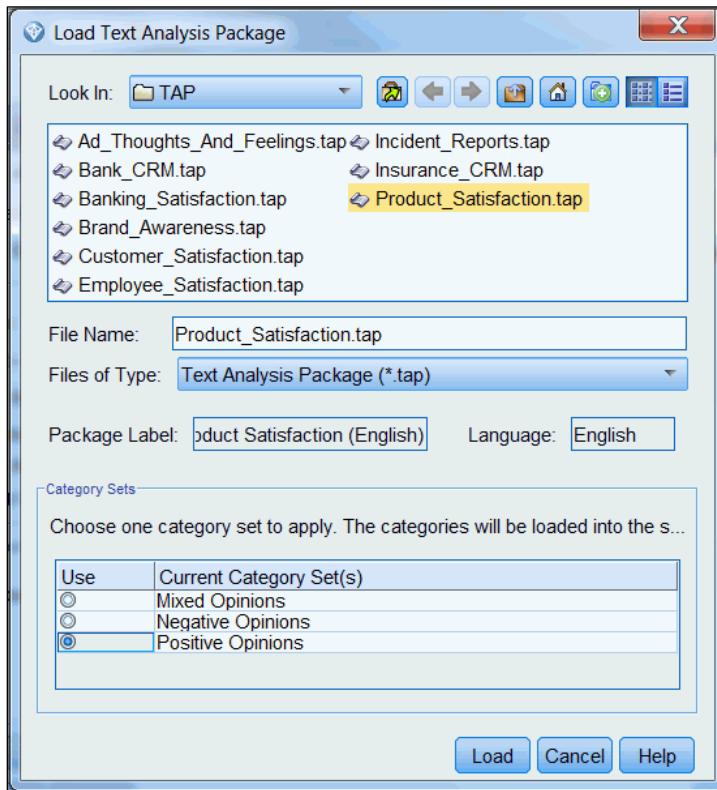


- Add a **Text Mining** node downstream from the **Statistics file**, and import the following field, **Q1_What_do_you_like_most_about_this_portable_music_player**.

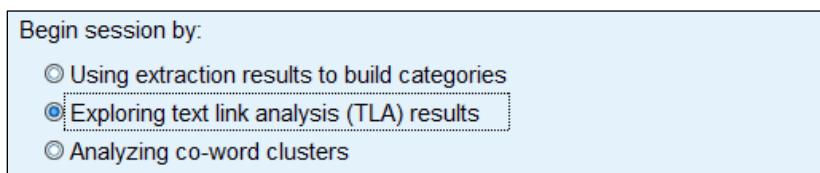


This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

- Use the **Product Satisfaction TAP**, and try using the **Positive Opinions** category set, since only the positive opinions are stored in this variable.

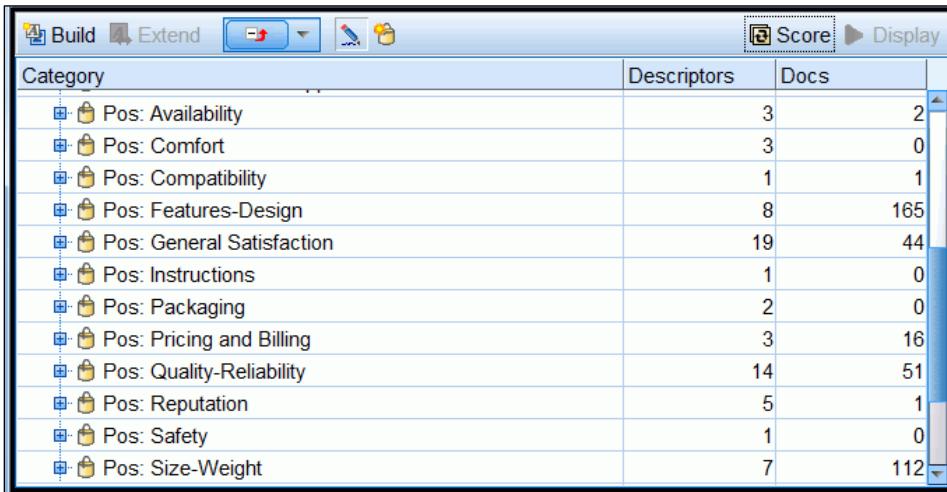


- Create text link analysis patterns.



- Run the Text Mining node.
- Select the **Categories and Concepts** view to get familiar with what people were saying about portable music players.

- Score the categories.



The screenshot shows a software interface with a menu bar at the top. The main area is a table with three columns: 'Category', 'Descriptors', and 'Docs'. The 'Category' column lists various product features and characteristics, each preceded by a small icon. The 'Descriptors' column contains numerical values ranging from 1 to 19. The 'Docs' column contains numerical values and some text labels indicating document counts or specific scores. The total count of documents is 112.

Category	Descriptors	Docs
Pos: Availability	3	2
Pos: Comfort	3	0
Pos: Compatibility	1	1
Pos: Features-Design	8	165
Pos: General Satisfaction	19	44
Pos: Instructions	1	0
Pos: Packaging	2	0
Pos: Pricing and Billing	3	16
Pos: Quality-Reliability	14	51
Pos: Reputation	5	1
Pos: Safety	1	0
Pos: Size-Weight	7	112

It appears that respondents are happiest about the Feature-Design of the portable music players. Many of them were also positive about the quality and reliability, and quite a few of them also said they were generally satisfied with the product.

- Exit from Modeler without saving.



Reading Text Data

IBM SPSS Modeler Text Analytics (v16)

Business Analytics software



© 2014 IBM Corporation

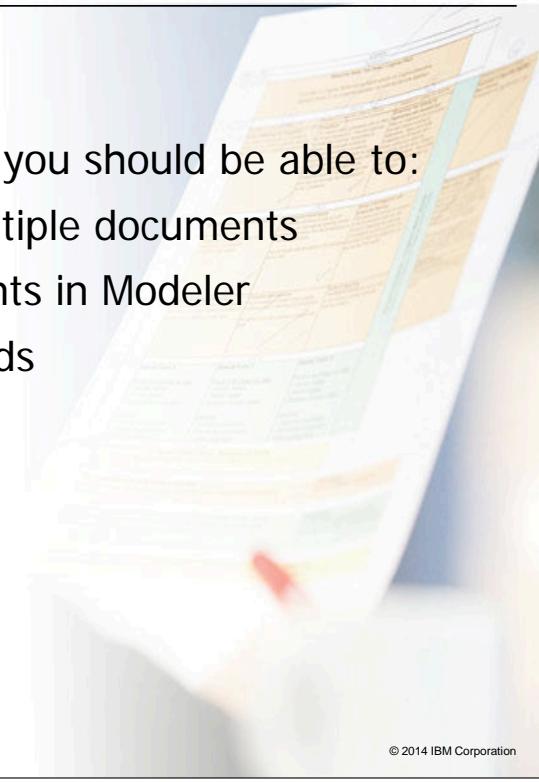
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Objectives

- At the end of this module, you should be able to:
 - read text data from multiple documents
 - view text from documents in Modeler
 - read text from Web feeds



© 2014 IBM Corporation

Text data now comes in a variety of formats, including the standard formats read by Modeler (text files, IBM SPSS Statistics files, various databases, or Excel files), but also in document formats, including Microsoft Word or PowerPoint, HTML, PDF, and other types. Additionally, text data can originate from Web feeds, including news feeds or blogs, in RSS or HTML formats.

Reading text fields from any of the standard data source formats used by Modeler is just as straightforward as when reading in numeric or other data fields. In this module, it will be demonstrated how to read data from non-standard sources including lists of files and Web feeds.

File List Node

- Used for reading text from multiple documents
- Generates a list of documents or folders as input to the text mining process
- Reads in text from unstructured documents saved in formats such as Microsoft Word, Microsoft Excel, and Microsoft PowerPoint, as well as Adobe PDF, XML, HTML, and others

© 2014 IBM Corporation



Text that is stored in one or more fields in a file is easy to read into Modeler. In some instances, the text of interest may be contained in multiple documents/files. The Text Mining node can read text in multiple documents, but a special node must be used to access and read the documents. This node is the File List node, located in the IBM SPSS Text Analytics palette.

The File List node actually generates a list of documents, or folders containing the documents, as input to the text mining process. This is necessary because unstructured text within a file cannot be represented as fields or records in the same way as standard data.

The output from the File List node is not data. Instead, the node outputs a single field, with one record for each document or folder listed. In essence, it creates a list of the documents. This field is, in turn, selected as input for the Text Mining node.

If the number of files is relatively small, then using a list of files works well. With a larger collection of files, it is best to create a list of directories, which speeds processing. The File List node provides this option as well.

Using the File List Node in Text Mining

- The following options must be set in both the Text Mining and Text Link Analysis nodes:
 - Text field represents – must be set to Pathnames to documents
 - Document type – must be set to either Full Text, Structured Text, XML
 - Textual unity – must be set to either Document mode or Paragraph node

© 2014 IBM Corporation



The Pathnames to documents option must be selected when using the File List node. A text field must also be specified, which will be Path.

There are three document types that can be selected:

- Full Text: This option is used for non-tagged documents and most Web pages. This usually includes Word and PowerPoint files, along with PDF formats.
- Structured Text: This option is used for bibliographic forms, patents, and any text pages that contain regular structures that can be identified and analyzed. If you select this option, you must click the Settings button and enter text separators in the Structured Text Formatting area of the Document Settings dialog box.
- XML: This option is used when the documents contain XML-tagged text. If you select this option, you must click the Settings button and explicitly specify the XML elements containing the text to be read during the extraction process in the XML Text Formatting area of the Document Settings dialog box.

Another setting is labeled Textual unity, which is used to specify the type of full text document. This is only available when Full Text is the document type. There are two choices:

- Document mode: Use this for documents that are short and semantically homogenous, such as articles from news services. That fits the data for the example, so leave the selection as Document mode.
- Paragraph mode: Use this for Web pages and other non-tagged documents. The extraction process semantically divides the documents, taking advantage of characteristics such as internal tags and syntax.

If you select the Paragraph mode, there is another setting for minimum and maximum extraction limits.

- Input encoding: This option is available only if the text field represents Pathnames to documents. It specifies the default text encoding. For all languages except Japanese, a conversion is done from the specified or recognized encoding to ISO-8859-1.

Business Analytics software

IBM

Demo 1

Using the File List Node to Read Text from Multiple Files



© 2014 IBM Corporation

This demo uses the following datasets coming from a (fictitious) telecommunications firm:

- C:\Train\0A105\03-Reading Text Data\bms-news, which is a folder that contains two sub folders called bms0108 and bms0109. Each of these subfolders contains multiple HTML format files containing news releases for two months concerning developments in the pharmaceutical industry.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

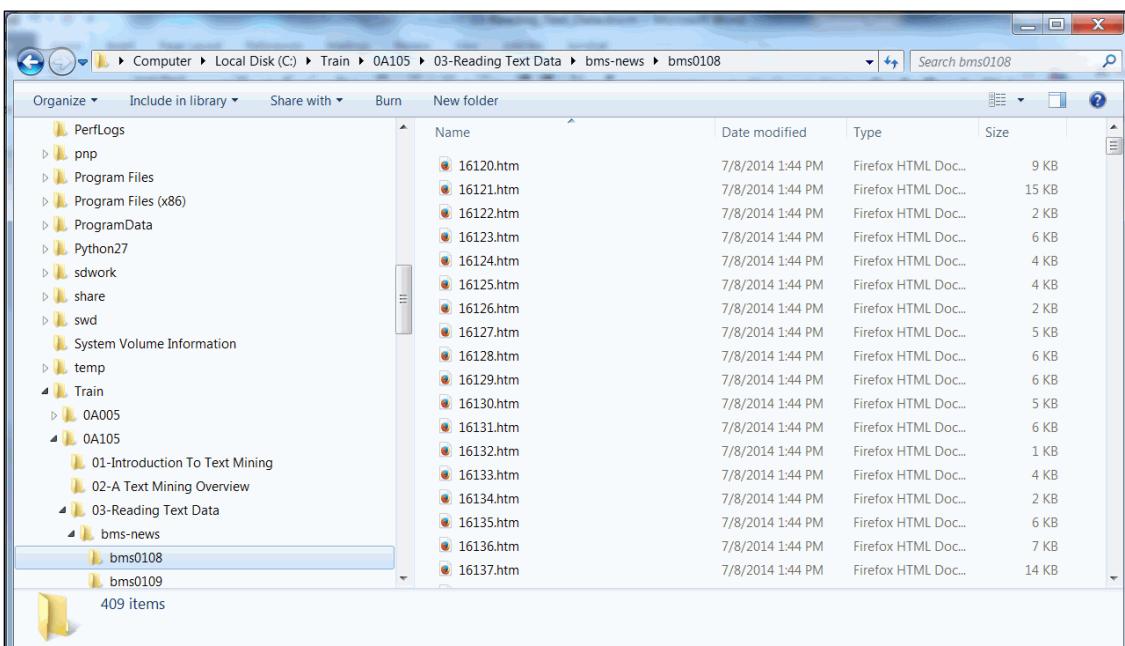
This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Demo 1: Using the File List Node to Read Text from Multiple Files

Purpose:

You would like to analyze multiple files containing news releases for two months concerning developments in the pharmaceutical industry. You would like to text mine all of these documents at once rather than one at a time.

The \bms-news folder contains a number of HTML-format files.

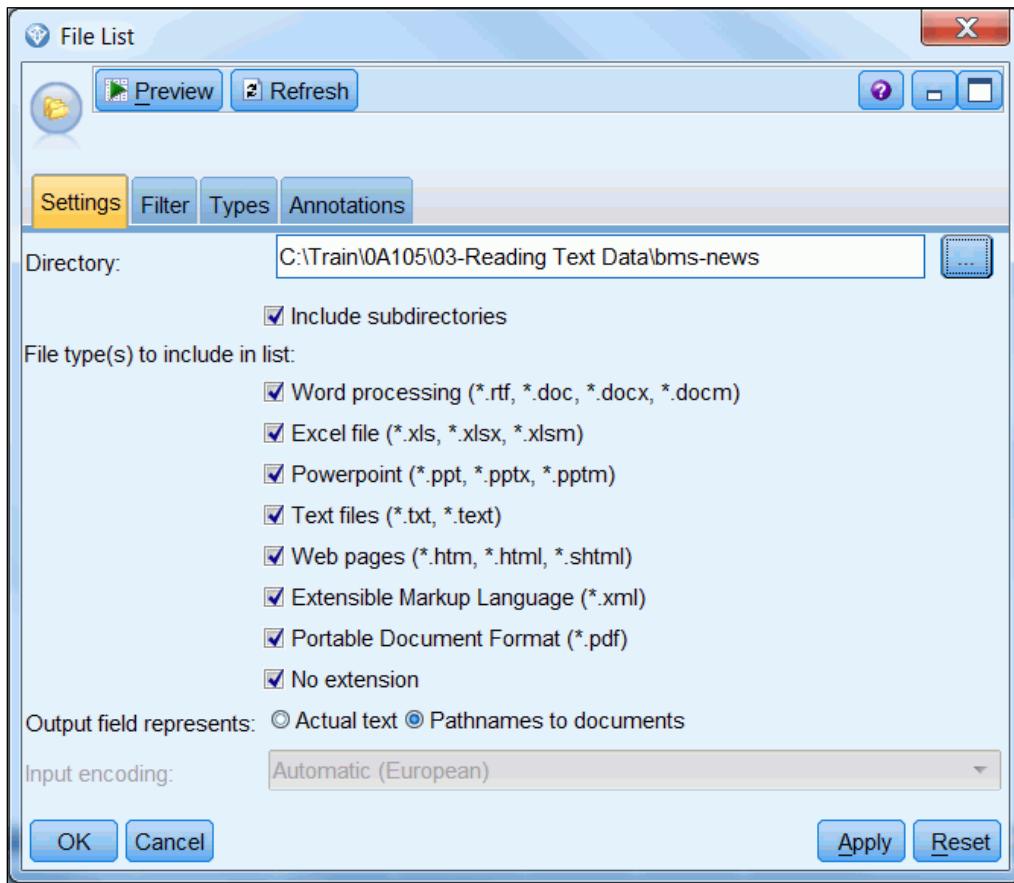


Task 1. Preparing the File List node.

1. From the **File** menu, click **New Stream**.
2. Add a **File List** node from the **IBM SPSS Text Analytics** palette to the stream canvas.
3. Edit the **File List** node.
4. Beside **Directory**, click the **Chooser**  button.
5. Navigate to **C:\Train\0A105\03-Reading Text Data\bms-news**.

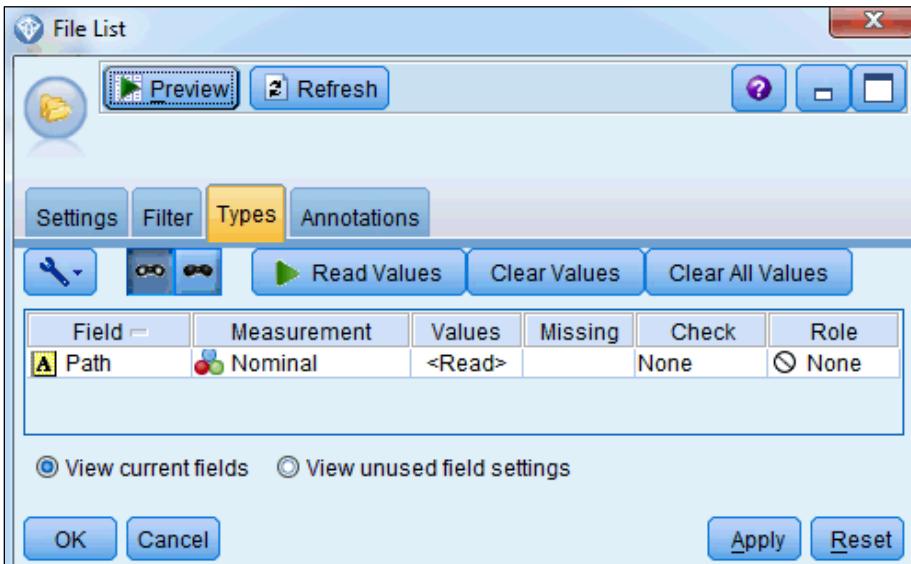
6. Click **Open**.

The results appear as follows:



Modeler will read files from the bms-news subfolder, including any subfolders under that. By default, all the types will be read from the directory and subdirectories listed. If a directory contains only one type of file, then the selection of file type is not critical. Otherwise, you should select only the appropriate file type. Notice that you cannot select specific files, only file types. Therefore if a directory includes files of the same type, and you want to only use some of them when text mining, you will need to move the files that you want to use into a separate directory.

7. Click the **Types** tab, and then click **Read Values**.



The Types tab is a standard Types tab as found on any Source node. Only one field is created, with the name Path. This field is of type Nominal since it will contain file names. You can attach a Table node to the File List node to view the output (or alternatively use the Preview to see a subset of the output).

8. Click **OK**.
 9. From the **Output** tab, add a **Table** node to the stream.
 10. Connect the **File List** node to the **Table** node, and then run the **Table** node.
- A section of the result appears as follows:

The screenshot shows the output of a Table node. It displays 10 records of file paths from 'Path' column:

	Path
1	C:\Train\0A105\03-Reading Text Data\bms-news\bms0108\010810975.htm
2	C:\Train\0A105\03-Reading Text Data\bms-news\bms0108\010810976.htm
3	C:\Train\0A105\03-Reading Text Data\bms-news\bms0108\010810977.htm
4	C:\Train\0A105\03-Reading Text Data\bms-news\bms0108\010810978.htm
5	C:\Train\0A105\03-Reading Text Data\bms-news\bms0108\010810979.htm
6	C:\Train\0A105\03-Reading Text Data\bms-news\bms0108\010810980.htm
7	C:\Train\0A105\03-Reading Text Data\bms-news\bms0108\010810981.htm
8	C:\Train\0A105\03-Reading Text Data\bms-news\bms0108\010810982.htm
9	C:\Train\0A105\03-Reading Text Data\bms-news\bms0108\010810983.htm
10	C:\Train\0A105\03-Reading Text Data\bms-news\bms0108\010810984.htm

There are 1055 records in the Table corresponding to the 1055 files in the two subfolders bms0108 and bms0109. The contents of Path are the respective file names.

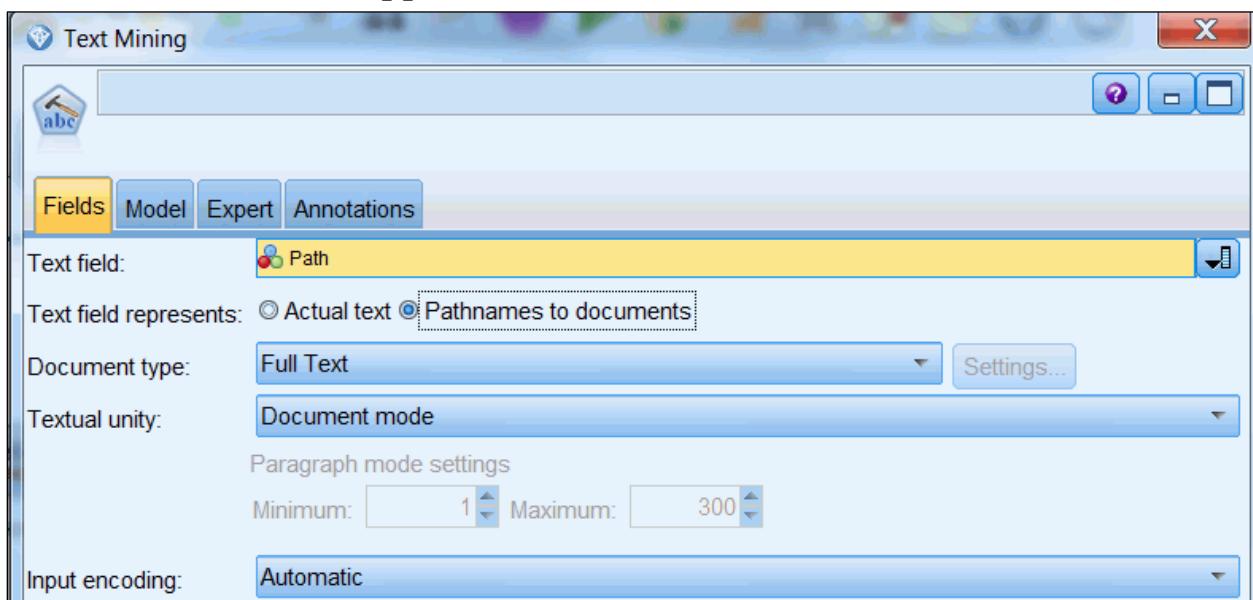
11. Close the **Table** window.

You are now prepared to create a text-mining model. In this demonstration, you will focus on the correct specifications in the Text Mining node for reading text from files and build a model automatically.

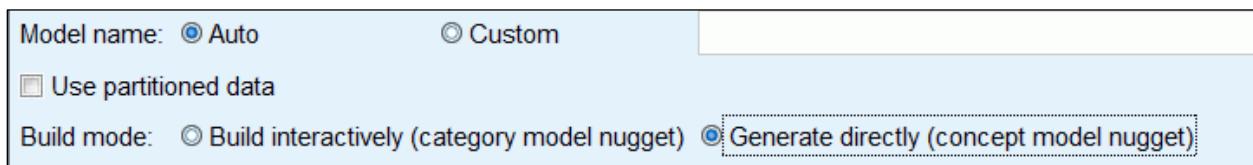
Task 2. Text mining the BMS-News documents.

1. Add a **Text Mining** model node to the stream.
2. Connect the **File List** node to the **Text Mining** node.
3. Edit the **Text Mining** node.
4. Use the field chooser to select **Path** as the **Text** field.
5. Beside **Text field represents**, click **Pathnames to documents**.
6. Leave **Document type** as **Full Text**.
7. Leave **Textual unity** as **Document mode**.

A section of the result appears as follows:



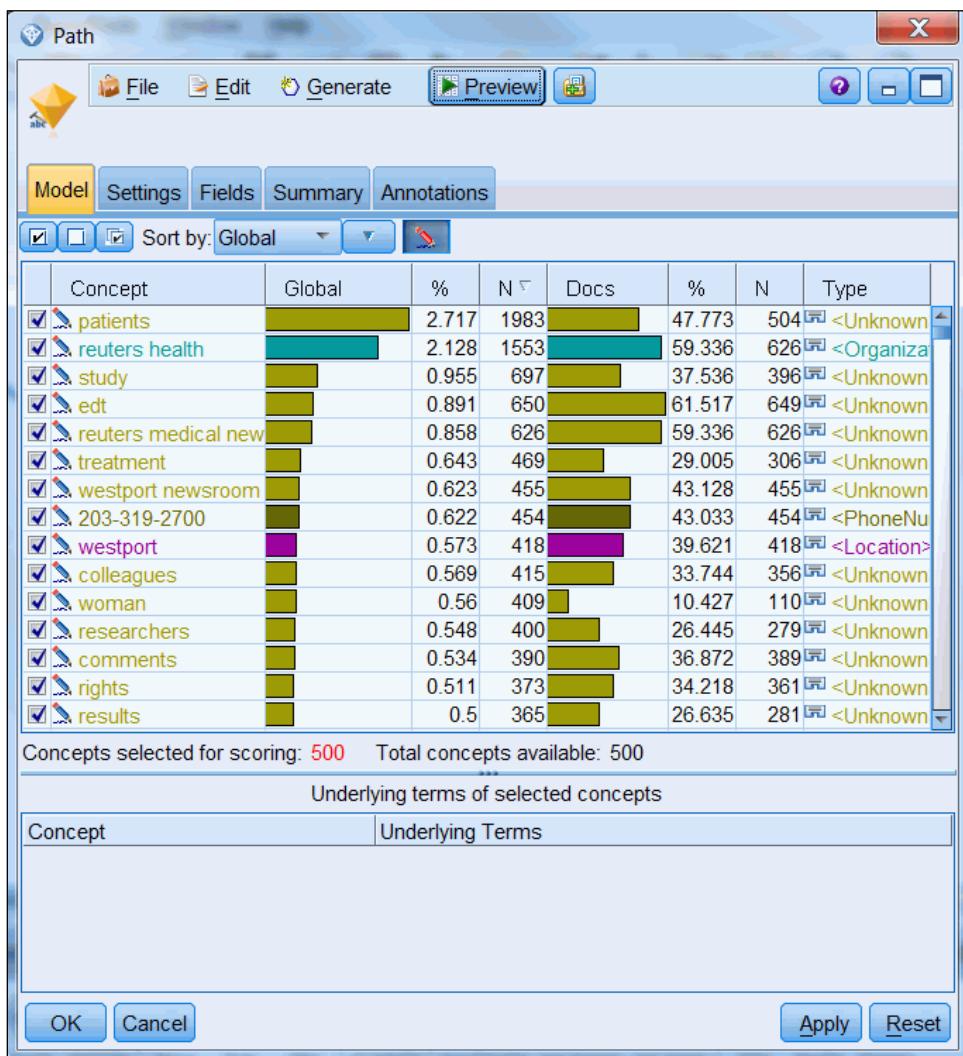
8. Click the **Model** tab.
9. Click **Generate directly (concept model nugget)**.



10. Click **Run**.

After the model is created you will examine the results.

11. Edit the **generated model** named **Path** that has been placed on the stream canvas.



You can observe from the names of the concepts that these data are about medical and biological subjects as well as drug effects. Once the correct specifications have been made in the File List node, and then in the Text Mining node, using text in documents is no different than using text in a field.

12. Close the **Model Browser**.

Do not close the stream. Leave it open for the next demo.

Results:

You have successfully analyzed multiple files containing news releases for two months concerning developments in the pharmaceutical industry.

File Viewer Node

- Allows you to review the contents of each document from within Modeler
- Makes it possible to better understand text extraction and categorization and to help you edit the linguistic resources

© 2014 IBM Corporation



Although text in documents can be successfully used for text mining, there is one salient difference compared to text in fields. If you would like to review specific text for a record, you can do so with the Table node for text in a field. But when a Table node is used to display the Path field for documents, all that is revealed is the file name: the text contained in that file is not displayed.

To view the text to better understand the results of text extraction or categorization, and to help you edit the linguistic resources, the File Viewer node provides direct access to the original data. Add this node to the stream after a File List node. The File Viewer node is also located in the Text Analytics palette.

The result of the File Viewer node is a window showing all of the document elements that were read. From this window, you can click a toolbar icon to launch the report in an external browser listing document names as hyperlinks. You can then click a link to open the corresponding document in the collection.

Business Analytics software

IBM

Demo 2

Using the File Viewer Node to View Documents in Modeler



© 2014 IBM Corporation

You will need to complete Demo 1 before starting this demo.

This demo uses the following datasets coming from a (fictitious) telecommunications firm.

- C:\Train\0A105\03-Reading Text Data\bms-news - a folder that contains two sub folders called bms0108 and bms0109. Each of these subfolders contains multiple HTML format files containing news releases for two months concerning developments in the pharmaceutical industry

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

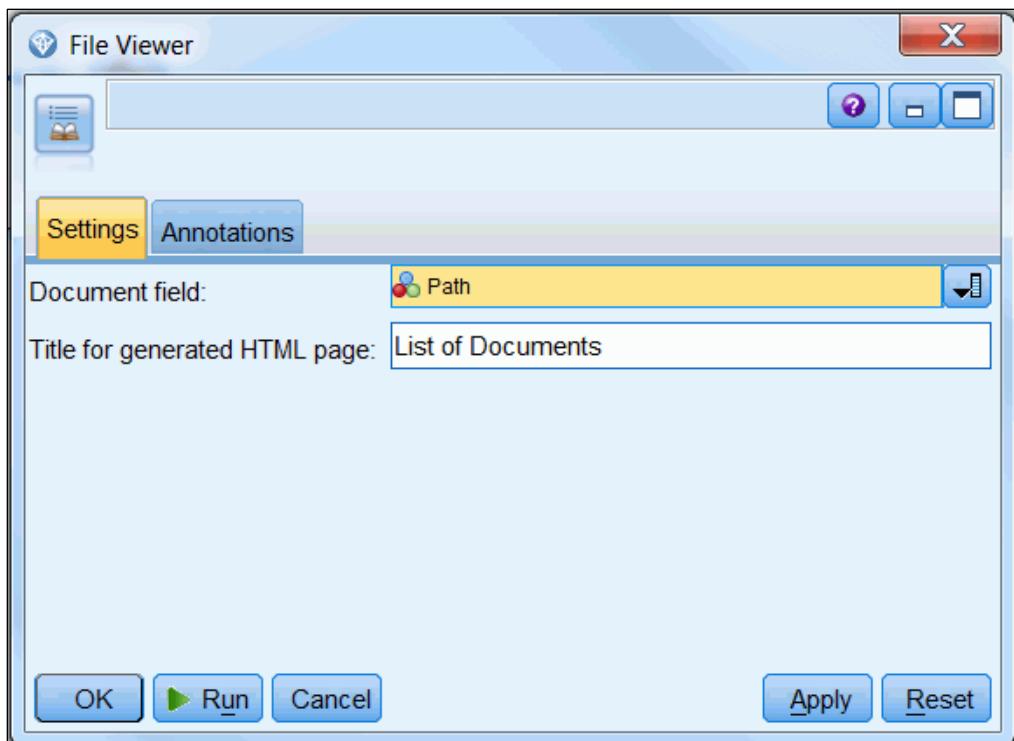
Demo 2: Using the File Viewer Node to View Documents in Modeler

Purpose:

Prior to text mining a set of documents, you should review the contents of the documents. It is often easier to understand the extraction process if you can review them within Modeler itself, rather than having to open them in another software application.

Task 1. Setting up the File Viewer Node.

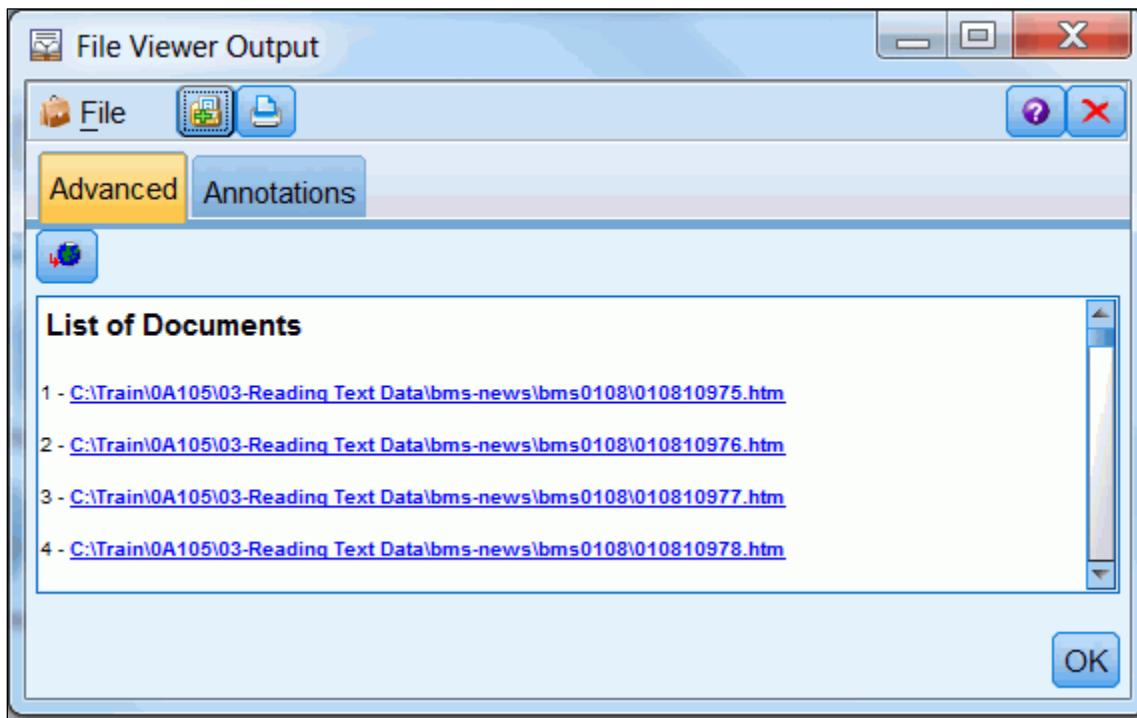
1. Add a **File Viewer** node to the stream created in Demo 1.
2. Connect the **File List** node to the **File Viewer** node, and then edit the **File Viewer** node.
3. Beside **Document field**, select **Path**.



4. Click **Run**.

Task 2. Viewing the documents.

When the node runs, a window opens with a hypertext link to each document. For the news service data, there are 1055 links for the 1055 files.



1. Click the **Launch** button  to start your external browser.

- After the browser opens, click the link to document number 1.

The screenshot shows a Mozilla Firefox browser window with the title bar "Poliovirus vectors may be useful for HIV vaccine - Mozilla Firefox: IBM Edition". The address bar shows the URL "file:///C:/Train/0A105/03-Reading Text Data/bms-news/bms0108/010810975.htm". The main content area displays a news article from Reuters Medical News dated August 10, 2001. The article discusses a study where simian immunodeficiency virus (SIV) vaccine based on Sabin 1 and 2 strains was used to protect macaques from SIV infection. It notes that four of six vaccinated monkeys were protected, while all control monkeys became infected. The text is presented in a standard web page layout with headings and paragraphs.

In this matter, you can read from within Modeler any document that you want from the original document list.

- Close the **Browser** window and the **Text Report** window.
- From the **File** menu, click **Close Stream**, and then click **No**.
- From the **File** menu, click **New Stream**.

Do not close Modeler. Leave it open for the next demo.

Results:

You have successfully reviewed the documents that you want to text mine from within Modeler, rather than having to open them in another software application.

Web Feed Node

- Used to read text from Web feeds, such as blogs or news feeds
- Accepts both RSS or HTML format Web feeds
- Prepares text data from Web feeds for the text mining process

© 2014 IBM Corporation



There is much interest today in blogs and Web feeds, and these sources contain lots of text data that can be mined. A blog is a website that typically provides commentary on one or more subjects, including politics, technology, sports, or hobbies. The blog entries are displayed in reverse chronological order, for example, the most recent at the top, not the bottom. Web feeds are a format used to automatically send content to subscribers. Content distributors, typically organizations, syndicate a Web feed. This makes it possible for people to subscribe to the Web feed. The stories or content are also typically in reverse chronological order.

Blogs are commonly in HTML format. Web feeds are usually in RSS (really simple syndication) or HTML format. Modeler can read blogs and feeds in RSS and HTML formats with the Web Feed node. There are differences depending on whether the format is RSS or HTML.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Web Feed Node – RSS Format

- RSS is a simple XML-based standardized format for Web content such as syndicated news sources and blogs
- Each linked article is automatically identified and treated as a separate record in the resulting data stream.
- RSS feeds have a known structure so it is unnecessary to define tags to identify the content fields

© 2014 IBM Corporation



The correct URL for the feeds you want to read is entered or pasted into the text box. You can enter multiple URLs. For RSS feeds, the URL will point to a page that has a set of linked articles. Each article can be automatically identified and treated as a separate record in the data, somewhat similar to the way a File List node can make each document a separate record. For HTML feeds, you will need to define the start tag for each record on a page (such as, each blog entry), as well as other delimiters identifying such things as date and the main content.

Fortunately, RSS feeds are XML-based and have a simple, standardized format. For an RSS feed, normally all you have to do is paste the correct URL from a browser. No further specifications will be required for Modeler to identify the records and text data. The option of Number of most recent entries to read per URL, set at 500 by default, controls the number of records to read (news stories for an RSS feed).

The option to Save and reuse previous Web feeds when possible tells Modeler to scan the feed and cache the processed results. Then, upon subsequent stream executions, Modeler can check to see whether the feed contents have been updated. If the contents of a given feed have not changed or if the feed is inaccessible (no connection to the Internet), the cached version is used to speed processing time. If you use this option, you also need to provide a label.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

3-19

Web Feed Node – HTML Format

- Harder to read from than RSS feeds because you must define the tags that delimit the elements of the HTML content
- These are the tags that will be output by the node
- The Description field is most commonly used since it contains the bulk of the text content
- Title and Short Description fields are also commonly used in text mining

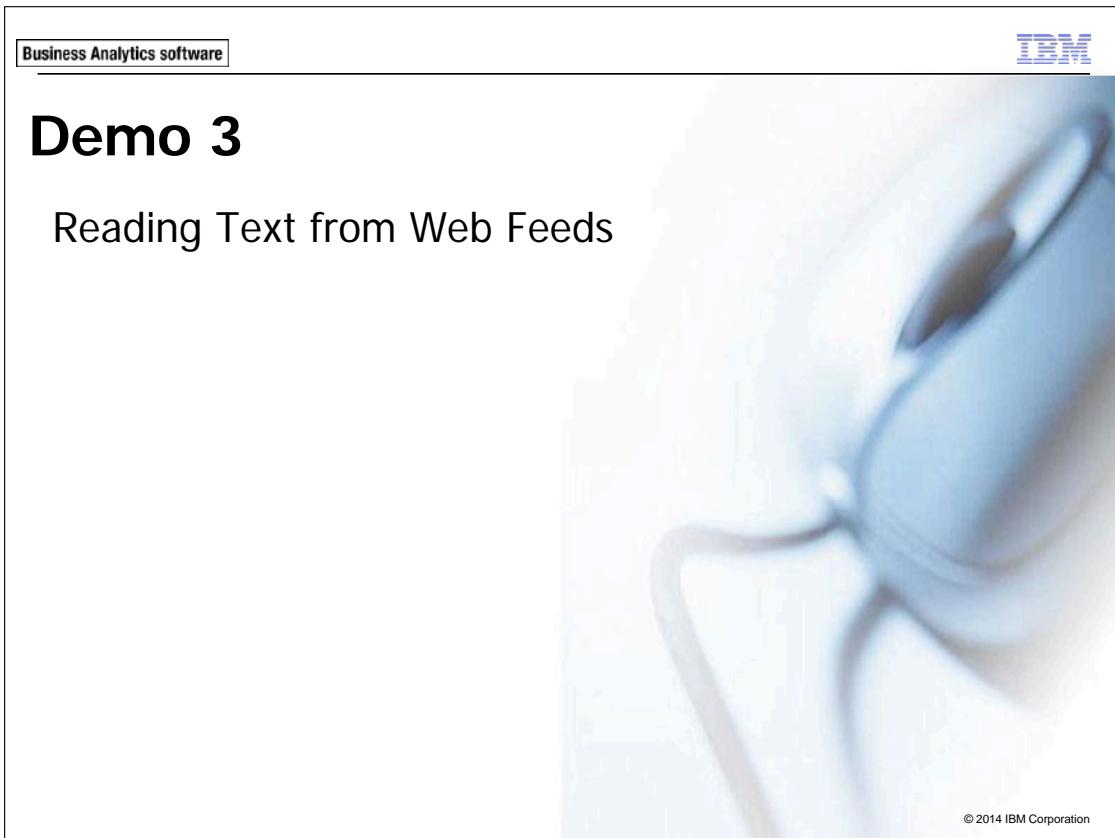
© 2014 IBM Corporation



The Records tab is used to define the HTML tags that will be used to process the input. Tabs must be defined for each HTML feed, but because an RSS feed has a known structure, the fields need no tags. These are the fields that will be output by the node. The Description field is generally the most commonly used since it contains most of the text from an RSS feed. Other fields often used for text mining are the Title and Short Description.

You can preview the Web feed by selecting a URL and clicking Preview. The Preview window lists each field, and then any text corresponding to that field. You can scroll down to the Description field to see the article itself. After the first record/article is complete, the next article begins.

The Content Filter tab allows some data cleansing from the feed. The options will not be covered in detail, but generally the filter allows removing of short lines of text, discarding certain tags and discarding lines with specific text (such as copyright or header lines).



The slide is titled "Demo 3" and subtitle "Reading Text from Web Feeds". It features the IBM logo in the top right corner and the text "Business Analytics software" in a small box at the top left. A large, blurry image of a person's face is visible in the background. At the bottom right, there is a small copyright notice: "© 2014 IBM Corporation".

Note: your computer will need to be connected to the Internet to perform this demonstration.

This demo uses the following two streams:

- C:\Train\0A105\03-Reading Text Data\Reading_Text_Data_Demo3_RSS_start.str
- C:\Train\0A105\03-Reading Text Data\Reading_Text_Data_Demo3_HTML_start.str.

The correct settings for the URLs you will be reading have already been entered in the Web Feed nodes in each stream.

Demo 3: Reading Text from Web Feeds

Purpose:

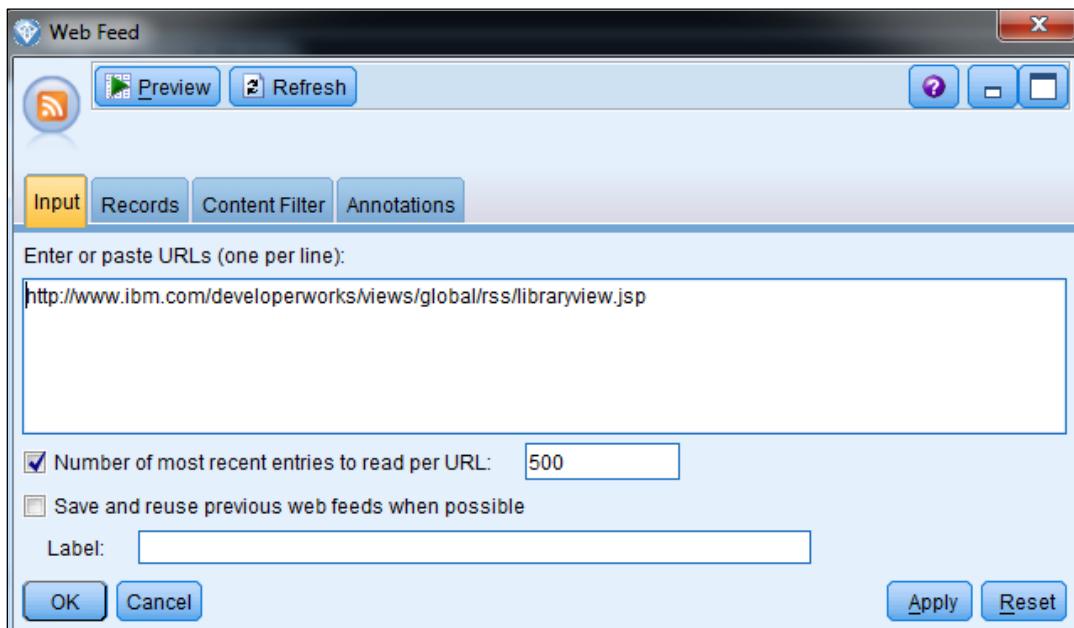
It is often extremely useful to text mine documents on the Web. For example, you may want to see what people who post on the Web are saying about your company and your competitors. You will read Web Feeds using IBM SPSS Text Analytics.

Task 1. Reading Web feeds in RSS format.

In this task, you will be reading an RSS feed from IBM developerWorks: Technical Library.

1. From the **File** menu, click **Open Stream**.
2. Navigate to the **C:\Train\0A105\03-Reading Text Data** folder, and then double-click **Reading_Text_Data_Demo3_RSS_start.str**.
3. Edit the **Web Feed** node.

The URL you will be reading has already been entered in the Input tab.



4. Click **OK**, and then add a **Table** node to the stream.

5. Connect the **Web Feed** node to the **Table** node, and then run the **Table** node. The results appear as follows:

Table (7 fields, 100 records)	
File Edit Generate	
Table	Annotations
Title	Short Description
1 Use code coverage tools in Rational Developer for i	Code coverage supports the common IBM i compiled languages such as RPG, COBOL, C, C++, and C#. Do you want to build a dynamic website with Node.js but are unsure where to start?
2 Build your first Node.js website. Part 3	Do you want to build a dynamic website with Node.js but are unsure where to start?
3 Build your first Node.js website. Part 2	Do you want to build a dynamic website with Node.js but are unsure where to start?
4 Build your first Node.js website. Part 1	Do you want to build a dynamic website with Node.js but are unsure where to start? This third part covers the basics of setting up a Node.js environment.
5 Processing and content analysis of various document types using MapReduce and InfoSphere BigInsights	Businesses often need to analyze large numbers of documents of various file types. In this four-part video tutorial, Jose Bravo demonstrates how to use QRadar Forensics to investigate IT security incidents with QRadar Forensics.
6 IBM Tivoli Provisioning Manager Express for Software Distribution V4.1	Download Tivoli Provisioning Manager Express for Software Distribution V4.1. Tivoli Provisioning Manager Express (TPE) is a component of the Tivoli Provisioning Manager suite.
7 IBM Bluemix includes a robust set of SDKs to interact between mobile devices and cloud services. Download a free trial of Tivoli Endpoint Manager V8.2, which includes four modules: Endpoint Management, Device Management, Application Management, and Security.	IBM Bluemix includes a robust set of SDKs to interact between mobile devices and cloud services. Download a free trial of Tivoli Endpoint Manager V8.2, which includes four modules: Endpoint Management, Device Management, Application Management, and Security.
8 Send a Push notification from a mobile app in 5 minutes or less	In this article you will find useful information for keeping WebSphere MQ channels secure. BLU Acceleration is a collection of technologies for analytic queries that was introduced in WebSphere MQ 8.0. Real-time detection of risks means that you can manage security vulnerabilities and threats proactively.
9 IBM Endpoint Manager	The database as a service (DBaaS) 1.1.0.8 component of the IBM PureApplication System. Download a free trial of IBM SmartCloud Provisioning, a true IaaS solution that provides a simple way to provision and manage virtual machines and storage resources.
10 Comparing BlockIOP2 with Channel Authentication Records for WebSphere MQ Security	Internet security is a major concern now-a-days. Internet Protocol Security (IPSec) is a security protocol that provides communication security between hosts that speak the IP protocol. Big SQL V3.0 supports federation to many data sources, including IBM DB2 for Linux, UNIX, and Windows.
11 DB2 monitoring enhancements for BLU Acceleration	Create this flight-tracking application that overlays the real-time location of flights as they move across the globe. Defects can be created automatically from an automation playback report. The list of failures is updated in real time.
12 Detecting security risks with IBM Security QRadar Vulnerability Manager	IBM Tivoli Service Automation Manager (TSAM) helps enable users to request, deploy, and manage software assets.
13 Use the DB2 with BLU Acceleration Pattern to easily deploy a database	IBM Tivoli Service Automation Manager (TSAM) helps enable users to request, deploy, and manage software assets.
14 IBM SmartCloud Provisioning	IBM Tivoli Service Automation Manager (TSAM) helps enable users to request, deploy, and manage software assets.
15 Exploring IKEv2 with ECDSA certificate in IBM AIX	IBM Tivoli Service Automation Manager (TSAM) helps enable users to request, deploy, and manage software assets.
16 Set up and use federation in InfoSphere BigInsights Big SQL V3.0	IBM Tivoli Service Automation Manager (TSAM) helps enable users to request, deploy, and manage software assets.
17 Follow air traffic with a Flight Status and Tracking app built on Bluemix	IBM Tivoli Service Automation Manager (TSAM) helps enable users to request, deploy, and manage software assets.
18 Create defects automatically from automation playback report	IBM Tivoli Service Automation Manager (TSAM) helps enable users to request, deploy, and manage software assets.
19 Explore new features in Tivoli Service Automation Manager Network Extension for Juniper	IBM Tivoli Service Automation Manager (TSAM) helps enable users to request, deploy, and manage software assets.

There are 100 records corresponding to the 100 new articles being pushed out by the IBM developerWorks Web feed. From this point, you can analyze the data with the Text Mining node, and you can analyze not just the description, but also the Title or other fields.

6. Close the **Table** node output.
 7. From the **File** menu, click **Close Stream**, and then click **No**.
 8. From the **File** menu, click **New Stream**.

Task 2. Reading Web feeds in HTML format.

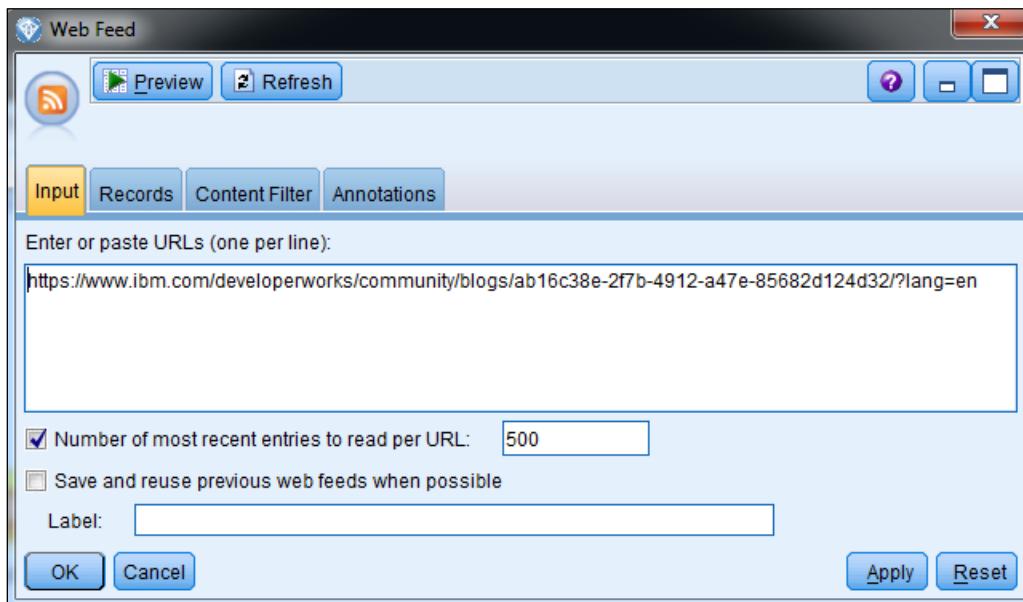
In this task, you will be reading from the SPSS Blog from IBM developerWorks.

1. From the **File** menu, click **Open Stream**.
 2. Double-click **Reading_Text_Data_Demo3_HTML_start.str**.

3. Edit the **Web Feed** node.

The URL you will be reading has already been entered in the Input tab.

The results appear as follows:



4. Click the **Records** tab.

5. Use the **URL** menu to select an address.

This drop-down list contains a list of URLs entered on the Input tab.

The screenshot shows a software interface with a top navigation bar containing tabs: Input, Records (which is highlighted in yellow), Content Filter, and Annotations. Below the navigation bar is a URL input field with the placeholder text '<Select an Address>'. A dropdown menu is open, showing two entries: '<Select an Address>' and a selected item, 'https://www.ibm.com/developerworks/community/...b16c38e-2f7b-4912-a47e-85682d124d32/?lang=en'. The URL is also displayed in a larger text area below the dropdown.

The specification for the "Non-RSS record start tag" signals the beginning of a record (such as an article or blog entry). If you do not define one for a non-RSS feed, the entire page is treated as one single record. In this example, the tag `<DIV class="entryContentContainer blogsWrapText" role="article">` signals the beginning of each blog article in this HTML Web feed.

The screenshot shows a 'Web Feed' configuration dialog. At the top, there's a toolbar with icons for Preview, Refresh, and Help. Below the toolbar is another set of tabs: Input, Records (highlighted in yellow), Content Filter, and Annotations. The URL is set to 'https://www.ibm.com/developerworks/community/...b16c38e-2f7b-4912-a47e-85682d124d32/?lang=en'. The 'Source' tab is selected, showing a 'Find:' input field and a 'Find' button. Below the tabs, a section titled 'Non-RSS record start tag:' contains the value '`<DIV class="entryContentContainer blogsWrapText" role="article">`'. A table labeled 'Output Fields' lists various fields and their corresponding start tags:

Output Fields	Start Tag For Content
Title	
Short Desc	
Description	<code><DIV class="entryContentContainer blogsWrapText" role="article"></code>
Author	
Contributors	
Published Date	
Modified Date	

At the bottom of the dialog are buttons for Clear All, Copy All, Paste All, OK, Cancel, Apply, and Reset.

To find the appropriate non-RSS record start tag, you may need to click the Source tab to view the html content, type a word such as "title" or "content" into the Find: box and use the Find button to help you locate the appropriate tag. The non-RSS record start tag may not be the same in every HTML formatted Web feed you are read. You could enter additional tags for Title, Author, etc. but you will skip that in this demonstration.

6. Click **OK**, and then add a **Table** node to the stream.
 7. Connect the **Web Feed** node to the **Table** node, and then run the **Table** node.
- The results appear as follows:

Title	Short Description	Description	Author	Contributors	Published Date	Modified Date
1		SPSS Statistics was, as far as I know, the first commercial software to deliver an integration with the R statistical language...			2014-07-30 00:02:11	2014-07-30 00:00
2		The SPSS Statistics Custom tables (CTABLES) procedure can test equality of column proportions and column means. T...			2014-07-30 00:02:11	2014-07-30 00:00
3		Suppose you want to generate all two-way crosstabs for a set of variables in your dataset but without creating any of the re...			2014-07-30 00:02:11	2014-07-30 00:00
4		The SPSS Statistics catalog of table types has over 500 entries (and all Custom Tables count as just one), but sometimes...			2014-07-30 00:02:11	2014-07-30 00:00
5		The SPSS Community website has a wealth of resources for users of SPSS Statistics and other SPSS products. In partic...			2014-07-30 00:02:11	2014-07-30 00:00
6		We are sometimes asked why we use Python as the main programmability and scripting language in IBM SPSS Statistics....			2014-07-30 00:02:11	2014-07-30 00:00
7		Case-control matching is a popular technique used to pair records in the "case" sample with similar records in a typically...			2014-07-30 00:02:11	2014-07-30 00:00
8		This is a quick note for people who produce a lot of tables with SPSS Statistics or use a lot of scripting code for formatting...			2014-07-30 00:02:11	2014-07-30 00:00
9		Users of traditional SPSS Statistics syntax are used to using the macro facility to parameterize blocks of syntax so that it...			2014-07-30 00:02:11	2014-07-30 00:00

8. Close the **Table** node output.
9. From the **File** menu, click **Edit** to end the Modeler session.

Results:

You have successfully been able to read Web feeds in both RSS and HTML format using Text Analytics for Modeler. The next step would probably be to connect a Text Mining node to the Web Feed node, so you could text mine the data, but that goes beyond the scope of this module.

Apply Your Knowledge

Purpose:

Test your knowledge of the material covered in this module.

Question 1: True or False: The File List node does not allow you to select specific files, only file types. Thus, if a folder includes files of the same type, only some of which you want to use in text mining, you will need to move the files that you want to use into a separate folder.

- A. True
- B. False

Question 2: True or False: When using the file List node, only the File Viewer node allows you direct access to the original data.

- A. True
- B. False

Question 3: True or False: When using the Web Feed node, only the File Viewer node allow you direct access to the original data.

- A. True
- B. False

Question 4: True or False: The Web Feed node allows you to use as many URLs as you want.

- A. True
- B. False

Question 5: True or False: You must define tags or records in order to read an RSS feed.

- A. True
- B. False

Apply Your Knowledge - Solutions

Answer 1: A. True

Answer 2: A. True

Answer 3: B. False

Answer 4: A. True

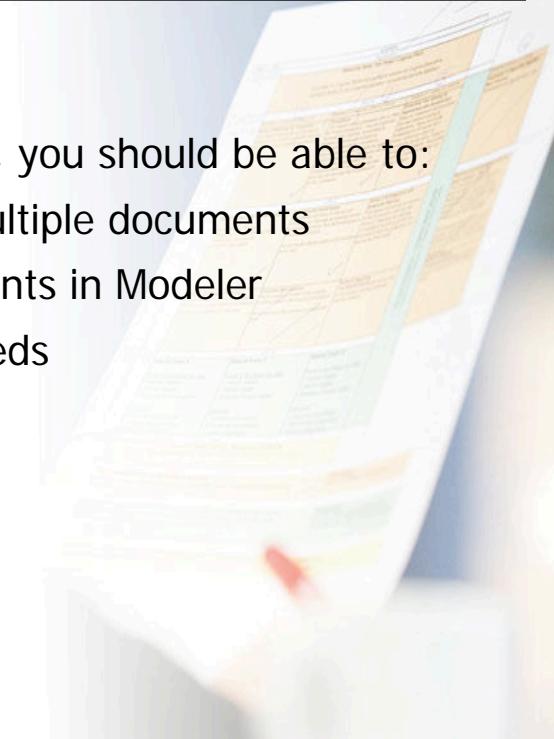
Answer 5: B. False

Business Analytics software

IBM

Summary

- At the end of this module, you should be able to:
 - read text data from multiple documents
 - view text from documents in Modeler
 - read text from Web feeds



© 2014 IBM Corporation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

3-29

Business Analytics software

IBM

Workshop 1

Text Mining Data from an RSS Feed



© 2014 IBM Corporation

This workshop uses the following file:

- C:\Train\0A105\03-Reading Text Data\Workshop_RSS_Feed.str

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Workshop 1: Text Mining Data from an RSS Feed

The Web often contains a vast amount of useful information companies would like to mine so they can analyze what customers are saying about them. Often this is in the form of blogs. In this workshop, you will be text mining an RSS Feed from the IBM developerWorks: Technical library.

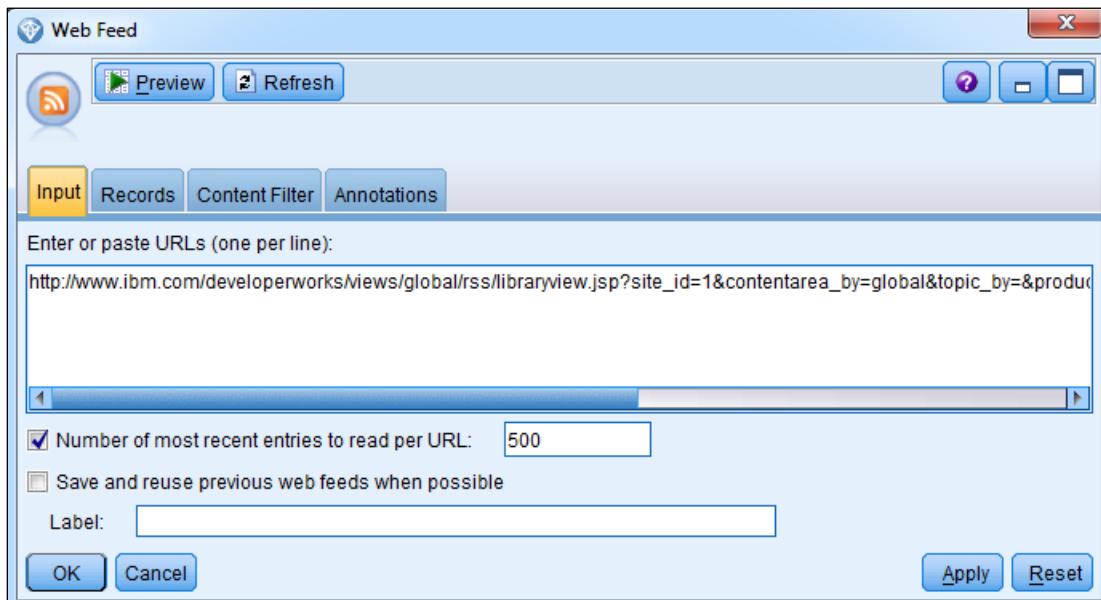
- Open Workshop_RSS_Feed.str.
- Edit the Web Feed node.
- Add a Table node downstream from the Web Feed node.
- Run the Table node and examine the results.
- Add a Text Mining node downstream from the Web Feed node.
- Edit the Text Mining node.
- Ensure that Actual text is checked in the Text Field represents option.
- Select the field Description in the Text field box.
- On the Model tab, select Generate directly (concept model nugget) in the Build mode section.
- Click Run.
- Edit the generated model named Description that has been placed on the stream canvas.

Workshop 1: Tasks and Results

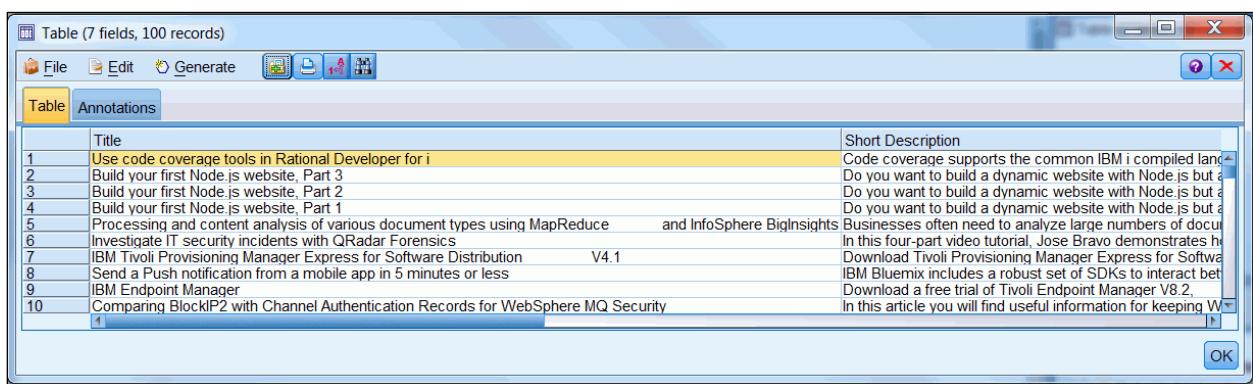
In this workshop, you will be accessing an RSS Feed from the IBM developerWorks : Technical library.

- Open the **Workshop_RSS_Feed.str** stream.
- Edit the **Web Feed** node.

The URL you will be reading has already been entered in the Input tab.



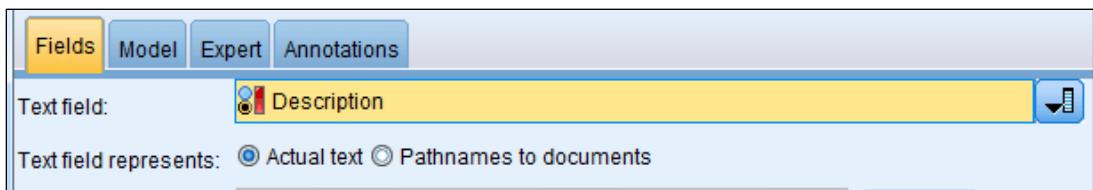
- Add a **Table** node downstream from the **Web Feed** node.
- Run the **Table** node, and examine the results.



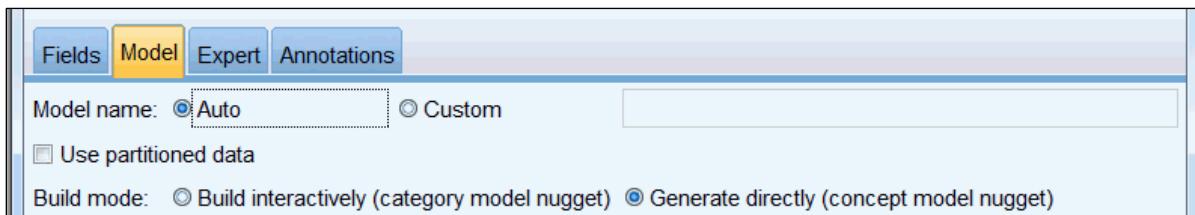
In this window you can see some of the Titles and Short Descriptions from the RSS feed.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

- Add a **Text Mining** node downstream from the **Web Feed**.
- Edit the **Text Mining** node.
- On the **Fields** tab, ensure that **Actual text** is selected.
- Beside **Text field**, select **Description**.



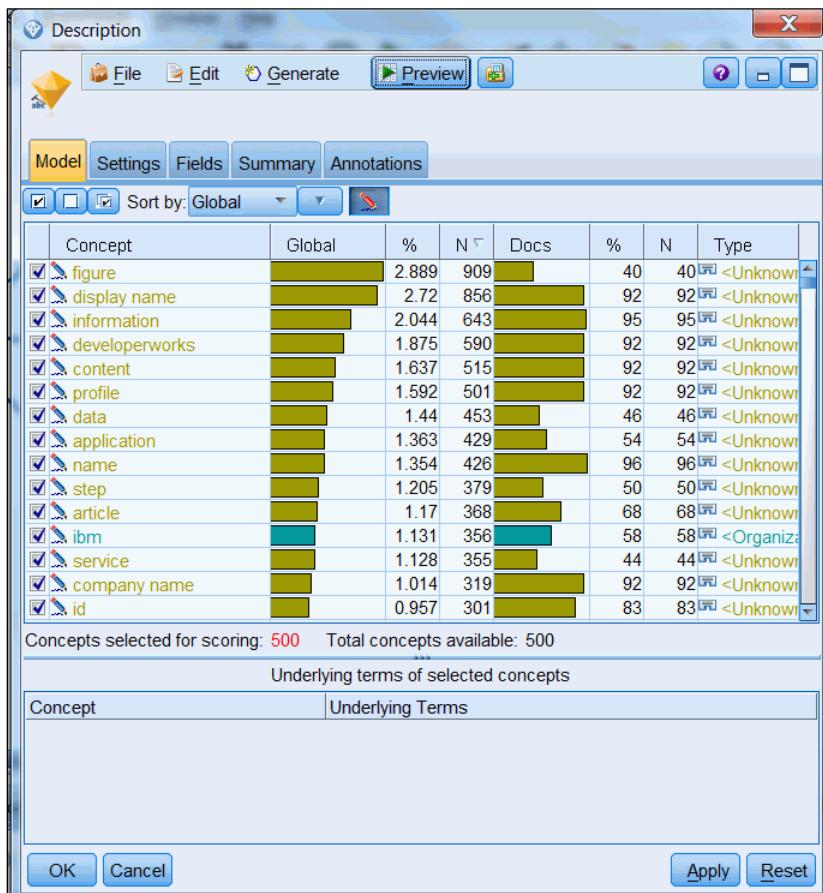
- On the **Model** tab, select **Generate directly (concept model nugget)**.



- Click **Run**.

- Edit the **generated model** named **Description** that has been placed on the stream canvas.

The results appear as follows:



- Exit from Modeler without saving.



Linguistic Analysis and Text Mining

IBM SPSS Modeler Text Analytics (v16)



Business Analytics software

© 2014 IBM Corporation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Objectives

- At the end of this module, you should be able to:
 - describe linguistic analysis in general
 - describe the process of text extraction
 - describe categorization of terms and concepts
 - describe templates and libraries
 - describe Text Analytics Packages

© 2014 IBM Corporation

This module will look in detail at how concepts are extracted from text in Modeler. Although every aspect of text mining is crucial, if the necessary text is not extracted, then appropriate categories cannot be created. Moreover, any analysis done using the extracted results, such as text link analysis, will be incomplete.

It is important that you understand the stages in extraction, the options available and the capabilities, and limits, of the extraction algorithms. Knowing this will make you more efficient at editing the linguistic resources so the right text is extracted. Part of extraction is making sure that the terms that you deem to be important are indeed extracted, but another part is grouping the extracted results with synonym and type definitions to simplify and consolidate the extracted results.

Extraction will be an iterative process, as you make changes to the linguistic resources, re-extract the text, review the results, make more changes to the resources, etc.

Because of this, and especially with large data files, you will most likely want to work with a sample of the data rather than the full data set. You must sample so that you have a full range of possible text variations and terms in the sample data, for obvious reasons. Even more so than with statistical models, if a term is missing from the sample, there is no guarantee that it will be extracted from the full data file.

Linguistic Analysis

- Linguistic analysis involves the study of the elements, structure, and meaning of language:
 - Morphology
 - Syntax
 - Semantics

© 2014 IBM Corporation



Text mining must take into account the universal fact that languages contain ambiguities. The same words can be different parts of speech (nouns, pronouns, verbs, adjectives, adverbs, etc.), and therefore play different roles in meaning. The same word, even when used as the same part of speech, can have different meanings depending on how it is used and the context within which it appears. In English:

- George could back (verb) his car into the back (noun) of his garage.
- George could also hurt his back (noun) getting out of his car.
- George could then relax in his back (adjective) garden to recuperate.

Linguistic analysis involves the study of the elements, structure, and meaning of language:

- Morphology: The study of a language's smallest grammatical units and the ways in which they combine into words. The word "cats" consists of two elements, or morphemes: "cat" means "feline animal" and "s" means "more than one." Prearrangement can be divided into the morphemes pre- (prior), arrange (to prepare), and "ment", a suffix that turns the verb into a noun.

- Syntax: The study of how words combine to make sentences. The order of words in sentences varies from language to language. English syntax generally follows a subject-verb-object order, as in the sentence The dog (subject) bit (verb) the man (object). The sentence The dog the man bit is not a sentence in English, and the sentence The man bit the dog has a very different meaning. In contrast, Japanese has a basic word order of subject-object-verb, as in watashi-wa hon-o kau, which literally translates to I book buy.
- Semantics: The study of the meaning of words, phrases, sentences, and texts. The best examples are synonyms and homonyms. For example to run (jog, move fast) versus to run (guide, manage) versus to run (ooze). Semantic analysis is the most difficult task for text mining, and involves in part the use of dictionaries, thesauruses, glossaries, lexicons, typologies, and so forth.

In summary, a linguistic approach to text mining does not treat text as merely a collection of words. IBM SPSS Modeler Text Analytics is able to recognize parts of speech within the clause or sentence structure, and extract compound words, phrases and idioms that would typically be treated as individual words by other approaches.

You can affect how extraction processes text, but you do so indirectly through editing the libraries and their associated dictionaries that are supplied with the program, or that you create (in addition to the settings in the Text Mining modeling node).

Parts of Speech

- Each part of speech not only explains what the word is, but how the word is used
- Traditional English grammar classifies words based on eight parts of speech
- Part of Speech (PoS) tagging in IBM SPSS Text Analytics using eleven tags

© 2014 IBM Corporation



Each part of speech explains not what the word is, but how the word is used. Traditional English grammar classifies words based on eight parts of speech: verb, noun, pronoun, adjective, adverb, preposition, conjunction, and interjection. However, part-of-speech (PoS) tagging in IBM SPSS Text Analytics uses the following tags:

- N: Noun. A word used to name a person, place, thing, quality, or action and can function as the subject or object of a verb.
- V: Verb. A word that expresses existence, action, or occurrence such as be, run, and conceive.
- A: Adjective. A word used to modify a noun by limiting, qualifying, or specifying it.
- B: Adverb. A word that modifies a verb, an adjective, or another adverb.
- O: Coordination. A conjunction such as "and" and "or".
- D: Determiner. A noun modifier including articles, demonstratives, possessive adjectives, and words such as any, both, or whose.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

- G: Gerund. A noun derived from a verb. In English ending in the suffix "ing", as in we admired his singing.
- P: Participle. A verb used as an adjective, most often ending in "ing" (present) or "ed" (past), as in jumping jack, and opened door.
- C: Preposition. A word placed before a noun that indicates the relation of that noun to a verb, an adjective, or another noun. For example, the word "of" is a preposition. Most prepositions are tagged as S or Stop words.
- X: Auxiliary. A verb such as is, have, can, could, or will that usually accompanies a main verb in a clause.
- S: Stop word. A very large category of words used to exclude from extraction. It contains all pronouns, particles, and prepositions (except "of")

For example, tagging for the phrase used in an earlier module as follows:

Innovative	solutions	from	SPSS	Inc.	enable	your	organization	to	both
A	N	X	?	N	V	X	N	X	X

uncover	concepts	hidden	in	text	and	use	them	to	predict
V	N	A	X	N	O	V	X	X	V

future	conditions	behavior	and	trends
NA	NV	N	O	NV

Results of derived terms in the sentence are in bold:

Innovative solutions from SPSS Inc. enable your **organization** to both uncover **concepts** hidden in **text** and use them to predict **future conditions**, **behavior**, and **trends**.

These tags are used during extraction to identify the concepts:

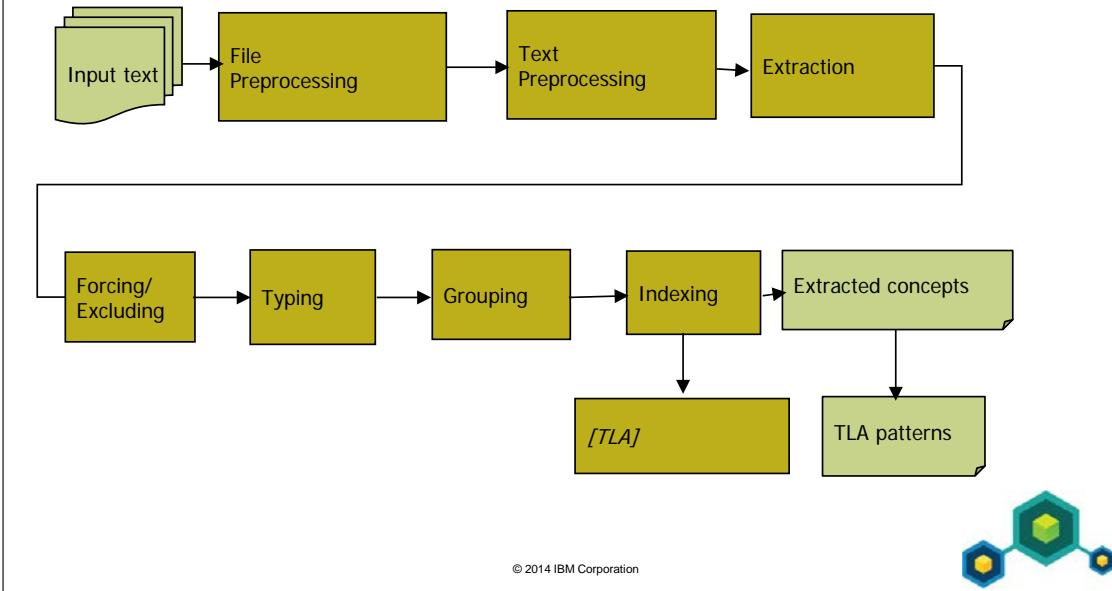
- Words "from", "your", "to", "both", "in", and "them" are eliminated as Stop words.
- Words "solutions", "inc.", "organization", "concepts", "text", and "behavior" are identified as nouns.
- Word "future" is tagged as noun and adjective. If a word is a noun and something else, noun is preferred.
- Words "conditions" and "trends" are nouns and verbs, but are identified as nouns for the same reason.
- Any word typed with a question mark (?) means that the word is absent from the dictionary and is thus tagged as a noun. *SPSS* is such a word.

Therefore:

- Pattern AN generates "innovative solutions" as a concept
- Pattern NN generates "spss inc." and "future conditions" as concepts
- Patterns N are uniterms and generate "organization", "concepts", "text", "behavior", and "trends".

As you can see, default extraction patterns are noun-oriented (noun noun, adjective noun, noun of noun, noun noun noun, and so forth). You can write your own patterns for the specific needs of an application. For example, you might want the extractor to be more predicate-oriented (everything in a sentence that might modify the subject), perhaps emphasizing verbs as in uncover and predict in the above example. Several of these topic-specific pattern resources are already available in the additional distributed libraries.

Extractor Component Workflow



The first step in text mining with Modeler is to run the imported data through the extraction engine. The extraction component workflow is outlined above.

Understanding how the extraction process works can help you make key decisions when fine-tuning your linguistic resources (libraries, types, synonyms, and more).

Text Preprocessing

- The extraction process begins after Modeler is able to:
 - detect the encoding of the data
 - filter and clean the text
 - determine the type of text - ASCII structured, HTML, XML, unstructured, Microsoft Word, etc.

© 2014 IBM Corporation



The process begins after a file is readied in the file preprocessing stage by detecting the encoding of the data and filtering and cleaning the text before sending it to the parts-of-speech tagger. Text preprocessing determines the type of text-ASCII structured, HTML, XML, unstructured, or Word, etc., and then cleaning it by deleting of some tags from HTML/XML code, replacement of some special characters with spaces, and determination of the end of sentences and end of paragraphs.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

4-10

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Identification of Candidate Terms

- Text tokenization
- Text normalization
- Candidate term extraction
- Part of speech tagging

© 2014 IBM Corporation



The extraction step begins by tagging each word and makes a first attempt at identification of candidate terms. Recall that terms are the building blocks of a text analysis, corresponding to single or compound words (usually noun phrases) that are deemed relevant or interesting. In certain cases, as with structured text, linguistic processing is not required to identify candidate terms for extraction. During this step, the following actions are taken:

- Text tokenization: A process that identifies character strings (tokens) from the input text, based on delimiters. Examples of delimiters are spaces, tabs, carriage returns, and punctuation marks.
- Text normalization: Helps manage poor punctuation in the text, such as improper use of a period, comma, semi-colon, colon, forward-slash, etc. The input text is "corrected" internally to place spaces around improper punctuation.
- Candidate term extraction: Identifies relevant words and compound words from the input text. An example of a compound word is "sports car".

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

4-11

- Part of speech tagging: Each token in the text stream is tagged with a Part of Speech (PoS) tag, which comes from the base dictionary. Upper-case unknown words are tagged as nouns. The extractor gives preference to a noun tag in the case of multiple choices.

Verbs are not extracted by themselves (although you can force their extraction except with nouns in compound terms, as in buying milk, which requires the use of advanced capabilities). Verbs are used to make sense of the meaning of sentences.

During the stage of identification of candidate terms, target terms are substituted for synonyms. For example, you might want to associate drug and pill with medicine so that medicine is the term extracted (the target), and drug and pill are associated (as synonyms) with medicine whenever they are encountered.

Note that terms are all extracted (and displayed) in lower case, which means that Medicine and MEDICINE are both medicine to the extractor.

Identification of Equivalence Classes

- Variants in separators
- Permuted components
- Comparison of inflected forms of words
- Omission of optional elements
- Fuzzy grouping
- Geographic variant

© 2014 IBM Corporation



After candidate terms are identified, the extractor uses a set of built-in algorithms to compare extracts and identify equivalence classes to ensure, for example, that "cancer of the thyroid" and "thyroid cancer" are not treated as separate terms. These processes include:

- Variants in separators: "stress free", "stressfree", and "stress-free" are extracted as "stress-free".
- Permuted components: Officials of the companies and "company officials" become "company officials". The lead term will be, in order:
 - User specified synonym
 - The most frequent form of the term
 - The shortest form of the term
 - The first one that is encountered

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

- Comparison of inflected forms of words: "American consumer" is equivalent to "American consumers". In general, inflection refers to the change of form in words to mark gender, number, or tense, and also includes prefixes and suffixes.
- Omission of optional elements: "Johnson and Johnson" is equal to "Johnson and Johnson Inc".
- Fuzzy grouping: Identifies spelling variants by removing vowels and double (or triple) consonants and then performing a comparison. However, if each term is assigned to a different type, excluding the <Unknown> type, the fuzzy grouping technique will not be applied.
 - **technical support** = **technical support** → tchnclsprt
 - **furniture** = **furniture** → frntr
 - **adidas** = **adidas** → add
- Geographic variant: "Color" and "colour" are extracted together as "color".

Forcing / Excluding

- Additional extraction can be performed by forcing words or phrases to be extracted that are not listed in the extracted results pane.
- Alternatively, you can prevent or exclude concepts from extraction that have no significance in your analysis.

© 2014 IBM Corporation



When reviewing the text data in the Data pane after extraction, you may discover that some words or phrases were not extracted. While, normally these words are verbs or adjectives that you are not interested in, sometimes you may want to use a word or phrase that was not extracted as part of a category definition. These may be terms that are specific to your industry as product names, acronyms, etc.

If you would like to have these words and phrases extracted, you can force a term into a type library. However, it is important to note that this method does not always work. Even though you have explicitly added a term to a dictionary, there are times when it may not be present in the Extraction Results pane after you have re-extracted or it does appear but not exactly as you have declared it. Although this rarely happens it can occur when a word or phrase was already extracted as part of a longer phrase. To prevent this, apply the Entire (no compounds) match option to this term in the type dictionary.

When reviewing your results, you may occasionally find concepts that you did not want extracted or used by any automated category building techniques. In some cases, these concepts have a very high frequency count and are insignificant to your analysis. In this case, you can mark a concept to be excluded from the final extraction. Typically, the concepts you add to this list are fill-in words or phrases used in the text for continuity but that do not add anything important and may clutter the extraction results. By adding concepts to the exclude dictionary, you can make sure that they are never extracted. By excluding concepts, all variations of the excluded concept disappear from your extraction results the next time that you extract.

Assigning of Types

- Types are assigned to extracted concepts.
- User-defined and compiled dictionaries are used to assign a semantic type to extracted terms.
- If an extracted term cannot be typed by one of the dictionaries, then it is typed as "Unknown" (U).
- Types are often used to force the extraction of terms or phrases that were not extracted by default.

© 2014 IBM Corporation



A type is a higher-level concept that contains one or more terms. There are default types for Organization, Product, Person, and Location in the Core English library. In the libraries associated with the Opinions (English) template, there are types for Budget, containing terms about prices, costs, and budgets. For example, there are a number of negative types containing many terms that are negative on their own, such as difficult or awful. There are corresponding positive types as well in this template.

After extracting and modifying terms as described above, the extractor next assigns each concept to a type, wherever possible. If an extracted concept cannot be matched to an existing type, it is given an Unknown type. In this way, no information is discarded at the extraction phase. Various types are searched for the specific term, and that term is "typed" under the type name if one is found. Both compiled resources and default type dictionaries are supplied with the product. You cannot view the compiled resources.

It is important to keep in mind that the compiled resources contain terms not listed in the default dictionaries. As an illustration, the number of organizations that can be typed automatically is much greater than is listed in the Organization dictionary because the compiled resources contain other organizations.

Categorizing Extracted Concepts

- Categories refers to a group of closely related ideas and patterns to which documents and records are assigned through a scoring process.
- There are two different ways to categorize the extracted concepts Text Analytics:
 - According linguistic techniques of classification such as concept inclusion, concept derivation or semantic networks
 - Manually

© 2014 IBM Corporation



The next step in text mining is normally the creation of categories, which represent the information that you consider to be important in the responses. The terms, types (and patterns) serve as the building blocks for the categories. There are three automatic categorization methods: concept inclusion, concept derivation, and semantic networks. You can also manually create categories from the extracted results.

Each technique is well suited for certain types of data and situations, but it is often helpful to combine techniques in the same analysis to capture the full range of responses. In addition, in the course of categorization, you will usually see other changes to make to the linguistic resources.

So what is a category? As used in text mining, it refers to a group of closely related concepts, opinions, or attitudes. Thus, if you are analyzing responses from consumers about a new laundry soap, you might construct a category labeled odor that contains all the responses describing the smell of the product. Such a category would not differentiate between those who found the smell pleasing and those who found it offensive or too strong. If you do want to capture the distinction, you could create two other categories to identify respondents who enjoyed the odor and disliked the odor.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Whether you create these additional categories depends entirely on the purposes of the analysis. If you want to capture the basic categories in a set of responses, the odor category is sufficient. But when you want to capture positive and negative opinions, you need to create different categories (and then may decide that the overall category is not needed).

Each category is defined by one or more descriptors, which are concepts, types, and patterns as well as conditional rules (presented later). Categories may be created from only a single concept or type, but it is common to combine several descriptors to create a category, with methods that take into account their root meaning and the relationship-both logical and linguistic-between sets of similar objects or opinions.

The three linguistic-based methods mentioned earlier provide a natural language-based approach to categorization. A fourth technique uses co-occurrence rules to discover and group concepts that are strongly related. A fifth method takes the top n number of types by frequency and turns them into categories.

None of the automatic techniques will perfectly categorize the text data. After applying one or more of these techniques, and reviewing the categories, you will invariably use manual techniques to refine the categorization, including merging and deleting categories, or creating new categories to which you then add terms. As with all data mining, the value of the results is dependent on the effort and time that you devote to the job.

Using Templates and Libraries

- A Library is a collection of dictionary resources, such as terms with types, synonyms, exclusions and so on
- A Resource template is a set of libraries
- There are a number of templates and libraries supplied with the Resource Editor
- Resource templates shipped with IBM SPSS Modeler Text Analytics have been fine-tuned for particular application areas

© 2014 IBM Corporation



IBM SPSS Modeler Text Analytics is shipped with a set of specialized resource templates that are designed for specific application areas. These templates include libraries, compiled resources, and some advanced linguistic resources. Libraries are comprised of dictionaries for types, terms, synonyms, optional elements, and excluded terms.

The templates are tuned for such areas as customer opinion, genomics, security intelligence, and CRM. Note that the compiled resources in any template, which are usually quite extensive, cannot be edited by the user.

You can load a template for an analysis, and you can load separate published libraries. It is important to understand that, when the software is first installed, the separate libraries are identical in content to the same library inside a template, but the templates have some advanced resources that offer more fine-tuning for a particular application, especially with text link analysis patterns. Thus, it is recommended working with the resources from a template as much as possible.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

When you load a template, a copy of its resources is stored in the text mining node, based on the state of the template at the moment that the template is accessed. However, the template itself is not linked to the node, which means that if the template is updated, changes are not reflected in the node. You can reload the template or make changes directly in the Resource Editor to ensure that the most recent resources are available.

When you make changes to the linguistic resources and want to reuse them in the future, on this data or new data, you can save the resources as a template. You can save to an existing template or to a new one. Templates and libraries can also be shared with other users.

After you have extracted terms and concepts, you will invariably make modifications to the resources. And then after the first attempt at categorization, you will probably make additional modifications to the resources. This will certainly be true when you are working with a data set for the first time, and as mentioned at the beginning of the module, when you use samples of data. Because of typically large file sizes in data mining, samples are often used to refine the linguistic resources before creating a final text mining model, and the sample should have a full range of textual variation so that the linguistic resources can be fully tuned.

Text Analysis Packages

- A Text Analysis Package (TAP) includes a template AND categories (at least, one)
- IBM SPSS Modeler Text Analytics is shipped with pre-built TAPs for satisfaction and CRM analysis (English).
- Shipped TAPs are especially useful when :
 - your business goal is to create a category model,
 - your field of application is covered by one of the shipped TAPs (opinions or CRM),
 - you do not have any predefined categories to import.

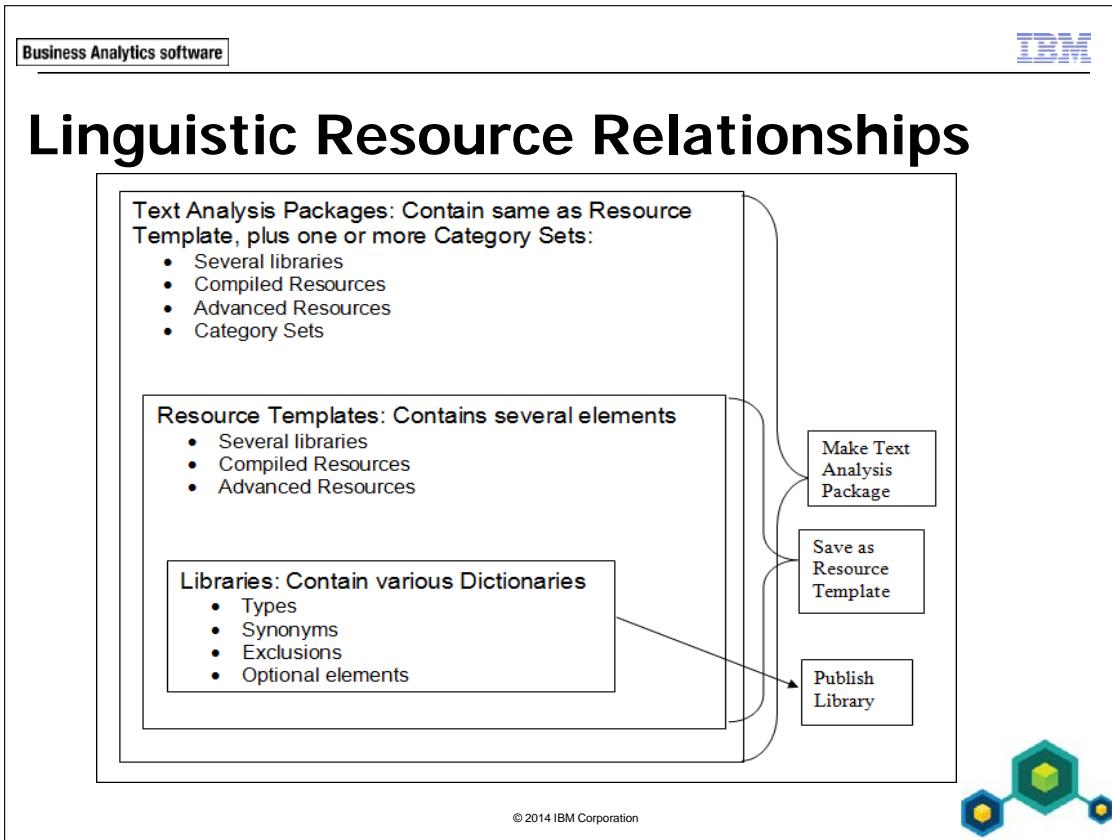
© 2014 IBM Corporation



IBM SPSS Modeler Text Analytics comes with an additional resource to aid in text categorization, a Text Analysis Package (TAP). A TAP contains the linguistic resources in a resource template in addition to one or more category sets, which are predefined categories (category names and code values), plus the set of descriptors for each category-concepts, patterns, rules, etc.-that will code responses into these categories from extracted results.

A TAP can be thought of as a quick way to build up categories as its template is the basis for all categorical extractions. IBM SPSS Modeler Text Analytics comes with several pre-packaged TAPs. The TAPs contain category sets which generally contain sentiment (such as positive or negative opinions) allowing quicker categorization of the text data.

A visual depiction of the relationship between libraries, resource templates, and text analysis packages is displayed below. Libraries are contained within Resource Templates, and TAPs contain the equivalent of a Resource Template, plus one or more category sets. Each of these elements can be saved separately.



If you have a finished project with linguistic resources and categories that you would like to save so they can be used on future text data, you can make a TAP from the project contents.

The TAPs are stored in individual files. This means that you can send a TAP you created to other users, who can then place it in the TAP folder on their computer (this is where the software looks for TAPs), and use it for new projects.

Apply Your Knowledge

Purpose:**Test your knowledge of the material covered in this module.**

Question 1: True or False: If the terms "arm" and "armm" were given the same type, fuzzy groupings could group these terms together.

- A. True
- B. False

Question 2: Which of the following processes of equivalence classes would help to ensure that the terms SPSS and SPSS Inc are not treated as separate terms?

- A. Comparison of inflected forms of words
- B. Omission of optional elements
- C. Fuzzy grouping
- D. Geographic variant

Question 3: True or False: Part of speech (PoS) tagging gives preferences to nouns.

- A. True
- B. False

Question 4: True or False: If the terms "riddle" and "ridde" were given the same type, fuzzy groupings could group these terms together.

- A. True
- B. False

Question 5: True or False: Text normalization helps manage poor punctuation.

- A. True
- B. False

Question 6: Which of the following processes of equivalence classes would help to ensure that the terms "behavior" and "behavior" are not treated as separate terms?

- A. Comparison of inflected forms of words
- B. Omission of optional elements
- C. Fuzzy grouping
- D. Geographic variant

Question 7: Which of these elements is contained in a Text Analytics Package but not in a Template?

- A. Types
- B. Synonyms
- C. Categories
- D. Excluded terms

Question 8: True or False: Verbs and adjectives are not automatically extracted from the text.

- A. True
- B. False

Question 9: Which of the following statements is true:

- A. Concepts are words or phrases extracted from your text data
- B. Types are semantic groupings of categories stored in the form of category dictionaries
- C. Categories can only be built using linguistic techniques of classification
- D. Each category is defined by one or more descriptor

Question 10: True or False: Categories may be used as Types but not vice-versa.

- True
- False

Apply Your Knowledge - Solutions

- Answer 1: A. True. armm → arm
- Answer 2: B. Omission of optional elements.
- Answer 3: A. True
- Answer 4: B. False. riddle → ridl is not the same as ridde → rid
- Answer 5: A. True
- Answer 6: D. Geographic variant
- Answer 7: C. Categories
- Answer 8: A. True. Verbs and adjectives are not extracted unless you force them to be extracted with a Type
- Answer 9: B. False. Types are semantic groupings of concepts, stored in the form of Type dictionaries.
- Answer 10: B. False. The correct statement is Types may be used as Categories but not vice-versa

Summary

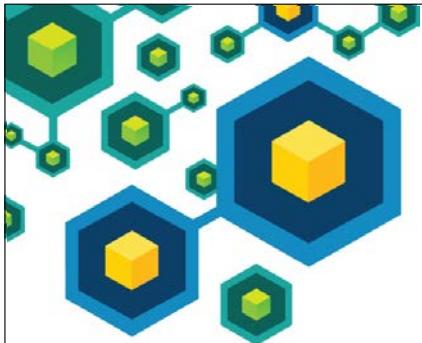
- At the end of this module, you should be able to:
 - describe linguistic analysis in general
 - describe the process of text extraction
 - describe categorization of terms and concepts
 - describe templates and libraries
 - describe Text Analytics Packages

© 2014 IBM Corporation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.



Creating a Text Mining Concept Model

IBM SPSS Modeler Text Analytics (v16)

Business Analytics software

© 2014 IBM Corporation



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Objectives

- At the end of this module, you should be able to:
 - develop a text mining concept model
 - compare models based on using different resource templates
 - score text data
 - analyze model results

© 2014 IBM Corporation

The Text Mining modeling node can create a concept- or category-based model, using the current state of the linguistic resources that are specified in the node. Building a model automatically assumes that the linguistic resources are sufficient for the model-building task at hand.

You might automatically build a concept-based model when you want to regularly (daily, weekly, etc.)-create a model with the latest text data that have been collected, after you previously edited and tuned the linguistic resources. Doing so will enable you to immediately use the concepts from a generated text mining model to make predictions on new and recent data records.

Additionally, creating a concept model, allows you to:

- see which concepts are extracted with the default resources
- see which concepts are most frequent and least frequent
- see which synonyms are being used
- look to see which text is not being extracted that could be important
- try several resource templates to compare the results

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Text Mining Concept Model

- Contains concepts that can be used to identify records or documents that also contain the concept (including any of its synonyms or grouped terms).
- Can be used to:
 - explore and analyze the concepts that were discovered in the original source text
 - apply this model to new text records or documents to quickly identify the same key concepts in the new documents/records

© 2014 IBM Corporation



During the extraction process, the text data is scanned and analyzed in order to identify interesting or relevant single words, such as election or peace, and word phrases, such as presidential election, election of the president, or peace treaties. These words and phrases are collectively referred to as terms. Using the linguistic resources, the relevant terms are extracted, and similar terms are grouped together under a lead term called a concept.

Models can be built with concepts to explore the text data at a detailed level to provide deep understanding. Many concepts are typically extracted from a data set, so as a rule, not all the concepts are used in the final model. Often the most frequent concepts are retained, but this will vary depending on the purposes of modeling and the interest you have in specific concepts.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Creating a Concept Model

- Select Generate directly (concept model nugget) in the Build mode section in the Model tab of the Text Mining modeling node.
- The model will be built using the model linguistic resources that are selected in the Resource template area.
- A concept model nugget will be placed directly in the Models palette after the Text Mining node has finished running.

© 2014 IBM Corporation



The options in the Generate Directly area become active when you select the option to build a model directly. These options allow you to set the maximum number of concepts that you want to include in the model, and how frequently a concept needs to occur before you want it used for scoring.

Specifying Model Options

- The following options in the Generate Directly area become active when you select the option to build a model directly:
 - maximum number of concepts to include in model
 - check concepts based on highest frequency
 - uncheck concepts that occur in too many records
 - optimize for speed of scoring

© 2014 IBM Corporation



By default, a concept-based model is created from up to the 500 most frequent concepts (by global frequency, which is the total number of times a concept appears in the text). This can create a large number of concepts to examine, so reduce the number here.

The Check concepts based on highest frequency option selects that number (75 by default) of the most frequent concepts for scoring in the model. Note that you can always select concepts for scoring later from the generated model.

The option to Uncheck concepts that occur in too many records specifies that those concepts appearing in more than the specified percentage of records (or documents) should not be selected in the model. Typically, text that appears too often is not useful because it consists of words such as "customer", "call", or a company name that are either redundant or don't contain any specific information.

Optimize for speed of scoring is selected by default. This option ensures that the model created is compact and scores at high speed. Deselecting this option creates a much larger model which scores more slowly.

Selecting a Resource Template / TAP

- When text mining, it is important to select the resource template or Text Analysis Package that is most appropriate for your data.
- These resources serve as the basis for how to handle and process the text during extraction in order to get the concepts, types, and sometimes patterns.

© 2014 IBM Corporation



One resource template is used for any text mining model. A number of templates are shipped with the software, and the available templates for the installation of IBM SPSS Text Analytics for Modeler are listed in the Load Resource Template dialog box. These shipped resources allow you to benefit from years of research and fine-tuning for specific languages and specific applications. Since the shipped resources may not always be perfectly adapted to the context of your data, you can edit these resource templates or even create and use custom libraries uniquely fine-tuned to the data of your organization.

For example, if you are dealing with customer survey data, there are at least two templates that would seem to be possible choices, including CRM (English) for customer relationship management, and Opinions (English), because they are designed to extract and code data from customer surveys. Other templates are designed specifically for Customer Satisfaction, Employee Satisfaction, Market Intelligence, and Insurance CRM data. There are several others besides these.

Selecting Expert Tab Options

- The Expert tab contains advanced parameters that control how text is extracted and handled
- Parameters in this dialog box represent only a portion of the options available to you for controlling basic behavior of extraction
- The resource template you select and the linguistic resources in the interactive workbench also affect extraction results

© 2014 IBM Corporation



The parameters in this dialog box control the basic behavior of the extraction process. However, they represent only a portion of the options available. The linguistic resources and options available through the interactive workbench impact the extraction results, and these are, in part, controlled by the resource template you select on the Model tab. To edit those, you would need to use the interactive workbench mode of the text-mining node.

By default, any concept that appears at least once will be extracted, as controlled by the Limit extraction to concepts with a global frequency of at least specification. This setting is as low as possible, so especially for larger files you may well choose to increase this value.

Accommodate punctuation errors: This option will apply a normalization technique to improve the extractability of concepts from text data containing many punctuation errors. These errors include the improper use of punctuation, such as a period, comma, semicolon, colon, and forward slash. This option is extremely useful when text quality may be poor (as, for example, in open-ended survey responses, e-mail, and CRM data) or when the text contains many abbreviations. As with all the extraction procedures, normalization does not permanently alter the text but modifies a working version of the text.

Accommodate spelling for a minimum root character limit of: This option is used to fix spelling errors. It uses a fuzzy grouping technique to find misspelled variants of words. The algorithm works by removing vowels and double/triple consonants from extracted words and comparing them to see if they are the same. However, if each term is assigned to a different type, excluding the <Unknown> type, the fuzzy grouping technique will not be applied.

By default, this option applies only to words with five or more root characters. The number of root characters in a term is calculated by totaling all of the characters and subtracting any characters that form inflection suffixes and, in the case of compound word terms, determiners and prepositions. For example, the term "stores" would be counted as 5 root characters in the form of store since the letter "s" at the end of the word is an inflection (plural form). Similarly, "manufacturing of cars" counts as 16 root characters (manufacturing car) because "of" is not counted. This method of counting is only used to check if the fuzzy grouping should be applied but does not influence how the words are matched.

If you find that using this option also groups certain words incorrectly, you can exclude word pairs from this technique by explicitly declaring them in the advanced resources editor in the Fuzzy Grouping / Exceptions section of the interactive workbench.

Extract uniterms: In most instances you will want to extract as many individual words as possible as terms; the Extract uniterms check box controls this option. Uniterms will be extracted when the word is not part of a compound word, the word is unknown to the extraction dictionary, or the word is identified as a noun in the dictionary.

Extract nonlinguistic entities: Text data often contains nonlinguistic information, meaning such things as phone numbers, dates, e-mail addresses, or customer ID codes. The extraction of these terms is controlled by the Extract nonlinguistic entities check box. You have additional control over the various types of nonlinguistic entities in the Nonlinguistic Entities section in Advanced Resources. You can disable the entities you do not need to decrease processing time in this area.

Uppercase algorithm: If the text data is not all in upper or lower case, then you should consider using the Uppercase algorithm option. This algorithm extracts simple words that are not in the extraction dictionary, as long as the first letter is in upper case. The algorithm also extracts sequences of words as compound words, provided they all start with a capital letter (for example, New York Stock Exchange).

Group partial and full person names together when possible: This option collects names that appear differently in the text together. This feature is helpful since names are often referred to in their full form at the beginning of the text and then only by a shorter version. This option attempts to match any uniterm with the <Unknown> type to the last word of any of the compound terms that is typed as <Person>. For example, if "smith" is found and initially typed as <Unknown>, the extraction engine checks to see if any compound terms in the <Person> type include "smith" as the last word, such as "john smith". This option does not apply to first names since most are never extracted as uniterms.

Maximum nonfunction word permutation: As you learned earlier, the extraction process permutes word order to group similar phrases that vary only because nonfunction words (for example, "of" and "the") are present, regardless of inflection. The Maximum nonfunction word permutation option specifies the maximum number of nonfunction words that must be present to apply the permutation technique. This technique would group together "sailing club officers" and "officers of the sailing club" in the same concept since both terms are the same if the word "of" is removed and the words are permuted.

Text Mining Nugget: Concept Model

- Edit the model to view the results.
- The extracted concepts are presented in a table format with one row for each concept.
- All concepts are selected for scoring by default:
 - a checked box means that the concept will be used for scoring.
 - an unchecked box means that the concept will be excluded from scoring.

© 2014 IBM Corporation



In the model nugget generated by running a concept model, the Model tab displays the set of concepts that were extracted. The concepts are presented in a table format with one row for each concept. To learn more about each concept, you can look at the additional information provided in each of the following columns:

- Check box in the leftmost column indicates whether or not a concept will be part of the output when data is read and the model applied.
- Concept is the concept name, generated from the text (always in lower case).
- Global is the total frequency count or total number of occurrences of that concept in the data, represented as a bar chart.
- % is the percentage the concept frequency is of the total number of concept occurrences (sum of all frequencies regardless of check box status).
- N is the number of occurrences.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

5-11

- Docs is either the total number of records or the total number of documents, in which this concept is found, represented as a bar chart.
- % is the percentage of the number of records or documents in which the concept occurred, compared to the total number of records or documents (5,220 total records).
- N is the number of records or documents in which the concept occurred.
- Type is the type to which the concept was assigned, where colors represent the types. Thus olive green means Unknown, while red signifies Negative.

Below the table, the number of checked concepts (Concepts selected for scoring) and the total number of concepts (Total concepts available) are shown. There is a Sort by: button to sort the results by other elements of the table, including the Docs percentage, alphabetically by the concepts, alphabetically by the types, or by the checked (selected) concepts.

There is a Sort by: button to sort the results by other elements of the table, including the Docs percentage, alphabetically by the concepts, alphabetically by the types, or by the checked (selected) concepts.

It is typical that there are many concepts with Unknown type when you use a resources template that has not been tuned to this particular text dataset.

Underlying Terms in Concept Models

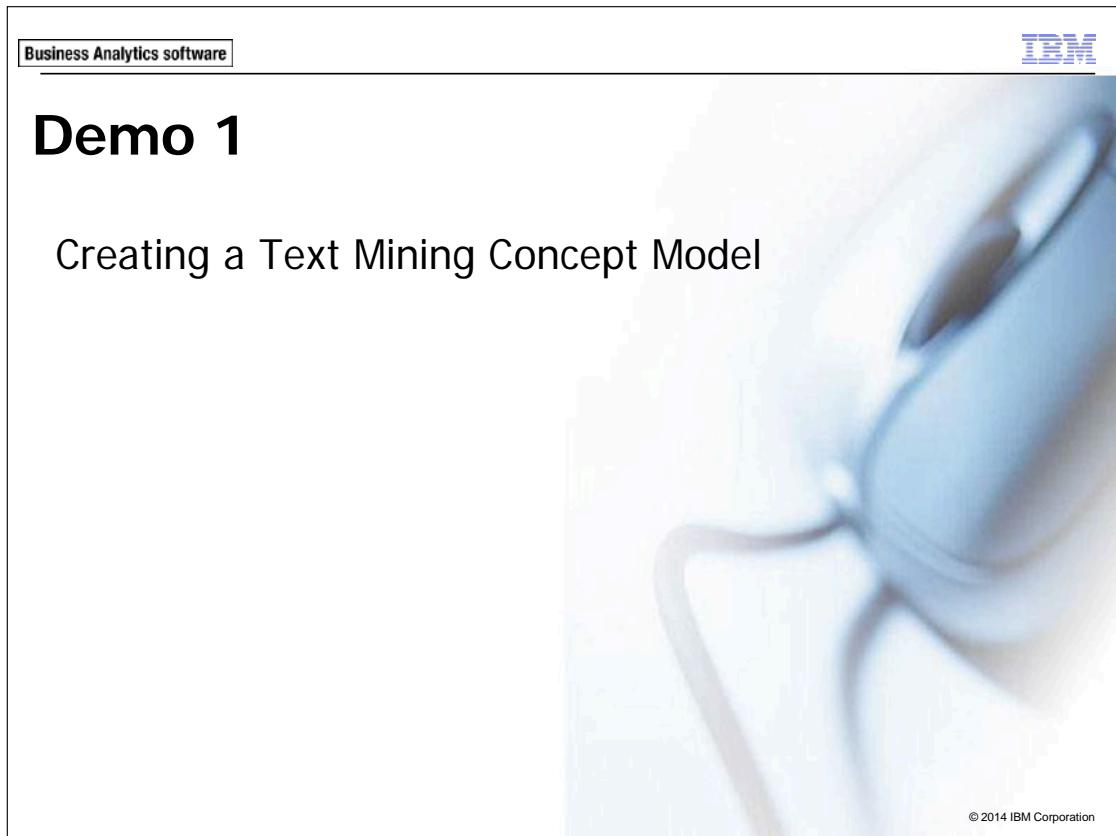
- Click on each concept to see the underlying terms that are defined for it.
- Displayed in a split pane at the bottom of the dialog.
- Include the synonyms defined in the linguistic resources (regardless of whether they were found in the text or not).
- Also include any extracted plural/singular forms found in the text used to generate the model nugget, permuted terms, terms from fuzzy grouping, etc.

© 2014 IBM Corporation



You can see the underlying terms that are defined for the concepts that you have selected in the table. By clicking the underlying terms toggle button on the toolbar, you can display the underlying terms table in a split pane at the bottom of the dialog. These underlying terms include the synonyms defined in the linguistic resources (regardless of whether they were found in the text or not) as well as any extracted plural/singular forms found in the text used to generate the model nugget, permuted terms, terms from fuzzy grouping, and so on.

Note: You cannot edit the list of underlying terms. This list is generated through substitutions, synonym definitions (in the substitution dictionary), fuzzy grouping, and more—all of which are defined in the linguistic resources. In order to make changes to how terms are grouped under a concept or how they are handled, you must make changes directly in the resources (editable in the Resource Editor in the Interactive Workbench or in the Template Editor and then reload the template in the node) and then reexecute the stream to get a new model nugget with the updated results.



The slide has a light blue background with a faint, abstract graphic of a person's head and shoulders. In the top left corner, there is a small rectangular box containing the text "Business Analytics software". In the top right corner, the "IBM" logo is displayed. The main title "Demo 1" is centered at the top in a large, bold, black font. Below it, the subtitle "Creating a Text Mining Concept Model" is also centered in a smaller, regular black font. At the bottom right of the slide, there is a small, faint copyright notice: "© 2014 IBM Corporation".

This demo uses the following datasets coming from a (fictitious) telecommunications firm.

- C:\Train\0A105\Astroserve0304.sav - a Statistics file storing call center data for March and April.

Demo 1: Creating a Text Mining Concept Model

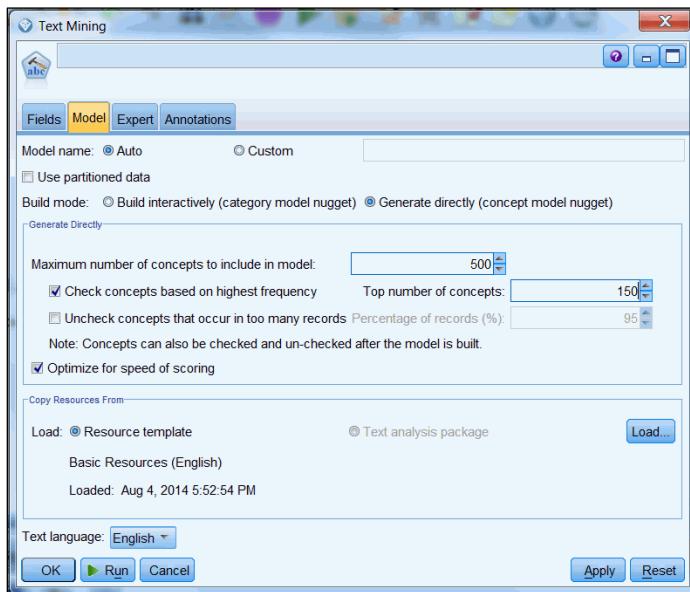
Purpose:

You are working at Astroserve and would like explore the call center data from March and April to find out what topics customers are calling about.

Task 1. Preparing to build a concept model.

1. Create a new stream, and then from the **IBM SPSS Statistics** tab, add a **Statistics File** node.
2. Edit the file, import **Astroserve0304.sav** from the **C:\Train\0A105** folder, and then click **OK**.
3. From the **IBM SPSS Text Analytics** palette, add a **Text Mining** node to the stream and connect it to the **Statistics File** node.
4. Edit the **Text Mining** node.
5. On the **Fields** tab, click the **Text Field** chooser, and then select **query**.
6. Click the **Model** tab, and then select **Generate directly (concept model nugget)**.
7. Select **Check concepts based on highest frequency**, and then beside **Top number of concepts**, enter **150**.

The results appear as follows:



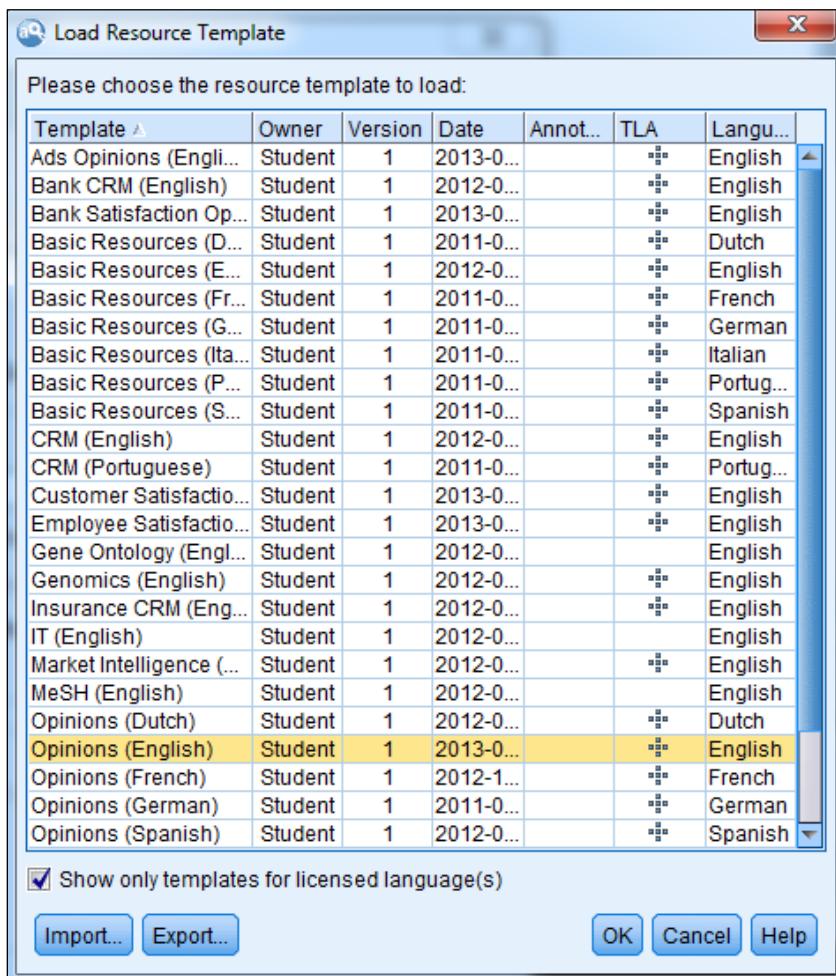
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

8. In the **Copy Resource From** section, click **Load**.
9. Click the **Opinions (English)** template to select it.

You are selecting this template because customers are calling with complaints and service issues, and often have strong opinions about their service. The results appear as follows:

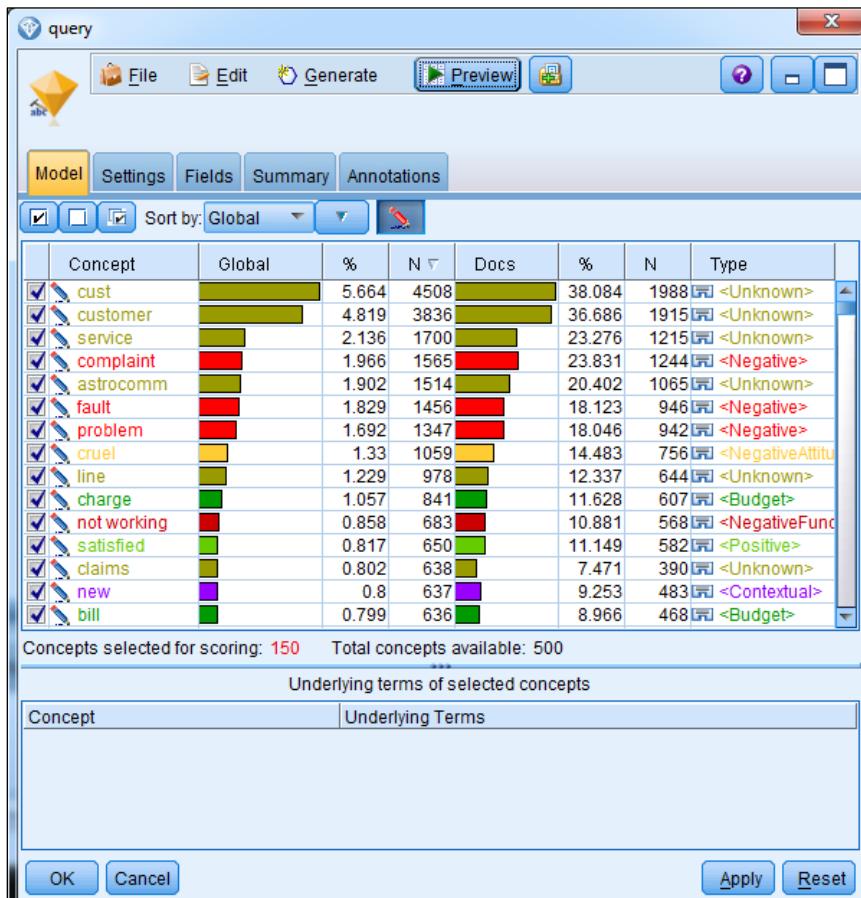


10. Click **OK**, and then click the **Expert** tab.
11. Select **Accommodate spelling errors for a minimum root charter limit of**, and then specify **4**.
12. Click **Run**.

Task 2. Exploring the extracted concepts.

- After the execution is complete, double-click the model nugget on the stream canvas.

The results appear as follows:

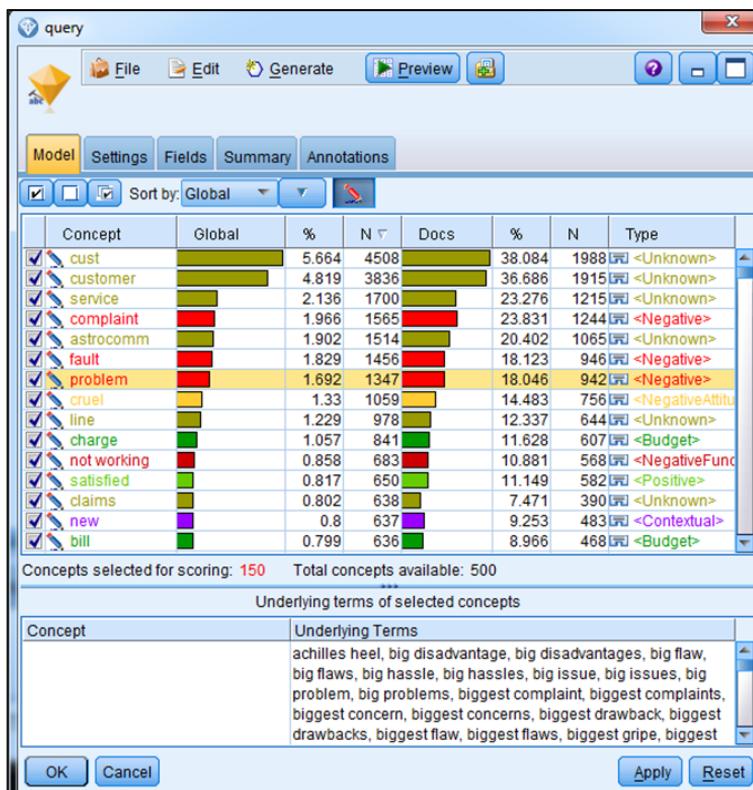


As was requested, the generated model contains 150 total concepts that will be scored, ordered by global frequency. Also listed is the document/record frequency.

As is typical when using a resources template that has not been tuned to this particular text dataset, there are many concepts with Unknown type. Also, the two most common concepts are "cust" and "customer", which as was noted earlier, are probably not useful for text mining. They both appear in 36-38% of all records.

- Click the concept, **problem** to select it.

The results appear as follows:



There are dozens of synonyms for problem, including such terms as difficulty, hassle, and trouble. Interestingly, there are also misspelled variants for these synonyms.

About 1.7% of all concepts that were extracted refer to a problem, which is mentioned in 18.046% of all records (not surprising given the nature of these data). Problem is typed as Negative.

- Click **OK** to save the model in this state, and then from the **File** menu, click **Save Stream As**.
- Name the stream **Creating_a_Concept_Model_Demo1_end**, and then click **Save**.
- From the **File** menu, click **Close Stream**.
- From the **File** menu, click **New Stream**.

Do not close Modeler. Leave it open for the next demo.

Results:

You explored the call center data from March and April to find out what topics customers were calling about.

Comparing Different Templates

- Using a different template will produce similar but not identical results.
- Each template is tuned to a specific type of data.
- Care should be taken to select the template that is most appropriate for your data.

© 2014 IBM Corporation



Text mining is a reliable method of data mining, meaning that the same linguistic resources used on the same data will yield identical results 100% of the time. However, there is no theory or set of principles upon which to base the creation and editing of the linguistic resources for a particular data set, and two users faced with the same text mining project will very likely construct somewhat different linguistic resources. As a consequence, they will create somewhat different text-mining models.

Given this, it can be very instructive to try different linguistic resources on the data. The results are summarized in the table below. Also displayed is the document percentage. Those concepts unique to one model (of the top ten concepts) are in bold.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Resource Template					
Opinions (English)		CRM (English)		Basic Resources (English)	
Concept	Frequency (%)	Concept	Frequency (%)	Concept	Frequency (%)
<u>cust</u>	38.1	customer	75.1	<u>cust</u>	38.3
customer	36.7	<u>requested</u>	39.0	customer	36.7
complaint	23.8	service	37.2	<u>astrocomm</u>	24.1
service	23.3	<u>advised</u>	33.2	service	22.8
<u>astrocomm</u>	20.4	<u>claim</u>	28.9	call	16.1
fault	18.1	<u>astrocomm</u>	27.4	complaint	14.7
problem	18.0	call	26.1	fault	12.3
cruel	14.5	<u>wrong</u>	25.8	line	11.9
line	12.3	<u>phone number</u>	22.8	<u>astroserve</u>	9.4
charge	11.7	received	19.5	bill	9.4

There is great similarity between the models. The concepts "customer", "astrocomm", and "service" are all near the top. However, there are dissimilarities as well. For instance, for the Opinions template the concept "problem" was placed in the top 10 (18.0%) but did not appear near the top of the list in the case of the Basic English or CRM templates. Also notice that for the Opinions template, the concept "complaint" appears higher in the list of important concepts and in a higher percentage of documents. This is because its linguistic resources are more specially tuned to locate customer opinions than the other two templates and thus can be extremely useful when performing sentiment analysis.

For other concepts, different templates have an advantage. The concepts "cust" and "customer" certainly refer to the same notion, but only the CRM template-based model groups them into one category under the label customer.

You should not be discouraged by the comparison of the three models. Text mining is not an exact science, and all three models may well be useful in particular circumstances. The general goal of text mining is to extract the great majority of information from the text data, with the understanding that there will be always be some variation caused by both the ambiguity of language and the interaction between various components of the linguistic resources.

Selecting Concepts for Scoring

- Scoring the data is the process of applying a model to the data
- Identifies the customers who expressed the various opinions mentioned in specific concepts
- All concepts that have a check mark on the Model tab will be included for scoring
- You can uncheck the ones you are not interested in so they will not be used in the scoring

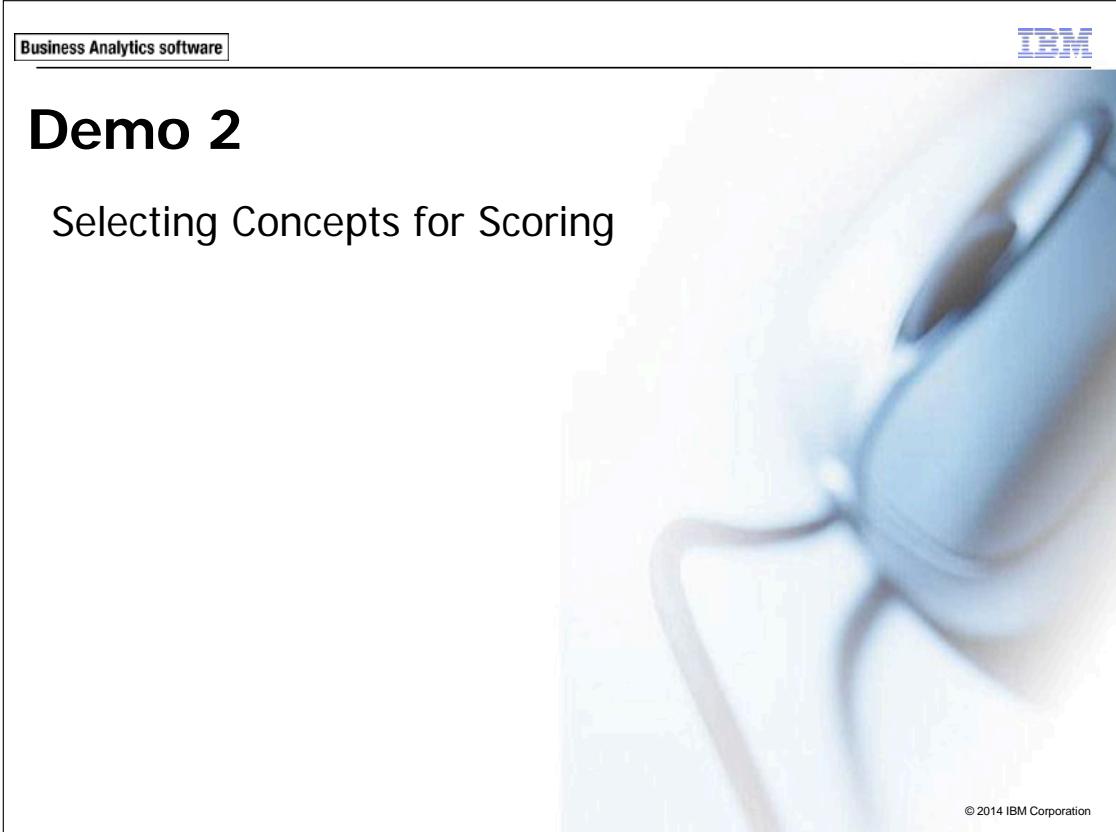
© 2014 IBM Corporation



Now return to the results for the model using the Opinions template and review the options for selecting concepts for scoring. Only those concepts whose boxes are checked will be used to score text data. You can certainly manually check each box for each concept that you want to use, but there is an available dialog box to provide some automatic options.

There are several options, only one of which can be selected at one time.

- Check concepts based on frequency: Selects the top n concepts based on the global frequency specified in the text box.
- Check concepts based on document count: Selects the concepts based on the minimum record/document count specified in the text box.
- Check concepts assigned to the type: Enables you to select concepts that are all of a certain type.
- Uncheck concepts that occur in too many records: Unchecks concepts with a record (or document) percentage higher than the number you specified.
- Uncheck concepts assigned to the type: This option will deselect concepts of a certain type.



The slide is titled "Demo 2: Selecting Concepts for Scoring". It features the IBM logo in the top right corner and the text "Business Analytics software" in the top left corner. A faint background image of a person wearing a bow tie is visible. The bottom right corner contains the text "© 2014 IBM Corporation".

This demo uses the following datasets coming from a (fictitious) telecommunications firm.

- C:\Train\0A105\05-Creating a Text Mining Concept Model\Creating_a_Concept_Model_Demo2_start.str

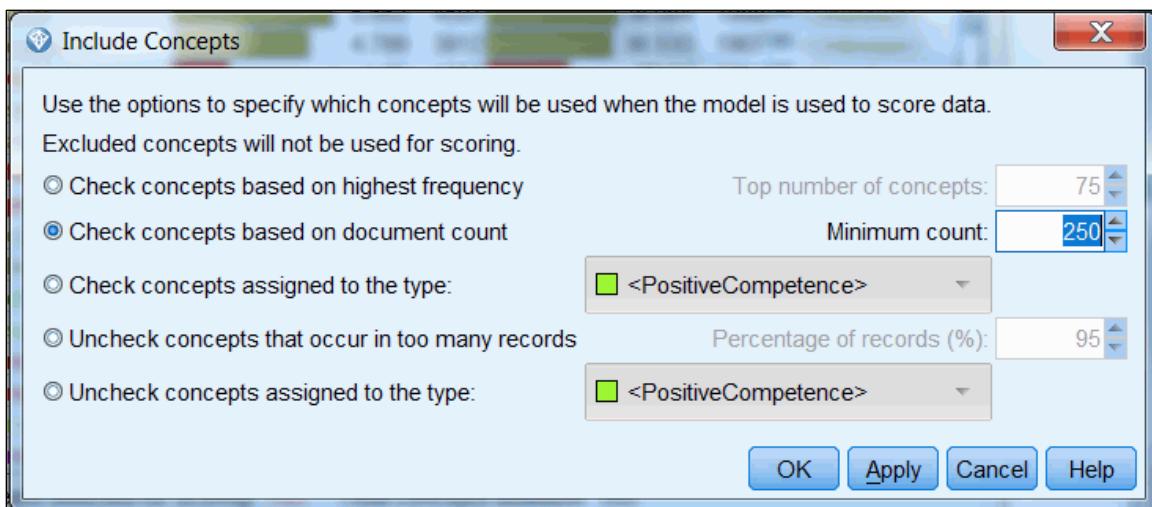
Demo 2: Selecting Concepts for Scoring

Purpose:

Before scoring the data, you should select a subset of concepts that you would like to use for scoring. You will select all the concepts that occur in 250 records (which is about 5% of all the records).

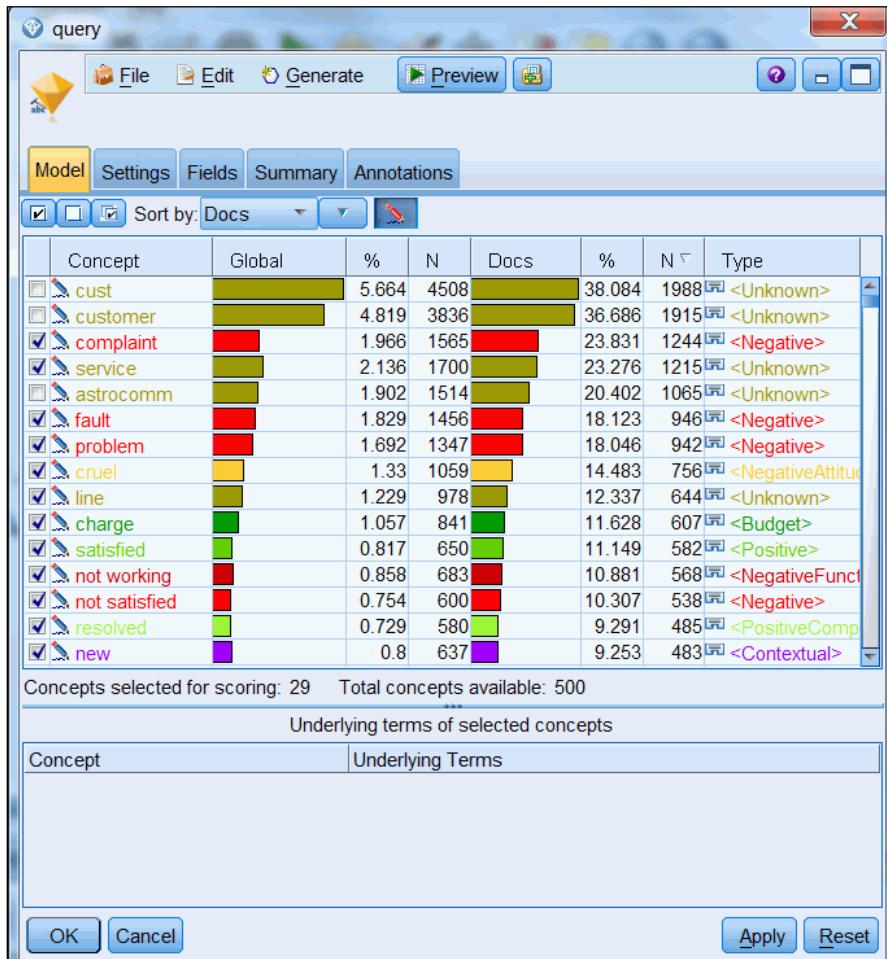
Task 1. Selecting the Most Frequently Occurring Concepts.

1. From the **File** menu, click **Open Stream**.
2. Select **C:\Train\0A105\05-Creating a Text Mining Concept Model\Creating_a_Concept_Model_Demo2_start.str**, and then click **Open**.
3. Double-click the **Text Mining** model nugget located on the stream canvas.
4. In the **Sort by** list, select **Docs**.
5. Deselect the concepts of **cust**, **customer**, and **astrocomm**.
6. Click **Include concepts using rules**  on the toolbar.
7. Select **Check concepts based on document count** and enter **250** in the text box.



8. Click **OK**.

9. Deselect the check boxes for **cust**, **customer**, and **astrocomm**.



The message below the list of concepts tells you that 29 concepts were selected with this specification.

10. Click **OK** to save the model in this state.
11. From the **File** menu, click **Save Stream As**.
12. In the **File Name** box, type **Create_a_Concept_Model_Demo2_end.str** and then click **Save**.
13. From the **File** menu, click **Close Stream**.
14. From the **File** menu, click **New Stream**.

Do not close Modeler. Leave it open for the next demo.

Results:

Before scoring the data, you selected a subset of concepts to use for scoring. You selected all the concepts that occur in at least 250 records.

Scoring Model Data

- After selecting the concepts you want to use:
 - add generated model to the stream
 - run the existing the data or new data through the model
 - identify which customers expressed various opinions and mentioned specific concepts

© 2014 IBM Corporation



The concepts of interest have been selected so that the data are scored. For simple understanding of the key concepts expressed by the Astroserve customers in their phone calls to the call center, the information you have already gleaned from the model may be helpful, especially concept frequency. But in text mining, it is very likely that you will add the generated model to a stream and score existing or new data so you know which customers expressed various opinions and mentioned specific concepts.

When you runs a stream that passes data through a generated model, new fields (the concepts) are added to the data. Data coming into the Text Mining model node need not be the data used to train the model. It could be new data to which the generated model is applied since it is the database of concepts which will be applied to the incoming text data, not new concept extractions. However, the data must contain the same input fields and field types as the training data used to create the model.

Note: there is usually no reason to retain the actual text once the model is applied (after the Generated model has been created). Therefore, it is a good idea to use a Filter node to remove it, which reduces file size and the size of Table displays.

Specifying the Scoring Mode

- There are two types of scoring modes available in the model:
 - concepts as fields – create one new field (column) on each record (row) for every concept
 - concepts as records – create a separate record (row) for each concept

© 2014 IBM Corporation

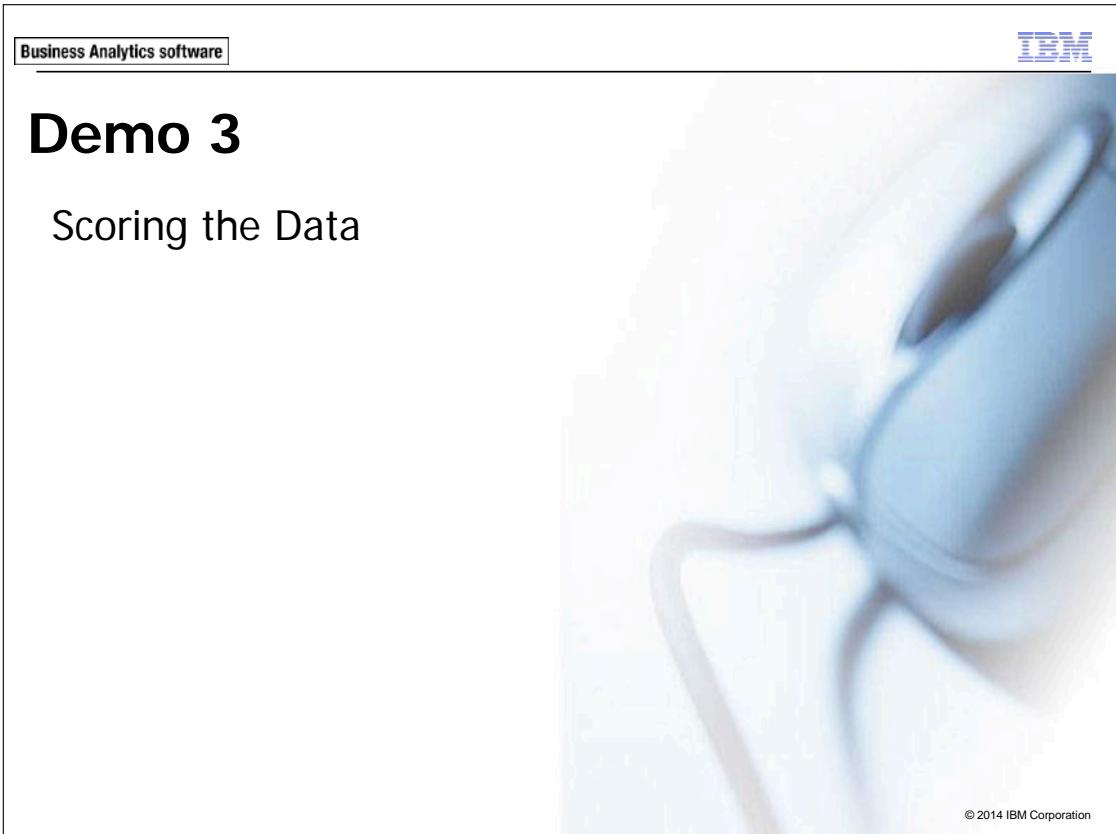


The Settings tab is used to define the type of output to be created when the text is scored. The model has two types of scoring modes:

- Concepts as fields: This creates one field per selected concept. There will be just as many records downstream as the data entering the model node. This mode allows you to concentrate on the original records (calls to the service center, in this example). When the concepts will be scored as fields, the new fields will be of type flag and have values of "T" (when the concept is present) and "F" (when it is not). If you prefer, you can change these values to something like "1" and "0". The new fields will be given a Field name extension of Concept_, which will be added as a prefix.
- Concepts as records: This creates one record per selected concept. The output contains all of the input fields plus three additional fields. Typically, there are more records in the output than there were in the input. This mode allows you to concentrate more on the concepts themselves.

As an alternative, the new fields can store the number of times that a concept occurred in each record or document, instead of just whether or not it was present. . For example, if the concept "bill" was mentioned twice by a customer, that record would have a value of 2 for the new field "Concept_bill".

When text data are scored, new synonyms are not created, and no additional fuzzy grouping is done. Instead, the concepts, with their associated synonyms, are located in specific records or documents for scoring.



The slide has a light blue background with a faint, abstract graphic of a person's head and shoulders. In the top left corner, there is a small rectangular box containing the text "Business Analytics software". In the top right corner, the "IBM" logo is displayed. The main title "Demo 3" is centered at the top in a large, bold, black font. Below it, the subtitle "Scoring the Data" is also centered in a smaller, regular black font. At the bottom right of the slide, there is a small, fine-print copyright notice: "© 2014 IBM Corporation".

This demo uses the following datasets coming from a (fictitious) telecommunications firm.

- C:\Train\0A105\05-Creating a Text Mining Concept Model\Create_a_Concept_Model_Demo3_start.str

Demo 3: Scoring the Data

Purpose:

Now that you have created the concept model, the next step is to discover which concepts customers mentioned the most. This will hopefully clarify the key topics customers are complaining about.

Task 1. Summarizing your findings.

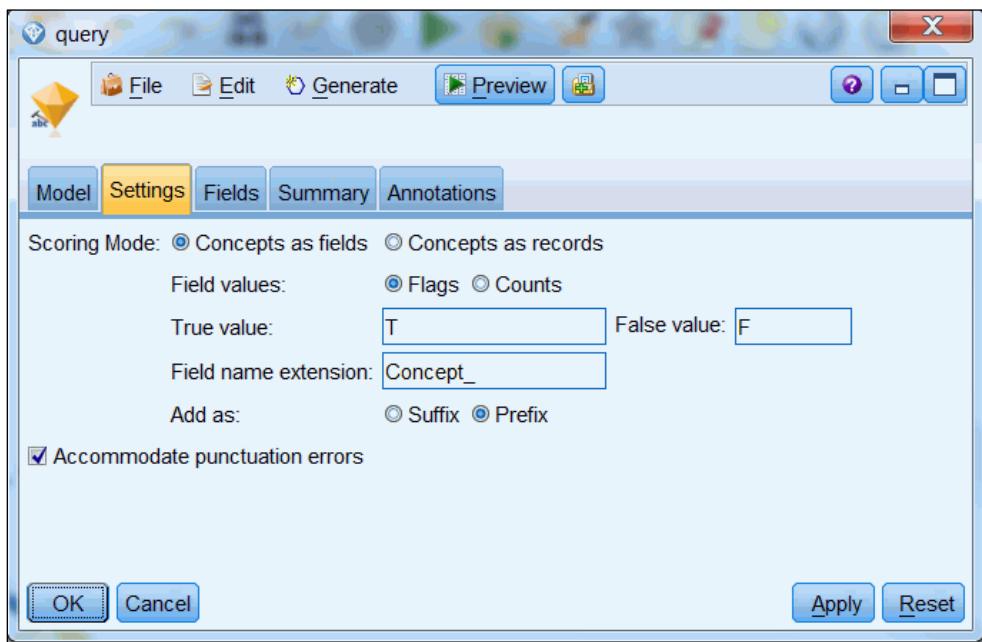
1. Click **File** and then **Open Stream**.
2. Select **C:\Train\0A105\05-Creating a Text Mining Concept Model\Create_a_Concept_Model_Demo3_start.str**, and then click **Open**.
3. Double-click the **Text Mining** model nugget located on the stream canvas.
4. Click the **Fields** tab.

When you apply a text-mining model to new text, the text field may not have the same name. Alternatively, you could apply this model to text stored in files rather than in a field. The Fields tab provides the ability to make the necessary changes to define the text to be scored. In this case, the same data will be used, so no changes need be made.

5. Click the **Settings** tab.

Take the defaults and create concepts as flags.

The results appear as follows:



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

5-29

6. Click **OK**.
7. Add a **Table** node downstream from the model nugget.
8. Run the **Table** node.

The results appear as follows:

This screenshot shows the output of a Table node in IBM SPSS Modeler. The window title is "Table (33 fields, 5,220 records)". The menu bar includes File, Edit, Generate, and various icons. The toolbar has buttons for Table, Annotations, and other functions. The main area is a grid with 20 rows and 14 columns. The first column is labeled with numbers 1 through 20. The columns are labeled: Concept_not satisfied, Concept_letter, Concept_pay, Concept_complaint, Concept_wrong, Concept_account, Concept_phone, Concept_astroserve, Concept_tia level, Concept_satisfied, Concept_charge, Concept_charge, Concept_charge, and Concept_charge. Most values are 'F', with several 'T's appearing in the Concept_phone and Concept_charge columns across different rows.

Twenty-nine new flag fields have been created, one for each concept checked for scoring. Looking at the first record, two of the flag fields, "Concept_phone" and "Concept_charge" have values of T. A review the text from that call confirms that the customer mentioned both "phone", and "charge" in their call. Thus, IBM SPSS Modeler Text Analytics correctly extracted those two concepts from that record.

This screenshot shows the output of a Table node in IBM SPSS Modeler. The window title is "Table (33 fields, 5,220 records)". The menu bar includes File, Edit, Generate, and various icons. The toolbar has buttons for Table, Annotations, and other functions. The main area is a grid with 20 rows and 14 columns. The first column is labeled with numbers 1 through 20. The columns are labeled: Query_ID, day, month, query, Concept_premises, Concept_bill, and Concept_charge. The "query" column contains detailed customer service logs. The "Concept_premises", "Concept_bill", and "Concept_charge" columns contain binary values (F or T) corresponding to the presence of specific concepts in the "query" text.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Oddly, this same customer also mentioned that she was upset, but the flag field "Concept_not satisfied" has a value of F. The linguistic resources have not been edited yet, so that type of association was not picked up by the model, especially at the concept level.

Because the option to create fields with the concepts model has been used, next you will examine the option to create records.

9. Close the table, and then edit the **Generated Model**.
10. Select **Concepts as records**, click **OK**, and then run the **Table**.

The results appear as follows:

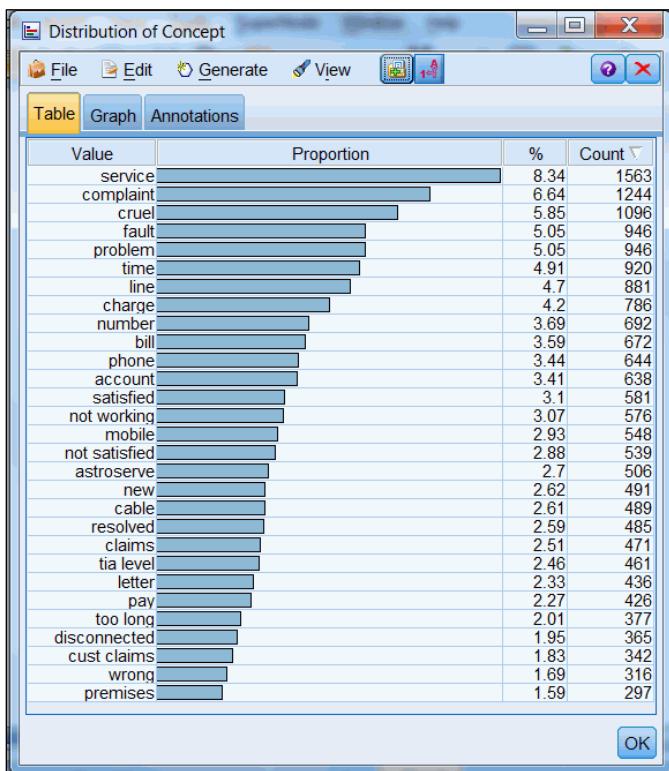
The 5,220 records in the source data have become 18,734 records. This means that there are over 3 concepts, on average, per input record ($18,734/5,220$). In fact, notice that the first two records have the same value for Query_ID, and so are from the same call to the call center. They therefore also have the same text for the query field, and the same dates. The Concept field lists which concepts were discovered in the text, which are charge, and phone for this call.

Data in this format allow you to create a report to determine the frequency with which the concepts occur. To do this:

11. Close the **Table** window.
12. Add a **Distribution** node from the **Graphs** palette to the stream.
13. Connect the **generated model** to the **Distribution** node.
14. Edit the **Distribution** node.
15. Beside **Field**, select **Concept**.
16. Beside **Sort**, select **By count**.

17. Select Proportional Scale, and then click Run.

The results appear as follows:



The percentages are not based on the number of records in the input data, but on the total number of times all the concepts have appeared. But the count is accurate, i.e., no count can be above 5,220, the number of input records. Notice that the most frequent concepts are "service" and "complaint". The fact that these two concepts are at the top of the list suggests that substantial numbers of people are calling in about service issues and because their phones are not working properly.

18. Click **OK** to close the Distribution output window, and then from the **File** menu, click **Save Stream As**.
19. Name the stream **Create_a_Concept_Model_Demo3_end**, and then click **Save**.
20. From the **File** menu, click **Close Stream**.
21. From the **File** menu, click **New Stream**.

Do not close Modeler. Leave it open for the next demo.

Results:

You viewed the data so you know which topics customers complained about the most. Clearly, service matters were what they complained about more than anything else.

Relating Concepts to Other Data

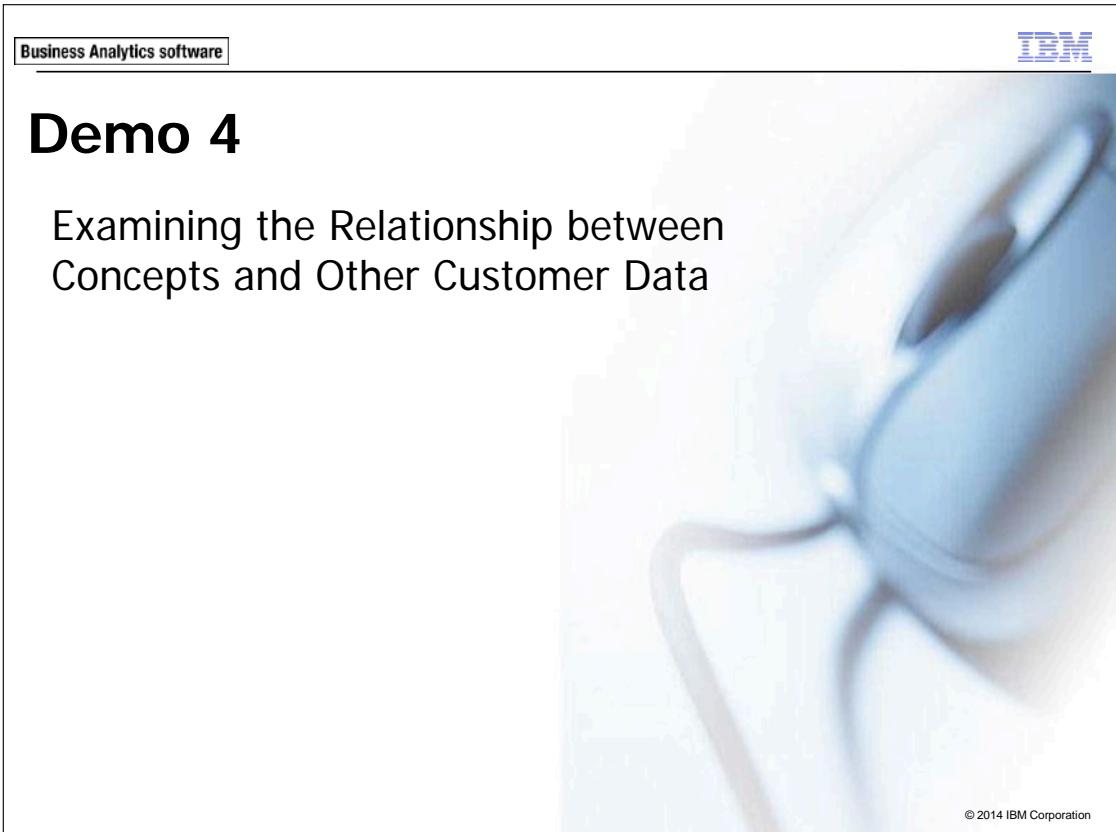
- Scoring helps you understand what customers are saying but it does not tell you much about who the customers are
- For example, what type of customer is most likely to complain? Are they males or females? Are they older customers or younger ones?
- To answer these questions, you will need to merge the concept data with additional information on the customers

© 2014 IBM Corporation



Although a concept model helps you understand what customers are saying, normally you want to do more with the scored text data than this type of simple, albeit valuable, report by looking at relationships between concepts and various other fields.

In the following demonstration, you will go through an example of how to do this.



The slide has a light blue background with a faint, abstract graphic of a person's head and shoulders. In the top left corner, there is a small rectangular box containing the text "Business Analytics software". In the top right corner, the "IBM" logo is displayed. The main title "Demo 4" is centered at the top in a large, bold, black font. Below it, the subtitle "Examining the Relationship between Concepts and Other Customer Data" is also centered in a smaller, black font. At the bottom right of the slide, there is a small, faint copyright notice: "© 2014 IBM Corporation".

This demo uses the following datasets coming from a (fictitious) telecommunications firm.

- C:\Train\0A105\05-Creating a Text Mining Concept Model\Create_a_Concept_Model_Demo4_start.str
- C:\Train\0A105\AstroCustomers.txt

Demo 4: Examining the Relationship between Concepts and Other Customer Data

Purpose:

Now that the data is scored, the next step is to examine the relationship between the concept data and additional information you have on the customers.

The following information on Astroserve customers is available in the file AstroCustomers.txt.

Field	Description
Query_ID	The query identification number
Cust_ID	The customer identification number
Priority (Y/N)	Whether customer is a Priority customer
Gender (F/M)	Gender of the primary account holder
A_cust (Y/N)	Whether customer has a landline account
A_years	Number of years as a landline customer.
A_lines	Number of land lines on the account
N_cust (Y/N)	Internet services (Astronet) account
N_years	Years as an internet services customer
N_type (B/D)	Type of account: Broadband or Dial-up
M_cust (Y/N)	Mobile phone account.
M_years	Years as a mobile phone customer
M_Period (MM/1Y/2Y)	Account Type: Monthly, 1 year, 2 years
Churn (Y/N)	Cancelled accounts by March or April

Task 1. Merging the model with customer data.

1. From the **File** menu, click **Open Stream**.
2. Select **C:\Train\0A105\05-Creating a Text Mining Concept Model\Create_a_Concept_Model_Demo4_start.str**, and then click **Open**.
3. Add a **Var File** node from the **Sources** palette to the stream near the generated categories model.
4. Edit the **Var File** node.
5. From **C:\Train\0A105** add the **AstroCustomers.txt** file.
6. In the **Field delimiters** area, select **Tab**, deselect **Comma**, and then click **OK**.
7. Add a **Table** node to the stream and connect it to the **Var File** node.
8. Run the **Table** node.

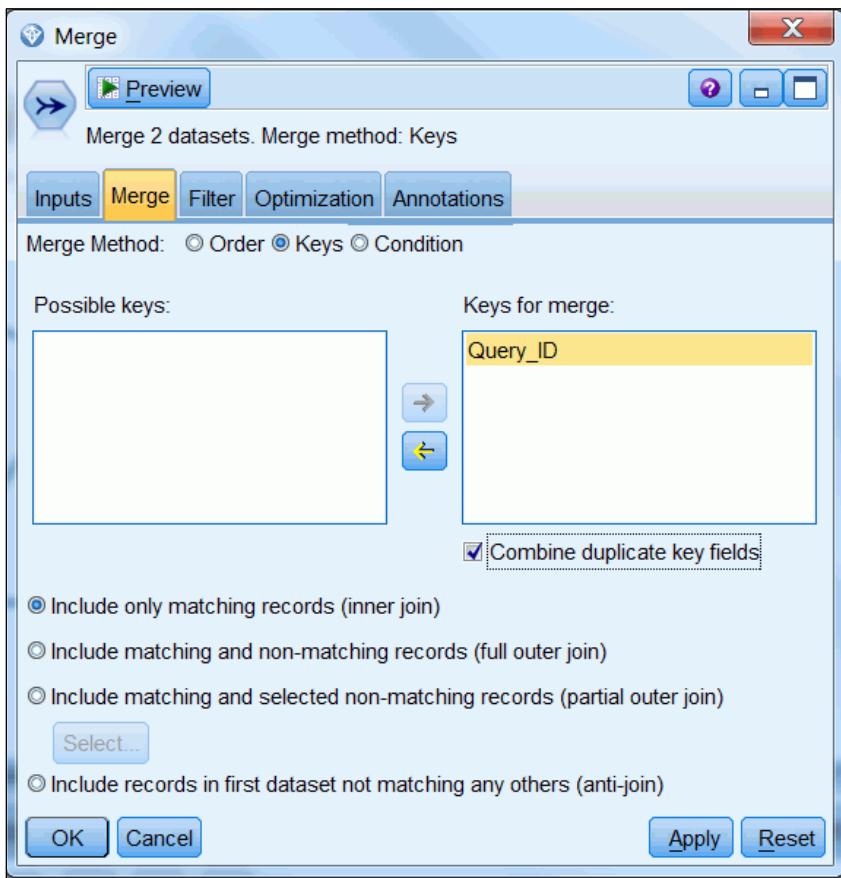
The results appear as follows:

	Query_ID	Cust_ID	Priority	Gender	A_cust	A_years	A_lines	N_cust	N_years	N_type	M_
1	386504	449 1429 27	N	F	N	\$null\$	\$null\$Y		2 B	Y	
2	386507	433 1970 25	N	M	Y	2	1 Y		2 D	N	
3	386510	438 1426 39	N	F	Y	13	3 N	\$null\$	N		
4	386513	434 1921 55	N	M	Y	4	1 N	\$null\$	N		
5	386514	432 1117 85	N	M	Y	0	3 N	\$null\$	N		
6	386517	449 1339 14	N	M	N	\$null\$	\$null\$Y		0 B	Y	
7	386519	434 1850 71	N	F	Y	5	3 N	\$null\$	Y		
8	386521	437 1631 26	N	M	Y	10	1 N	\$null\$	N		
9	386524	449 1139 84	N	M	N	\$null\$	\$null\$N	\$null\$	Y		
10	386526	436 1152 33	Y	M	Y	8	1 Y		1 D	Y	

There are two ID fields in this file. **Query_ID** references the call ID in the call center records. **Cust_ID** is the actual customer ID used to identify a customer. This data file is the link between the two IDs so customer data can be matched to call center data. There are more records in this file than in the call center data from March and April because the customer file also includes details for customers who called in May.

Now the two files need to be matched, but before you can do this, you need to switch the generated model to output flag variables.

9. Close the **Table** node.
10. Edit the model nugget named **query** located on the stream canvas.
11. Click the **Settings** tab, select **Concepts as fields**, and then click **OK**.
12. From the **Record Ops** tab, add a **Merge** node to the stream.
13. Connect the **AstroCustomers.txt** node to the **Merge** node.
14. Connect the model nugget named **query** to the **Merge** node.
15. Edit the **Merge** node.
16. Beside **Merge Method**, select **Keys**.
17. Use the arrow to move **Query_ID** to the **Keys for merge** box.



18. Click **OK**.
19. From the **Output** palette, add a **Table** node to the stream and connect it to the **Merge** node.

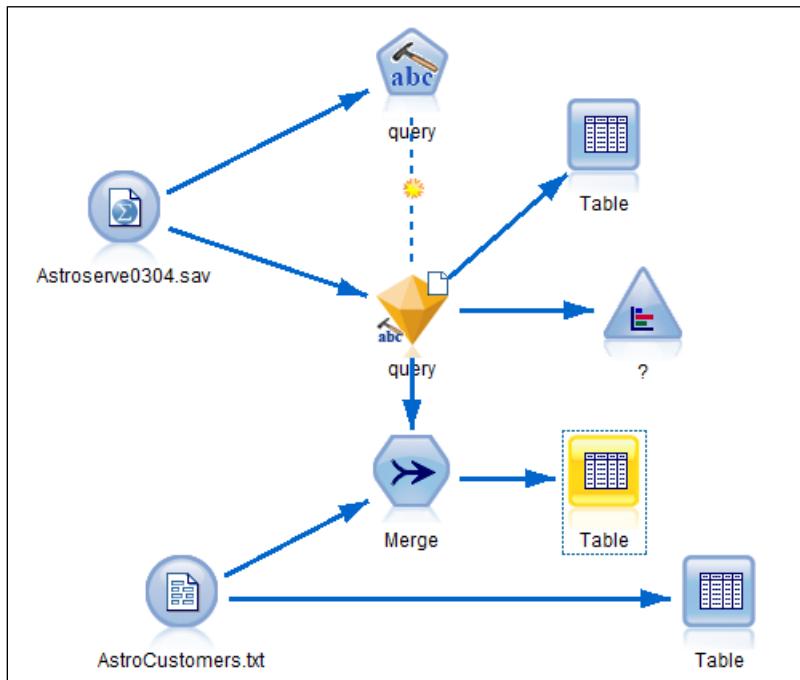
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

20. Right-click on the model nugget named **query**, point to **Cache**, and then click **Enable**.

This will store the scored data at the point of the cache so that text data does not have to be rescored every time you run the stream.



21. Run the **Table** node connected to the **Merge** node.

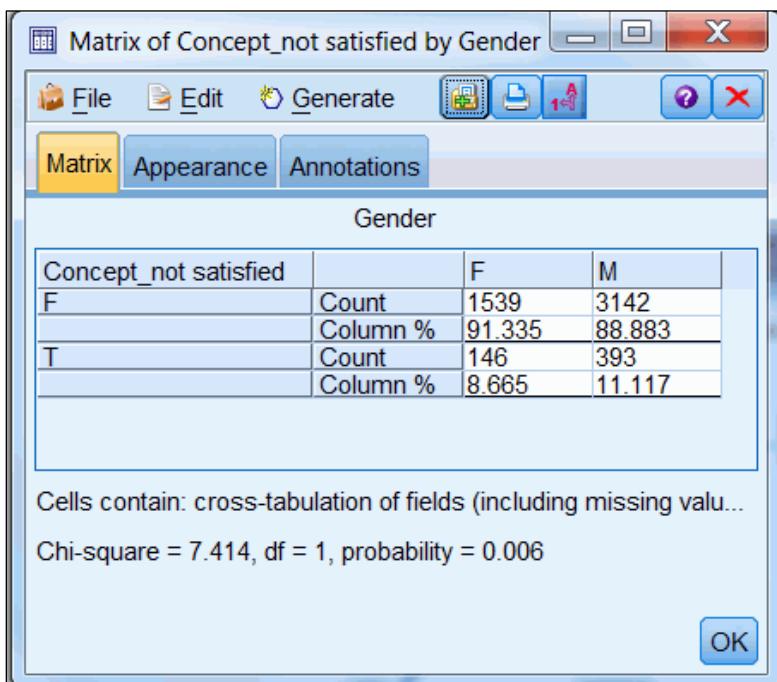
Table (46 fields, 5,220 records)																			
Annotations																			
Query_ID	Cust_ID	Priority	Gender	A_cust	A_years	A_lines	N_cust	N_years	N_type	M_cust	M_years	M_period	Churn	day	month	query			
1	386504449	142927	N	F	N	\$null\$	\$null\$Y	2B	Y	4MM	N	1...mar	Cust is upset that her						
2	386507433	197025	N	M	Y	2	1Y	2D	N	\$null\$	N	1...mar	Cust is dissatisfied as						
3	386510438	142939	N	F	Y	13	3N	\$null\$	N	\$null\$	N	1...mar	Customer advised that						
4	386513434	192155	N	M	Y	4	1N	\$null\$	N	\$null\$	N	1...mar	This number and dedic						
5	386514432	111785	N	M	Y	0	3N	\$null\$	N	\$null\$	N	1...mar	Customer called advi						
6	386517449	133914	N	M	N	\$null\$	\$null\$Y	0B	Y	11Y	N	1...mar	Priority cus cust can n						
7	386519434	185071	N	F	Y	5	3N	\$null\$	Y	62Y	Y	1...mar	Cust called in as she d						
8	386521437	163126	N	M	Y	10	1N	\$null\$	N	\$null\$	N	1...mar	Lightening damaged w						
9	386524449	113994	N	M	N	\$null\$	\$null\$N	\$null\$	Y	0MM	N	1...mar	Due to an order proce						
10	386526436	116233	N	M	Y	8	1Y	1D	Y	2MM	N	1...mar	Simon is unhappy with						
11	386527438	136479	Y	M	Y	12	1N	\$null\$	Y	11Y	N	1...mar	Mr bobsman rand after						
12	386531436	124529	Y	M	Y	9	1N	\$null\$	N	\$null\$	N	1...mar	Cust needs to complai						
13	386534436	112019	N	M	Y	8	2N	\$null\$	N	\$null\$	N	2...mar	Comp logged for valda						
14	386538437	162777	Y	F	Y	10	1N	\$null\$	N	\$null\$	N	2...mar	Customer advised that						
15	386541449	135680	N	M	N	\$null\$	\$null\$Y	0D	Y	12Y	N	2...mar	Mr bobil is upset that n						
16	386544432	110546	N	M	Y	0	1N	\$null\$	Y	0MM	N	2...mar	The customer is unha						
17	386545432	110557	Y	F	Y	0	1N	\$null\$	N	\$null\$	N	2...mar	Cust has had net fault						
18	386548436	133925	Y	F	Y	9	4Y	0B	Y	1MM	N	2...mar	Customer has had no						
19	386551436	124387	N	M	Y	9	1N	\$null\$	N	\$null\$	N	2...mar	Customers nephew ca						
20	386552437	161711	Y	M	Y	11	1N	\$null\$	Y	0MM	N	2...mar	Cust has had to wait fo						

22. Close the **Table** window.

Now you are ready to produce some reports.

Task 2. Creating reports and graphs from the results.

1. Add a **Matrix** node from the **Output** palette to the stream to the right of the **Merge** node.
2. Connect the **Merge** node to the **Matrix** node, and then edit the **Matrix** node.
3. Beside **Rows**, select **Concept_not satisfied**.
4. Beside **Columns**, select **Gender**.
5. On the **Appearance** tab, select **Percentage of column**, and then click **Run**.

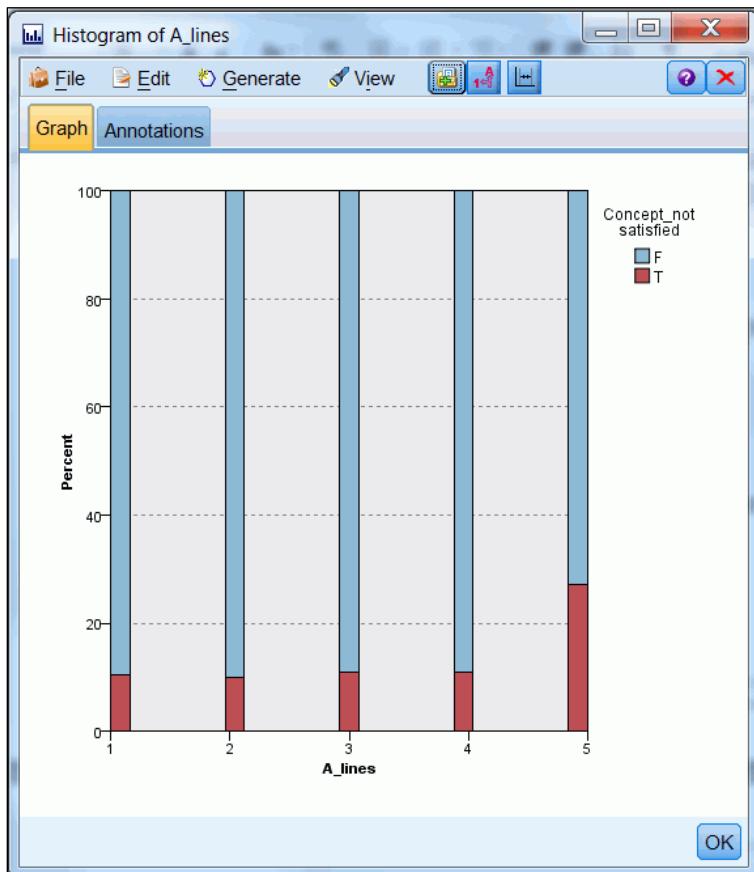


The results indicate that Male customers (11.1%) were slightly more dissatisfied with Astroserve than their Female counterparts (8.7%).

6. Close the **Matrix** dialog box.
7. Add a **Histogram** node from the **Graphs** palette to the stream, and then connect the **Merge** node to the **Histogram** node.
8. Edit the **Histogram** node.
9. Beside **Field**, select **A_lines**.
10. Under **Overlay**, beside **Color**, select **Concept_not satisfied**.

11. On the **Options** tab, select **Normalize by color**, and then click **Run**.

The results appear as follows:



The results show that for some reason, customers with the most land lines were more dissatisfied with Astroserve than customers with fewer lines.

12. Click **OK** to close the **Histogram** output window.
13. Click **File** and then **Save Stream As**.
14. Name the stream **Concepts Model**.
15. Click **Save**.
16. From the **File** menu, click **Exit**, and then click **Exit** again to end the Modeler session.

Results:

You examined the relationship between the concept data and the demographic information you have on the customers. One thing you learned from the analysis is that gender is not a very good predictor of churn. Also, you discovered that customers with more land lines tend to be less satisfied with Astroserve than those with fewer lines.

Apply Your Knowledge

Purpose:

Test your knowledge of the material covered in this module.

Question 1: True or False: Concept models allow you to score data as fields or records.

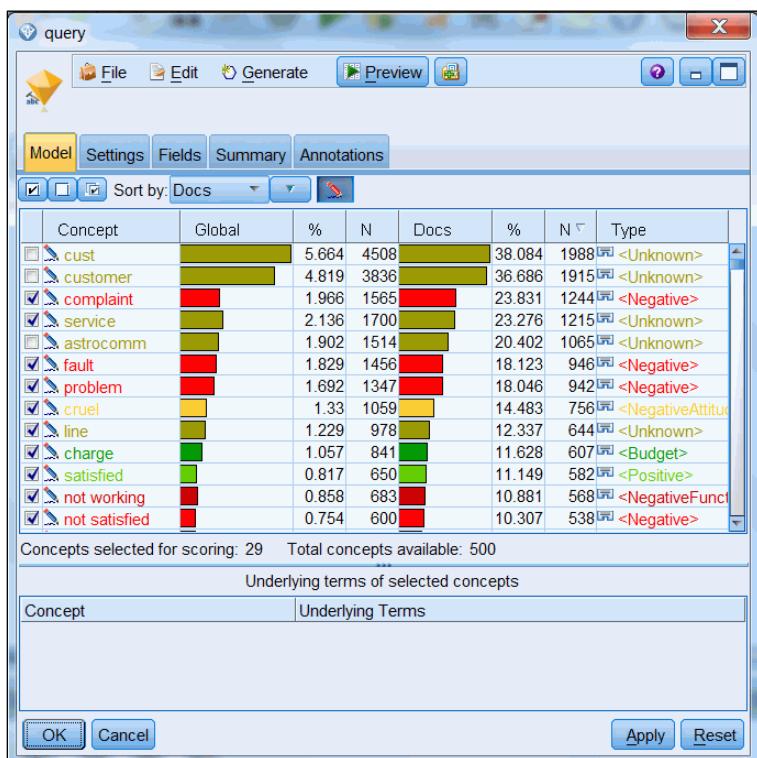
- A. True
- B. False

Question 2: True or False: Templates extract the same concepts just the frequency is different?

- A. True
- B. False

Question 3: In the concept model table, what percent of customers reported that something was not working?

- A. .858%
- B. 10.88%



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Question 4: True or False: In text mining, the word Concept is synonymous with the word Term.

- A. True
- B. False

Question 5: What scoring mode should you use if you want to create a table that rank orders the concepts by percentage?

- A. Concepts as fields
- B. Concepts as records

Apply Your Knowledge - Solutions

- Answer 1: A. True
- Answer 2: B. False. While you may get similar results with a different template, the frequencies probably will not be the same.
- Answer 3: B. 10.88%. This percentage is the percentage of Documents (customers) who reported that something was not working. The Global percentage of .858% is based on the total number of concept occurrences. The global numbers take into account that certain customers mentioned the same concept more than once, as opposed to the document numbers which only measure whether or not they mentioned the concept.
- Answer 4: B. False. Terms are words that are initially identified as relevant words by the extraction engine. Once these relevant terms are extracted, linguistic resources are used to group these terms along with similar terms under a lead term called a concept. While it is possible that in some instances a term and a concept are the same, this will not hold in general.
- Answer 5: B. Concepts as records.

The background of the slide shows a screenshot of the IBM SPSS Text Analytics software interface. A large, bold title 'Summary' is displayed at the top left. Below it, there is a bulleted list of learning objectives. In the top right corner of the slide, the 'IBM' logo is visible.

Business Analytics software

Summary

- At the end of this module, you should be able to:
 - develop a text mining concept model
 - compare models based on using different resource templates
 - score text data
 - analyze model results

© 2014 IBM Corporation

Business Analytics software

IBM

Workshop 1

Creating a Text Mining Concept Model



© 2014 IBM Corporation

The following file will be used:

- C:\Train\0A105\Music_Survey.sav - a Statistics file containing customer likes and dislikes about portable music players.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

5-45

Workshop 1: Customer Likes and Dislikes about Portable Music Players

In this workshop, you will learn what customers like and dislike about portable music players.

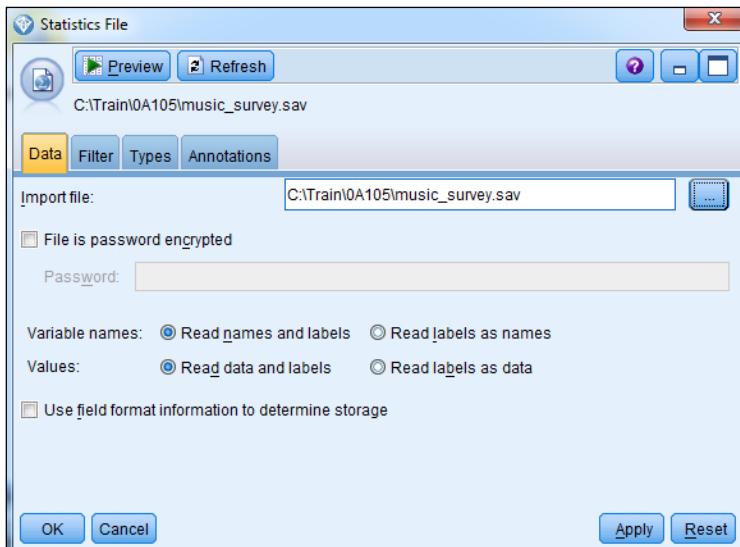
- Start a new stream.
- Import the data from C:\Train\0A105\music_survey.sav (a Statistics file).
- Add a Text Mining node downstream from the Statistics import node.
- Import the following field, "What do you like most about this portable music player."
- Create a concept model using the Basic Resources (English) template.
- Next create a concept model using the Product Satisfaction Opinions (English) template.

Which model do you prefer and why? Explore the results and get familiar with the concept model.

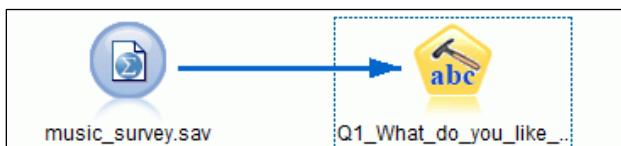
Workshop 1: Tasks and Results

Task 1. Creating a Concept Model.

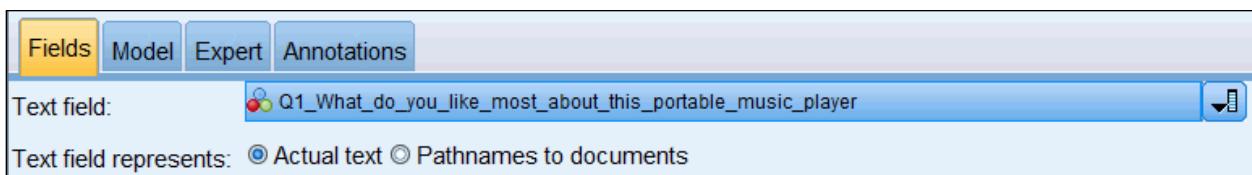
- Start a new stream.
- Add a **Statistics File** to the stream, and then import the data from C:\Train\0A105\music_survey.sav.



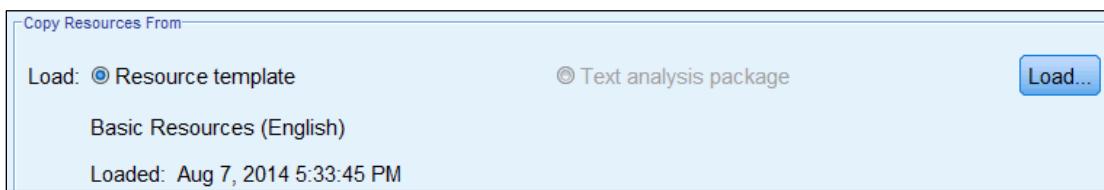
- Add a **Text Mining** node downstream from the Statistics import node.



- Import the following field, "What do you like most about this portable music player".



- Create a concept model using the **Basic Resources (English)** template.



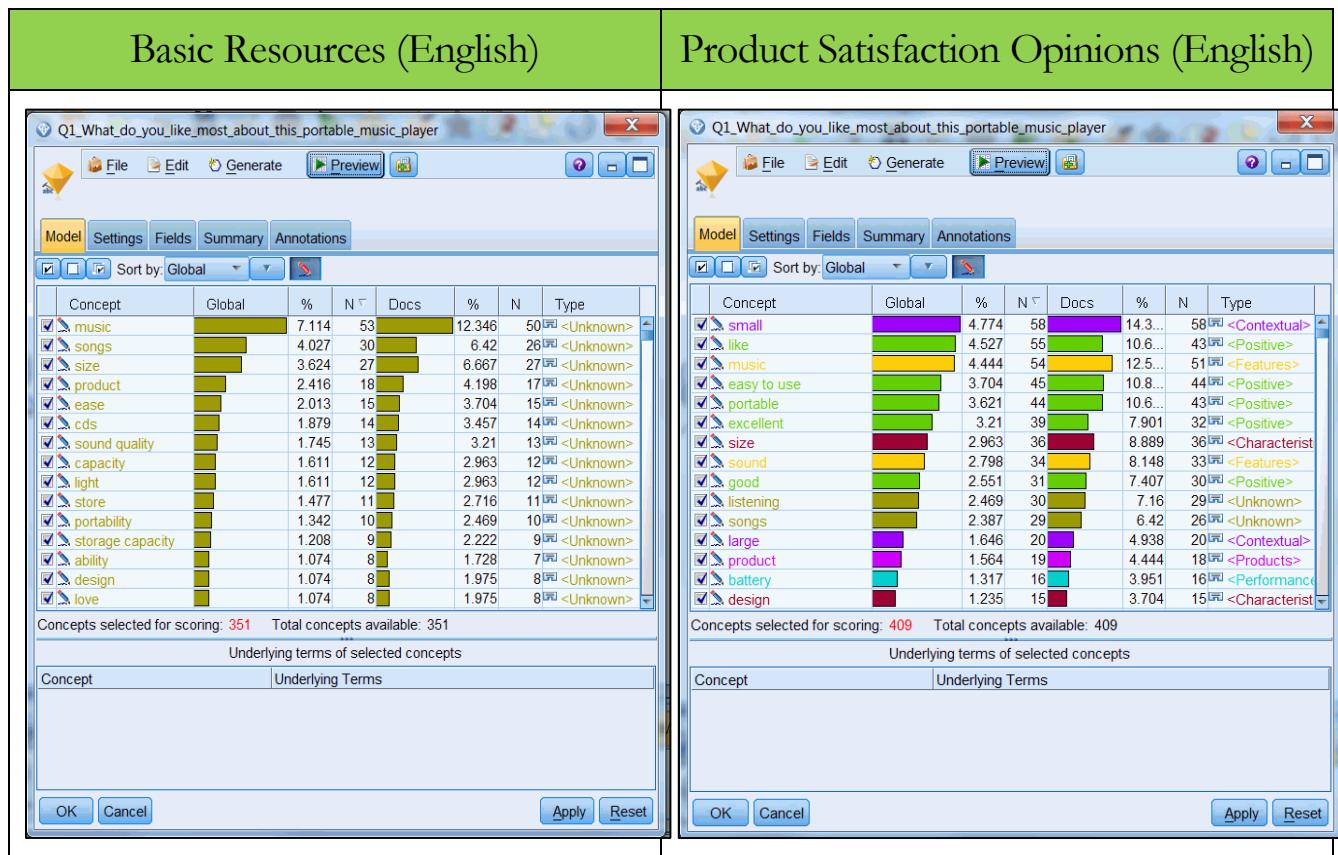
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

- Next create a concept model using the **Product Satisfaction Opinions (English)** template.
 - Click the **Load** button
 - Select **Product Satisfaction Options (English)** from the list of templates
 - Click **OK**

The results appear as follows:



- Which model do you prefer and why? Explore the results and get familiar with the concept model.

While the templates extracted similar concepts there is a big difference in how the concepts were typed by the two templates. All the concepts in the model that used the Basic Resources (English) template were typed as Unknown. In contrast, several of the concepts were typed as Positive in the model that used the Product Satisfaction Opinions (English) template which makes it much easier to pick out the characteristics that customers liked most about portable music players.



Reviewing Types and Concepts in the Interactive Workbench

IBM SPSS Modeler Text Analytics (v16)

Business Analytics software



© 2014 IBM Corporation

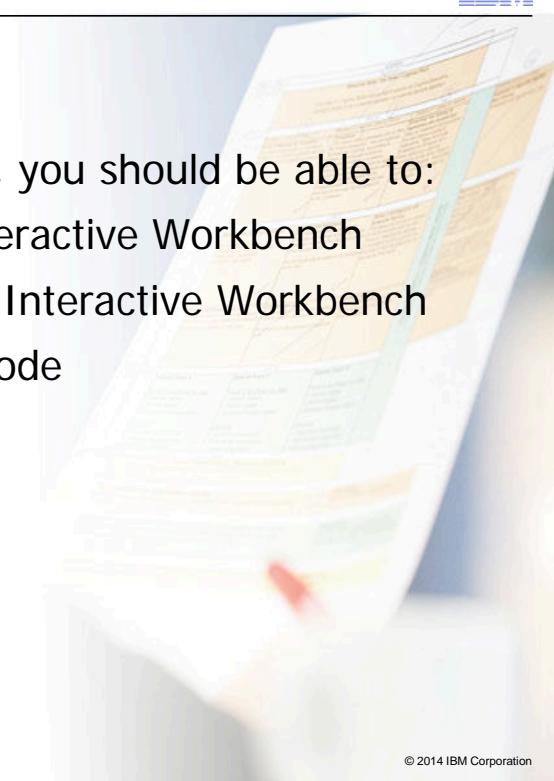
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Objectives

- At the end of this module, you should be able to:
 - review types in the Interactive Workbench
 - review concepts in the Interactive Workbench
 - update the modeling node



© 2014 IBM Corporation

One method of building text mining models as has been shown in the previous module is to build the model automatically from the text mining node. As an alternative, you have the option of launching an Interactive Workbench session. In the workbench, you work with extracted concepts and perform interactive text mining. You can create a categories-based text-mining model, and you can also explore the text with clustering and text link analysis. Most importantly, it is in the workbench where you edit the linguistic resources to tune the dictionaries specifically for a particular set of text data. You then re-extract and re-categorize the data to see the effect of modifying the resources. This feature makes the workbench environment truly interactive.

The Interactive Workbench was used briefly in an earlier module, in the quick example of text mining in Modeler. This module will begin by using a more in depth use of the workbench by using and modifying the extracted results.

Interactive Workbench Views

- The Interactive workbench window has four different view:
 - Categories and Concepts view
 - Text Link Analysis view
 - Clusters view
 - Resource Editor view

© 2014 IBM Corporation



The Interactive Workbench window has four different views for different types of analysis or for editing the dictionary resources. You can start your session on the Model tab in Categories and Concepts, Text Link Analysis or Clusters view. From any one of these three views you can switch to another view and perform that type of analysis. Thus, the mode that you use to launch the Interactive Workbench does not restrict you in how you choose to analyze the text data. The fourth view, the Resource Editor, cannot be opened until after the interactive session has begun.

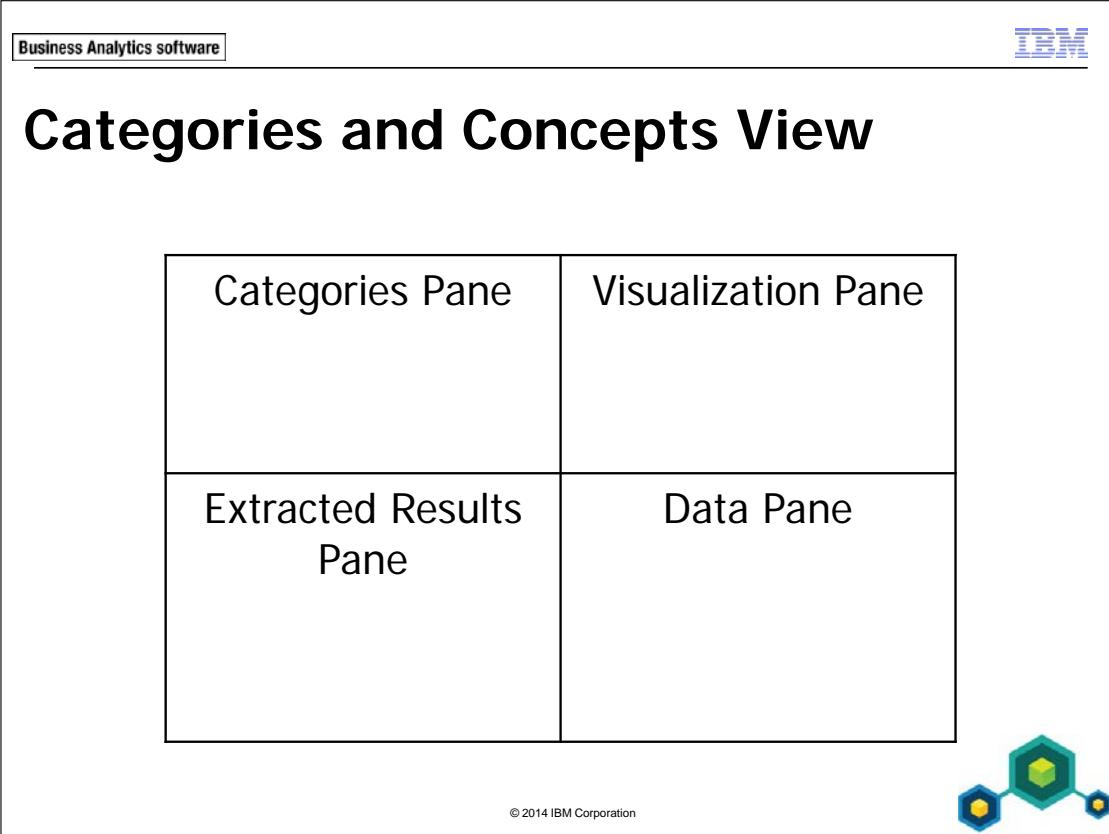
- Categories and Concepts view: (Using extraction results to build categories): This choice performs an extraction and opens the workbench in the Categories and Concepts view. By default, all interactive sessions begin in this view. The Categories and Concepts view is the window in which you can create and explore categories as well as explore and tweak the extraction results. Categories refer to a group of closely related ideas and patterns to which documents and records are assigned through a scoring process. Concepts refer to the most basic level of extraction results available to use as building blocks, called descriptors, for your categories.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

- Text Link Analysis view: (Exploring text link analysis (TLA) results): Used to discover relationships between the concepts within the text. In order to use this option, you must select a resource template that has pattern rules or have already created TLA pattern rules in the linguistic resources. Patterns can be used to create categories. Patterns are most useful when you are attempting to discover relationships between concepts or opinions about a particular subject. Some examples include wanting to extract opinions on products from survey data, genomic relationships from within medical research papers, or relationships between people or places from intelligence data. Once you have extracted some TLA patterns, you can explore them in the Data or Visualization panes and even add them to categories in the Categories and Concepts view. There must be some TLA rules defined in the resource template or libraries you are using in order to extract TLA results.
- Clusters view: (Analyzing co-word clusters): Enables you to discover relationships between concepts by clustering the concepts based on the strength of the link value between them. The link value is based on their co-occurrence in a record or document, and you can use the clusters to create categories. The goal of clusters is to group concepts that co-occur together while the goal of categories is to group documents or records based on how the text they contain matches the descriptors (concepts, rules, patterns) for each category. The more often the concepts within a cluster occur together coupled with the less frequently they occur with other concepts, the better the cluster is at identifying interesting concept relationships. Two concepts co-occur when they both appear (or one of their synonyms or terms appear) in the same document or record.

- Resource Editor view: Opens an environment in which the linguistic resources can be modified. You can view and fine-tune the linguistic resources used to extract concepts, group them under types, discover patterns in the text data, and much more. The operations that you perform in the Resource Editor view revolve around the management and fine-tuning of the linguistic resources. These resources are stored in the form of templates and libraries. The Resource Editor view is organized into four parts: Library Tree pane, Type Dictionary pane, Substitution Dictionary pane, and Exclude Dictionary pane. Because these resources may not always be perfectly adapted to the context of your data, you can create, edit, and manage your own resources for a particular context or domain in the Resource Editor. To simplify the process of fine-tuning your linguistic resources, you can perform common dictionary tasks directly from the Categories and Concepts view through context menus in the Extraction Results and Data panes.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE



The Interactive Workbench window is organized in a set of panes, each of which can be hidden, or resized. The panes differ depending on the selected view.

The Extracted Results pane is located in the lower left corner and displays the extracted concepts and types. The Data pane, located in the lower right corner, is used to view the text data corresponding to selections in the other panes. The Data pane is not populated automatically but will be when the Display button in another pane is clicked. The Categories pane, located in the upper left corner, presents the categories that have been created along with their frequency in the text. The categories can be managed and edited from this pane. The Visualization pane, located in the upper right corner, provides various graphical representations of categorization. The graphs include a bar chart of categories, a Web graph showing category relationships, and a table displaying the same in a more traditional format. As with the Data pane, the visual display corresponds to what is selected in the other panes.

Reviewing Extracted Concepts

- Working with large data sets can produce a list of thousands of extracted concepts.
- The high volume of concepts makes the task of going through them overwhelming.

© 2014 IBM Corporation



When you are working with very large datasets, the extraction process could produce millions of results. For many users, this amount can make it more difficult to review the results effectively. By default, only the top 5,000 concepts are displayed by their document frequency but in all likelihood, probably far more than 5,000 concepts were extracted. It would be nearly impossible to go through all these concepts one-by-one to find the ones you want.

You can increase the display limit in the Filter dialog, which can be found in the Tools menu. However, this is usually not desirable because 5,000 concepts is already difficult to go through to begin with. Also, the concepts you would be adding are the ones that occur less often and so may not be of as much interest.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

6-8

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Filter Concepts Dialog

- Filtering reduces the number of concepts displayed in the extraction pane.
- Filtering allows you to focus your attention on the concepts that most interest you.

© 2014 IBM Corporation



In order to zoom in on those that are most interesting, it helps to filter these results through the Filter dialog box available in the Extraction Results pane. You may want to reduce this number further, or filter the concepts by other criteria.

There are four choices available to filter the concepts.

- By Frequency: Concepts can be filtered by their global and/or document frequency to display only concepts that occur more often than a specified frequency.
- By Type: Concepts can be filtered by their type. You can choose more than one type to display.
- By Match Text: Concepts can be filtered by text that they contain. For example, you may request a match on the text "serv" to display all instances of concepts with the word "service" or variants of it. There are several match conditions available.
- By Rank: Concepts can be filtered based on global or document frequency. Using this option, you can display only the top number of frequently occurring concepts.

These settings are used together, so they are cumulative. Thus, if you select to display by Organization type and where the global frequency is greater than 100, then only concepts that meet both these restrictions will be displayed.

Updating the Text Mining Node

- It usually is not possible to complete the changes you need to make to the Resource Editor in one session.
- If you exit from the Interactive Workbench without updating the Text Mining node, you will lose all the changes you just made.
- This would include any new types or synonyms you just made, any new libraries you just added, etc.
- Updating the Text Mining node stores all the changes you made for later work.

© 2014 IBM Corporation



Although the Interactive Workbench is a special environment, it is still another type of output so far as Modeler is concerned. There are three choices of how to proceed when you exit from the Interactive Workbench:

- Update: This option allows you to first save the work back into the originating modeling node for future sessions. After updating the Text Mining node, the session window is closed, and the session is deleted from the Output manager in the Modeler window.
- Exit: This option will discard any unsaved work, close the session window, and delete the session from the Output manager in the Modeler window.
- Close: This option will not save or discard any work. This option closes the session window but the session will continue to run. You can open the session window again by selecting this session in the Output manager in the Modeler window.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

In other words, you can close an interactive session but leave it available from within a Modeler session for later work. This saves time in re-extracting the data. One disadvantage of doing so is that interactive sessions use lots of memory, and this option will not free up that memory. It is certainly not recommended having multiple interactive sessions open, which is possible, without having at least 2 GB of memory on the computer.

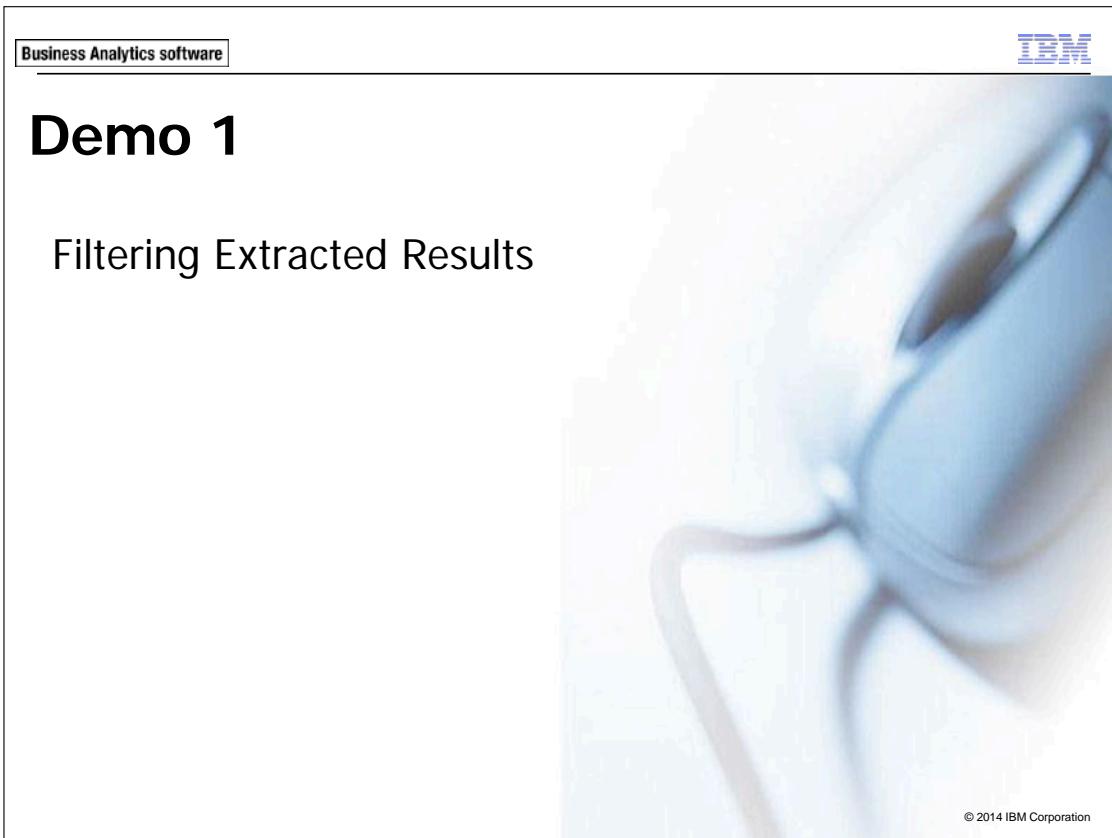
Saving the workbench information back into the modeling node does not create a model. Instead, it updates the modeling node so you can later recreate the exact state of the results, including the categories that you created, with the current data or a new data file. But updating a modeling node is not creating a Text Mining model. That has to be done from within the Interactive Workbench.

When you click Update, a second dialog appears asking whether you want to update the Text Mining node and publish the libraries in the project. Doing so will allow you to share the libraries from this session with other projects that have similar text data.

There are two methods to update the modeling node:

- Keep only session work: This option retains the changes you made to the linguistic resources, text link analysis rules, or any categories that were created during the interactive session. These changes are added to the Text Mining node so they will be available the next time you run that node to create a model or work in the Interactive Workbench.
- Keep session work and cache text data: This option does the same as the first, with the addition of caching the text data itself, along with the extracted results, in the Text Mining node. Using this option makes it very quick to return to a previous interactive session without having to re-extract the data.

The node can also be updated from within the Interactive Workbench session by selecting File and then Update Modeling Node from the menus.



The slide has a light blue background with a faint, abstract graphic of a person's head and shoulders. In the top left corner, there is a small rectangular box containing the text "Business Analytics software". In the top right corner, the "IBM" logo is displayed. The main title "Demo 1" is centered at the top in a large, bold, black font. Below it, the subtitle "Filtering Extracted Results" is also centered in a smaller, regular black font. In the bottom right corner of the slide area, there is a small copyright notice: "© 2014 IBM Corporation".

This demo uses the following datasets coming from a (fictitious) telecommunications firm.

- C:\Train\0A105\06-Reviewing Types and Concepts in the Interactive Workbench\Concepts_Model_Demo1_start.str - a Modeler stream that reads a file containing call center data for March and April

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Demo 1: Filtering Extracted Results

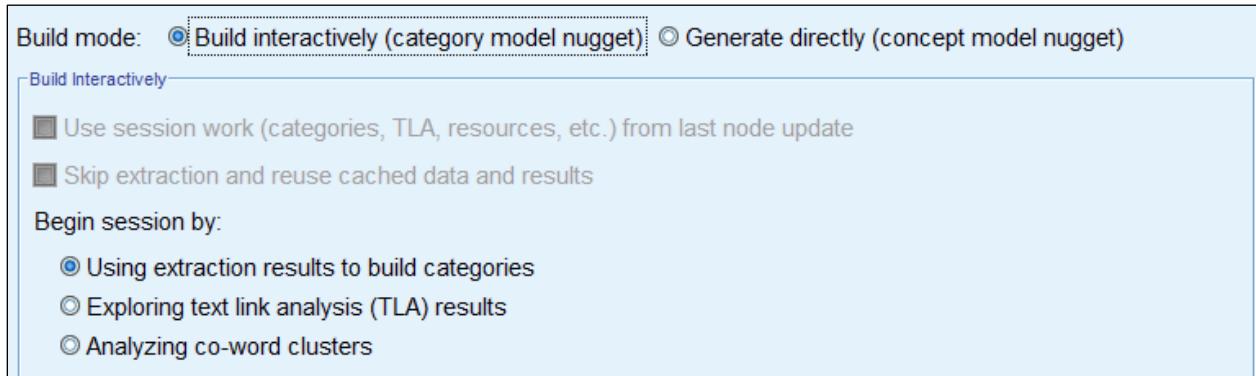
Purpose:

You are in the Interactive Workbench and would like to examine what phone customers are saying about their service. Because there are so many concepts in the Extraction pane, you need an easy way to display only the customer records that mentioned the word "service" and exclude all the rest from the Extraction pane.

Task 1. Starting the interactive session.

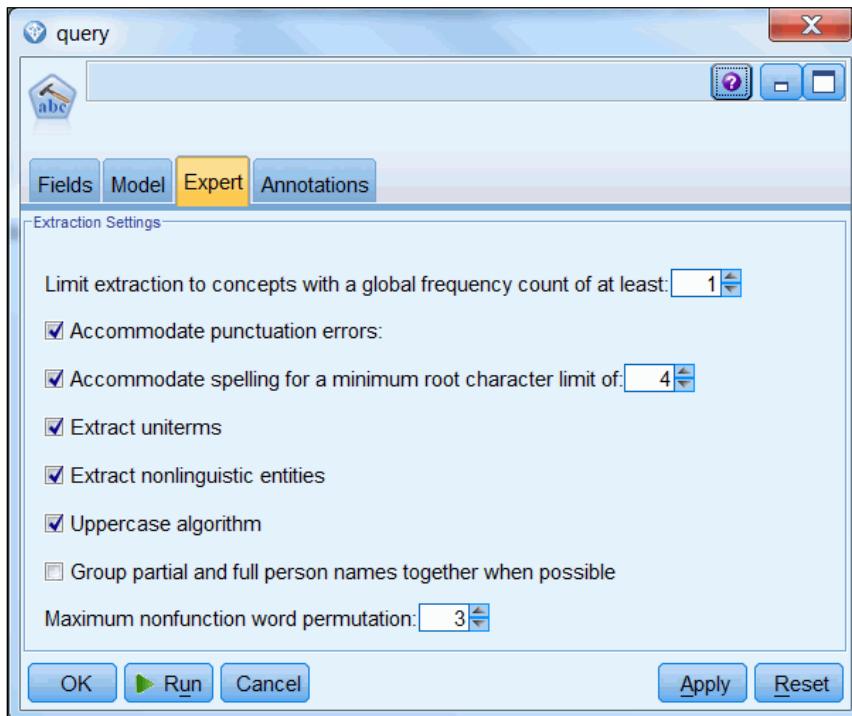
1. From the **File** menu, click **Open Stream**.
2. Navigate to **C:\Train\0A105\06-Reviewing Types and Concepts in the Interactive Workbench** and double-click **Concepts_Model_Demo1_start.str**.
3. Edit the **Text Mining** modeling node.
4. Click the **Model** tab, and ensure that **Build Interactively (category model nugget)** is selected.
5. Click **Using extraction results to build categories** in the **Begin session by** area (if necessary).

This setting will open the Interactive Workbench in Categories and Concepts view.



6. Click the **Expert** tab.

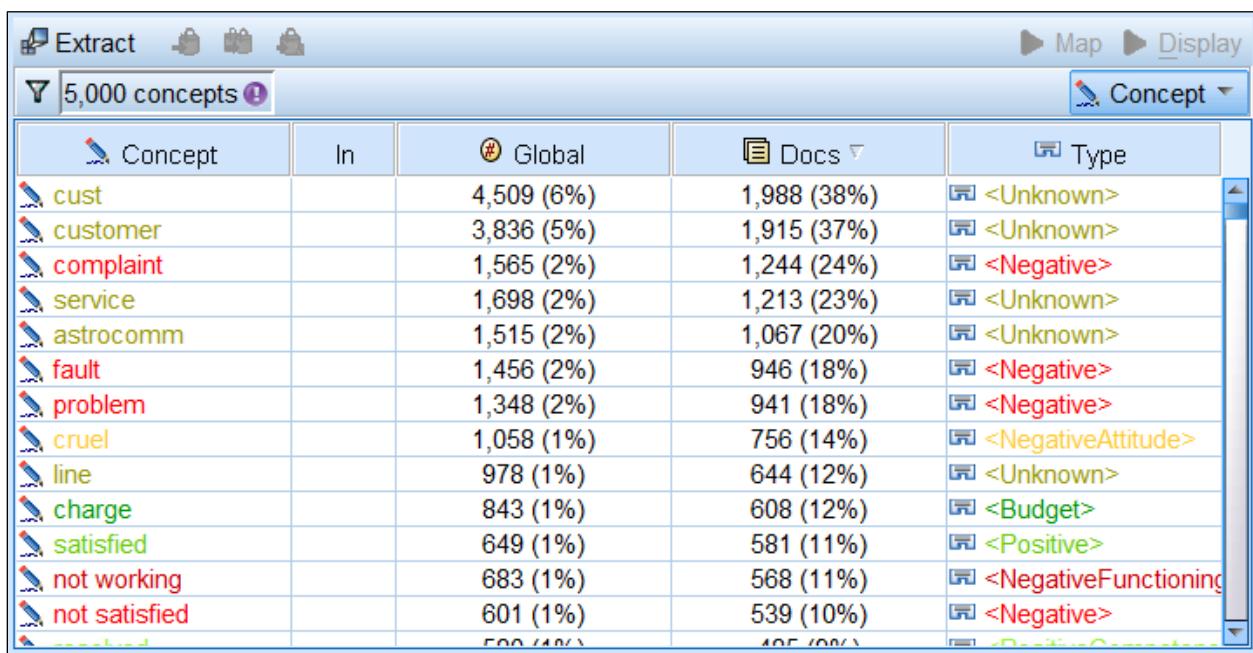
The following settings will be used when the Interactive Workbench is launched to extract text.



7. Click **Run**.

When the workbench is launched, you will receive a feedback window as the text is extracted. After it has completed, you will see the Interactive Workbench in Categories and Concepts view. You will focus on the Extraction pane in the lower left corner.

The results appear as follows:

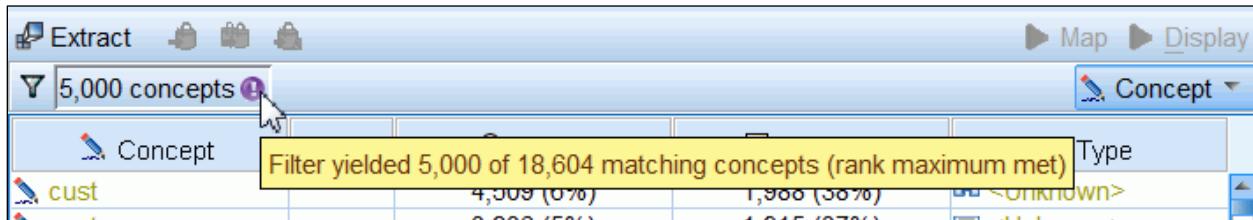


A screenshot of the IBM SPSS Interactive Workbench Categories and Concepts view. The interface has a toolbar at the top with icons for Extract, Map, and Display. Below the toolbar is a header bar with a dropdown menu set to 'Concept' and a count of '5,000 concepts'. The main area is a table with five columns: Concept, In, Global, Docs, and Type. The table lists 15 concepts, each with a small icon to its left. The 'Global' column shows the count and percentage of documents containing each concept. The 'Docs' column shows the count and percentage of documents for each concept. The 'Type' column indicates the category of each concept. The table is scrollable, with a vertical scrollbar on the right side.

Concept	In	Global	Docs	Type
cust		4,509 (6%)	1,988 (38%)	<Unknown>
customer		3,836 (5%)	1,915 (37%)	<Unknown>
complaint		1,565 (2%)	1,244 (24%)	<Negative>
service		1,698 (2%)	1,213 (23%)	<Unknown>
astrocomm		1,515 (2%)	1,067 (20%)	<Unknown>
fault		1,456 (2%)	946 (18%)	<Negative>
problem		1,348 (2%)	941 (18%)	<Negative>
cruel		1,058 (1%)	756 (14%)	<NegativeAttitude>
line		978 (1%)	644 (12%)	<Unknown>
charge		843 (1%)	608 (12%)	<Budget>
satisfied		649 (1%)	581 (11%)	<Positive>
not working		683 (1%)	568 (11%)	<NegativeFunctioning>
not satisfied		601 (1%)	539 (10%)	<Negative>
...		500 (1%)	405 (8%)	<Unknown>

In the upper left corner of the window, it reads 5000 concepts. By default, only the top 5,000 concepts are displayed by their document frequency. This limit exists because in large files with rich text, you can easily extract tens of thousands of individual concepts, which can be overwhelming to review and use.

8. Point to the  symbol to get an exact number of concepts that were extracted.



Notice that the total was 18,604 concepts. This is far too many concepts to go through manually. You may want to reduce this number further, or filter the concepts by other criteria. Filtering is available from the dialog box of that name.

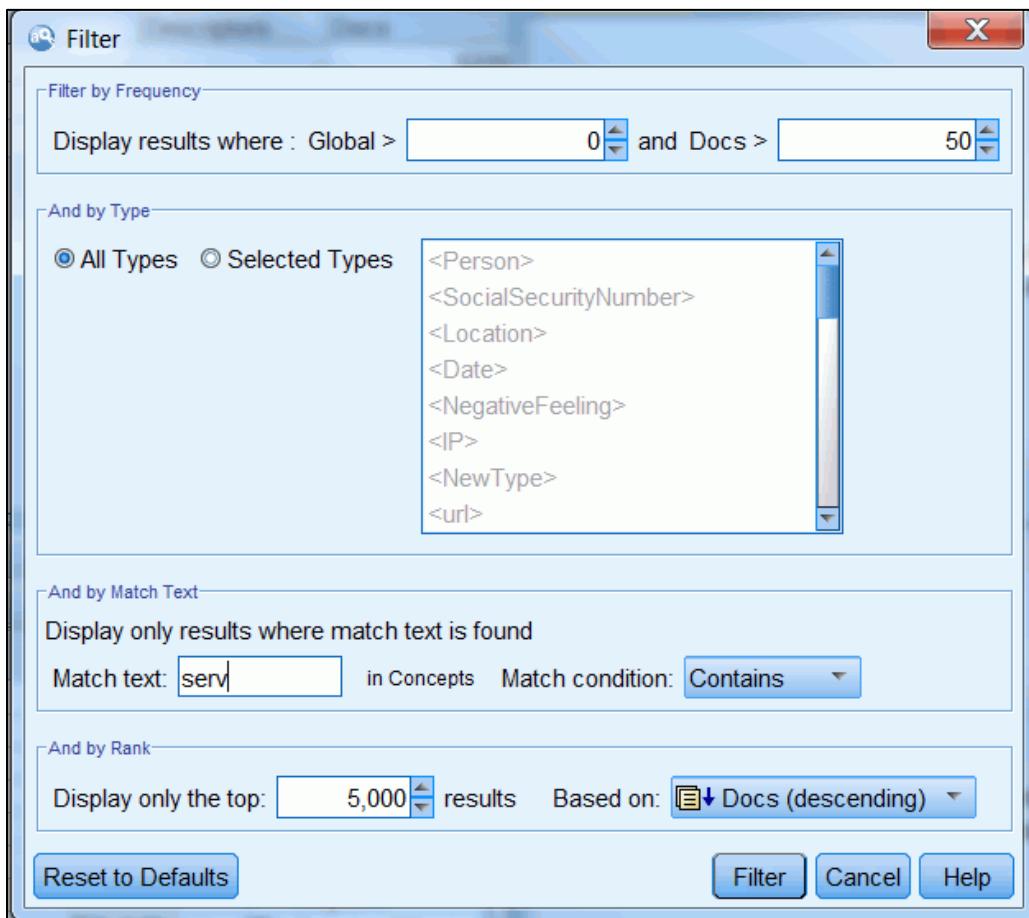
Task 2. Filtering to focus on a single concept.

- From the **Tools** menu, click **Filter**.

Alternatively you could click on the Filter button  in the upper left corner of the window.

- Beside **Display results where**, set the **Docs >** value to **50**.
- Beside **Match text**, type **serv**.

Notice at the bottom that there is an option to change the display number to something other than 5,000.



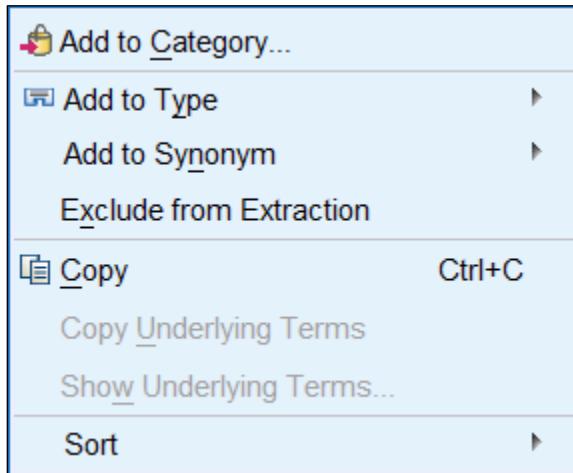
4. Click Filter.

The Extracted Results pane now displays only eight concepts. All have the text string "serv" included in their text at some point, and all occur in at least 51 records.

Concept	In	Global	Docs (>50)	Type
service		1,698 (2%)	1,213 (23%)	<Unknown>
astroserve		488 (1%)	431 (8%)	<Unknown>
service number		75 (0%)	68 (1%)	<Unknown>
supervisor astroserve		66 (0%)	66 (1%)	<Unknown>
phone service		64 (0%)	63 (1%)	<Unknown>
service from astrocomm		60 (0%)	60 (1%)	<Unknown>
customer service		55 (0%)	55 (1%)	<Unknown>
mobile service		58 (0%)	55 (1%)	<Unknown>

Now you can easily find all the concepts that have the string "serv". Notice that two, "astroserve" and "supervisor astroserve", refer to the company and have nothing at all to do with service. You can instruct Modeler that you do not want to extract these two concepts.

5. Right-click the **astroserve** concept.



The Exclude from Extraction option instructs Modeler to do exactly that. Notice the two choices on the context menu that allow you to easily edit the linguistic resources by adding this concept as a synonym or to a type.

6. Select **Exclude from Extraction**.

The results appear as follows:

Concept	In	Global	Docs (>50)	Type
service		1,698 (2%)	1,213 (23%)	<Unknown>
astroserve		488 (1%)	431 (8%)	<Unknown>
service number		75 (0%)	68 (1%)	<Unknown>
supervisor astroserve		66 (0%)	66 (1%)	<Unknown>
phone service		64 (0%)	63 (1%)	<Unknown>
service from astrocomm		60 (0%)	60 (1%)	<Unknown>
customer service		55 (0%)	55 (1%)	<Unknown>
mobile service		58 (0%)	55 (1%)	<Unknown>

The concept "astroserve" did not disappear; nor is it specially marked. Instead, the background of the Extraction pane has turned yellow. This is a visual indicator that a) the linguistic resources have been modified, or b) changes have been made to the extracted concepts. In either case, to display the correct list of extracted concepts, the text data need to be re-extracted.

7. From the **Tools** menu, click **Extract**.

The results appear as follows:

Concept	In	Global	Docs (>50)	Type
service		1,698 (2%)	1,213 (23%)	<Unknown>
service number		75 (0%)	68 (1%)	<Unknown>
phone service		64 (0%)	63 (1%)	<Unknown>
service from astrocomm		60 (0%)	60 (1%)	<Unknown>
customer service		55 (0%)	55 (1%)	<Unknown>
mobile service		58 (0%)	55 (1%)	<Unknown>

Only six concepts are displayed after the extraction is completed. Astroserve has disappeared, but so too has the concept "supervisor astroserve". This is because it also contains the word "astroserve", and any concept containing an excluded word is itself excluded.

When you are reviewing the extracted concepts you will invariably want to review some of the text comments that contained a concept. You will select the concept phone service, which occurs in 63 records.

8. Click concept **phone service** to select it.

9. Click **Display**.

The results appear as follows:

	query (63)	Categories	Text Preview
1	Cust is upset that her phone was out for a week and she is being charged for outgoing mobile calls and now her phone service is out again.		Cust is upset that her phone was out for a week and she is being charged for outgoing mobile calls and now her phone service is out again.
2	Customer has complained that he experiences difficulty with the eftpos service on his business line. when he attempts to settle his transactions for the day the screen display on his machine disappear...		
3	Customer user name solar email address sbob@astronet.com.au customer claims he churned back from optitel to ASTROCOMM as he was advised that he would now only receive a singel bill for the DSL and no...		

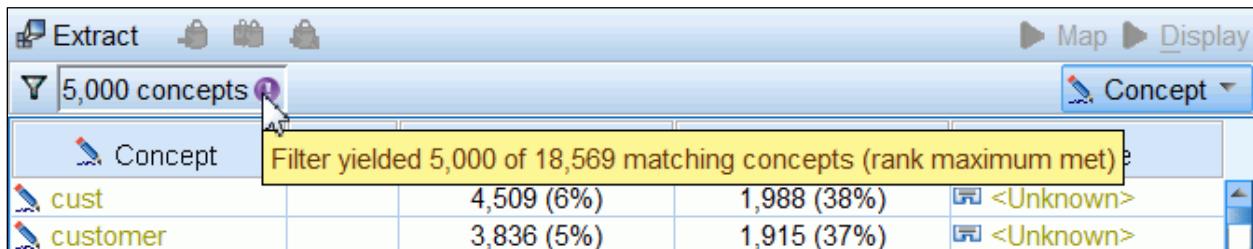
The first record contains the text "phone service". The text associated with the concept is highlighted in yellow. Other text is in various colors, corresponding to a specific type to which that text has been assigned. Thus the negative type is in a shade of red, while the Budget and Positive types are in green.

You certainly cannot view the actual text data for most, or even a substantial fraction of, the concepts. That would take too long. Instead, use the display of data selectively for important concepts, for concepts that you need to understand better (such as the context in which they are used), and to help you decide how to edit the linguistic resources.

10. From the **Tools** menu, click **Filter**.
11. Change the **Docs >** value to **0**.
12. Delete the text **serv** in the **Match text** box.

13. Click Filter.

By removing the filter, the Extracted Results pane will again display the top 5,000 concepts although the total number of concepts extracted has been reduced from 18,604 to 18,569 as a result of excluding the concept "astroserve". The results appear as follows:



14. From the **File** menu, click **Close**, and then click **Update** to save your work. If you do not update the Text Mining node before exiting the Interactive Workbench, astroserve will no longer be excluded from extraction the next time the stream is run.
15. Click **OK** to **Keep only session work (categories, patterns, resources, etc.)**.
16. Click **OK** to the informational message about the node being updated.
17. From the **File** menu, click **Save Stream As**.
18. Name the stream **Concepts_Model_Demo1_end.str**, and then click **Save**.
19. From the **File** menu, click **Close Stream**.
20. From the **File** menu, click **New Stream**.

Do not close Modeler. Leave it open for the next demo.

Results:

You reviewed what customers are saying about their service. You displayed only the customer records that mentioned the word "service" and excluded all others from the Extraction pane.

Reviewing Extracted Types

- The higher-level grouping can be useful in a variety of ways:
 - frequency of a type can be an indicator of whether the linguistic resources are going to do a good or bad job finding text for that type
 - because the concepts are grouped by types, it is possible to see general trends in the data before doing any categorization

© 2014 IBM Corporation



Types are semantic groupings of concepts, stored in the form of type dictionaries. Reviewing the frequency of a specific type can be a good indicator of general trends in the data. For example, the Opinions (English) resource template contains the following Types which are designed to capture positive, negative, and uncertain opinions.

Positive Types	Negative Types	Other Types
Positive	Negative	Uncertain
Positive Attitude	Negative Attitude	Contextual
Positive Budget	Negative Budget	
Positive Competence	Negative Competence	
Positive Feeling	Negative Feeling	
Positive Functioning	Negative Functioning	
Positive Feeling Emoticon	Negative Feeling Emoticon	

As an example, suppose you are analyzing call center data for a major software company. Given the nature of call center data, you would expect to get more negative than positive comments because customers generally call with a problem or a complaint. Thus, you would expect that the negative types should have higher frequencies than the positive types. If the ranking of these types was not consistent with expectations, you would definitely need to do a detailed review some of these comments to check on the extraction.

If the types are living up to expectations, then the frequencies should give you some idea of what customers are negative about. For instance, if the frequency for the Negative Functioning type is relatively high, it would strongly suggest that customers did not think the software worked very well. If Negative Competence got a high count, that would suggest that customers are not happy with the service that they are getting.

It is important to note that with the default resources, the vast majority of extracted concepts will probably be types as Unknown. In other words, text mining was unable to recognize a large percentage of words/terms it extracted. This is not a cause for concern. It may suggest that you need to create some new types to group terms together that share something in common with each other. In other instances, you should probably just leave them as Unknown. The fact that a term has been typed as Unknown is useful information. For instance, in Text Link Analysis, it might clue you into things that had not occurred to you before, if many of the concepts customers are negative about were typed Unknown.

It is also important to note that concepts do not have to be typed as something other than Unknown to be categorized. Typing has several purposes, and you will invariably create some types in projects, but most of the concepts you use in categories may well be typed as Unknown. Again, this is not a cause for concern.

Business Analytics software

IBM

Demo 2

Reviewing Extracted Types



© 2014 IBM Corporation

This demo is a continuation of Demo 1. If you have not completed Demo 1, a start file has been provided for you.

This demo uses the following Modeler stream.

- C:\Train\0A105\06-Reviewing Types and Concepts in the Interactive Workbench\Concepts_Model_Demo2_start.str - a Modeler stream that reads a file containing call center data for March and April

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

6-23

Demo 2: Reviewing Extracted Types

Purpose:

Along with reviewing the extracted concepts, you should also examine the extracted types to see if they are performing well and if you can see any general trends in the data.

Task 1. Re-starting the interactive session.

1. From the **File** menu, click **Open Stream**.
2. Navigate to **C:\Train\0A105\06-Reviewing Types and Concepts in the Interactive Workbench**, and then double-click **Concepts_Model_Demo2_start.str**.
3. Edit the **Text Mining** modeling node.
4. Click **Run**.

Task 2. Reviewing the extracted concepts and types.

1. In the top right corner of the **Extracted Results** pane, click the list, and then select **Type**.

Concept	In	Global	Docs	Type
cust		4,509 (6%)	1,988 (38%)	<Unknown>
customer		3,836 (5%)	1,915 (37%)	<Unknown>
complaint		1,565 (2%)	1,244 (24%)	<Negative>

The results appear as follows:

Type	In	Global	Docs
<Unknown>		60,048 (76%)	5,201 (100%)
<Negative>		8,888 (11%)	3,662 (70%)
<Budget>		5,860 (7%)	2,509 (48%)
<Positive>		3,741 (5%)	2,291 (44%)
<Contextual>		2,680 (3%)	1,708 (33%)
<Date>		2,357 (3%)	1,393 (27%)
<NegativeFunctioning>		1,874 (2%)	1,339 (26%)
<Person>		1,863 (2%)	1,210 (23%)
<NegativeAttitude>		1,329 (2%)	946 (18%)
<Period>		1,108 (1%)	870 (17%)
<PositiveCompetence>		906 (1%)	736 (14%)
<Currency>		531 (1%)	419 (8%)
<Uncertain>		368 (0%)	329 (6%)
<Organization>		351 (0%)	261 (5%)
<PositiveAttitude>		271 (0%)	239 (5%)
<email>		215 (0%)	200 (4%)
...

In the case of the Astroserve data, you can observe the following:

- As would be expected, the number of records with negative comments is greater than those with positive comments. This makes sense because customers are generally calling with a problem or complaint. If the ranking of these types was not consistent with expectations, you would definitely need to do a detailed review of the concepts included for each type.
- On the other hand, there are still many positive comments, almost too many. It will be worthwhile reviewing some of these comments to check on the extraction.
- An Organization type is only mentioned in about 5% of the records, but this seems to be too low. And this is because the linguistic resources are not tuned to recognize the particular organizations mentioned by Astroserve customers.
- As always with the default resources, the vast majority of extracted concepts are typed as Unknown (76% by global frequency). In other words, text mining was unable to recognize seven in ten of the words/terms it extracted.

You will review two of those positive comments.

2. Click the **Positive** type, and then click **Display**.
3. Click record 8.

The screenshot shows the IBM SPSS Modeler interface with the 'Text Analytics' node open. The 'Text Preview' pane on the right displays a snippet of text from record 8, which contains several highlighted words in yellow and green, such as 'faults', 'mobil', 'happy', 'dean', 'consultant', 'wanted', 'use', 'pay', 'phone', 'ordinary', 'advised', 'connected', 'two', 'sockets', 'premises', 'active', 'reqd', '3rd', 'socket', 'tech', 'attended', 'site', 'only', 'connected', 'one'. The text discusses a customer's service issue where they had two sockets in their premises and needed both active, requiring a 3rd socket to be installed because the tech attended site only connected to one.

query (1000 - Max)		Categories	Text Preview
7	Customers nephew called to enquire whwn his aunts service would be fixed . Advised Mr bobin () was not too happy . He says his aunt is 86 years old and lives alone and needs a...		Mobil cnt nbr cust has had two faults have div call to mobil cust not happy at paying mobil rates on mobil for outgoing call while there is a fault last time fault dean what told by consultant if wanted loc rates would need to use pay phone dean felt this was a bit ordinary
8	Mobil cnt nbr cust has had two faults have div call to mobil cust not happy at paying mobil rates on mobil for outgoing call while there is a fault last time fault dean what told by consultant if wan...		
9	Cust advised when asked for service connected that she had two sockets in the premises and needed both active and also reqd a 3rd socket to be installed. When tech attended site only connected to one...		

The yellow highlighted text is "happy", which is a word which could indeed be a positive comment or evaluation. However, the customer was not expressing satisfaction, but instead is "not happy" with paying while there is a problem. This is a typical example of initial extracted results not coding the data accurately.

4. Click record 21.

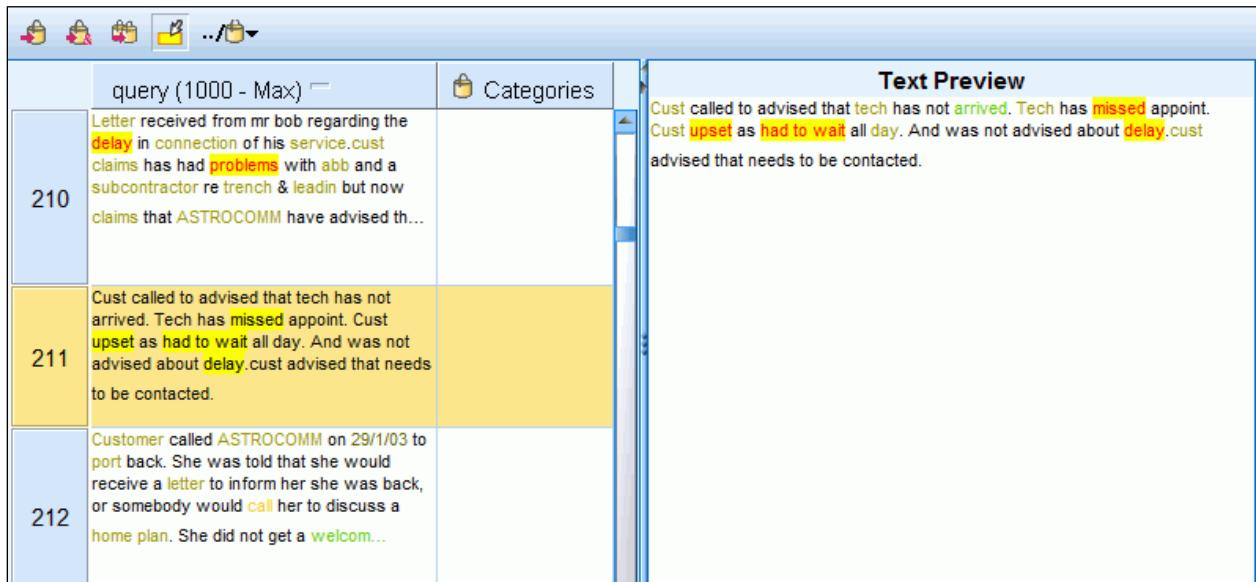
The screenshot shows the IBM SPSS Modeler interface with the 'Text Analytics' node open. The 'Text Preview' pane on the right displays a snippet of text from record 21, which contains several highlighted words in yellow and green, such as 'unhappy', 'ASTROCOMM', 'contractors', 'skilled', 'numbers', 'voice', 'fax', 'duet', 'seperate', 'independent', 'long', 'multiple', 'contact', 'order', 'reimbursement', 'months', 'free', 'internet', 'acceptable', 'disgust', 'competence', 'change', 'name', 'company', 'unskilled', 'techy', 'hooked', 'bargain', 'cheap', 'labor', 'skill', 'necessary', 'reimbursement'. The text discusses a customer's dissatisfaction with service offered by ASTROCOMM and its contractors, mentioning multiple lines and技工 issues.

query (1000 - Max)		Categories	Text Preview
20	Rcvd email, email adress: bremmimbob@astronet.com: kana case number 9942: subject: unhappy customer.to whom it my concern,iam writing in regards to my astronet account. On the 27th February i received...		Cust is unhappy about service offered by ASTROCOMM and its contractors skilled . Cust advised he was meant to have 2 numbers connected one a voice line (which was connected finally by ASTROCOMM and not skilled) and a fax line which is not yet connected. I advised cust that the fax line was a duet system with the main number. Cust advised he did not order that he ordered 2 seperate independent lines. Cust advised he is a long standing customer of ASTROCOMMs and has multiple lines with them. Cust wants contact by 04/03 or he is withdrawing all his accounts with ASTROCOMM. Cust is refusing to pay any of his connection fees until he is actually connected and wants reimbursement . Cust advised in the past for ASTROCOMMs stuff up they gave him a months free internet access which cust agreed was acceptable . Cust needs also to point out his disgust with the lack of skill and competence with the contractors skilled cust advised they should change the name of the company to unskilled cust advised he was happy that a ASTROCOMM techy hooked it up for him and that it should be done by ASTROCOMM (trained) techs and not bargain basement cheap labor that do not have the skill necessary to complete a job cust needs reimbursement
21	Cust is unhappy about service offered by ASTROCOMM and its contractors skilled. Cust advised he was meant to have 2 numbers connected one a voice line (which was connected finally by ASTROCOMM and no...		
22	Samantha rang about surname45 - she said that surname45 never signed a transfer of ownership to the company name on his mobile 0419xxxxxx & wanted it changed back to surname45 tonight. I explained th...		

In this instance, the term "happy" has been assigned to the Positive type, and that is clearly a correct classification. Of course, notice that this customer was actually very dissatisfied with the lack of skill of the contractors. This is typical for call center data, and in general for text data that is more than a few words in length. Normally text for one record or document can be assigned to several types, and eventually several categories.

Next you will examine how well text was classified into the Negative type.

5. Click the **Negative** type, and then click **Display**.
6. Click record 211.



The screenshot shows the IBM SPSS Text Miner interface. At the top, there are icons for saving, opening, and closing files. Below that is a toolbar with various buttons. The main area is divided into three columns: a left column with record numbers 210, 211, and 212; a middle column labeled 'Categories' containing short descriptions of the text; and a right column labeled 'Text Preview' containing a detailed summary of the extracted concepts for record 211. Record 211 is highlighted with a yellow background in both the categories and preview sections.

	query (1000 - Max)	Categories	Text Preview
210	Letter received from mr bob regarding the delay in connection of his service.cust claims has had problems with abb and a subcontractor re trench & leadin but now claims that ASTROCOMM have advised th...		Cust called to advised that tech has not arrived . Tech has missed appoint. Cust upset as had to wait all day. And was not advised about delay .cust advised that needs to be contacted.
211	Cust called to advised that tech has not arrived. Tech has missed appoint. Cust upset as had to wait all day. And was not advised about delay .cust advised that needs to be contacted.		
212	Customer called ASTROCOMM on 29/1/03 to port back. She was told that she would receive a letter to inform her she was back, or somebody would call her to discuss a home plan . She did not get a welcom...		

Four terms were identified here as negative in connotation: "missed", "upset", "had to wait", and "delay". Given that the data generally concern calls to Astroserve about repair, service, and other issues, when a comment is typed as Negative, that is very likely an accurate judgment.

These few examples have illustrated the power of extraction and the need to edit the results. Given how many concepts are extracted from typical text data, you will want to focus the work first and foremost on the concepts, categories and linguistic resources in those areas with a) more responses, and b) on those concepts and categories which are most crucial for business or organizational goals. Follow this approach in working with the Astroserve data.

7. From the **File** menu, click **Close**, and then click the **Exit** button.

Because you did not make any further changes to the Resource Editor, there is no need to update the Text Mining node again.

8. From the **File** menu, click **Save Stream As**.
9. Name the stream **Concepts_Model_Demo2_end.str**, and then click **Save**.
10. From the **File** menu, click **Close Stream**.
11. From the **File** menu, click **New Stream**.

Do not close Modeler. Leave it open for the next demo.

Results:

You reviewed the extracted concepts, and examined the extracted types to see if they performing well and identified general trends in the data.

Using an Updated Modeling Node

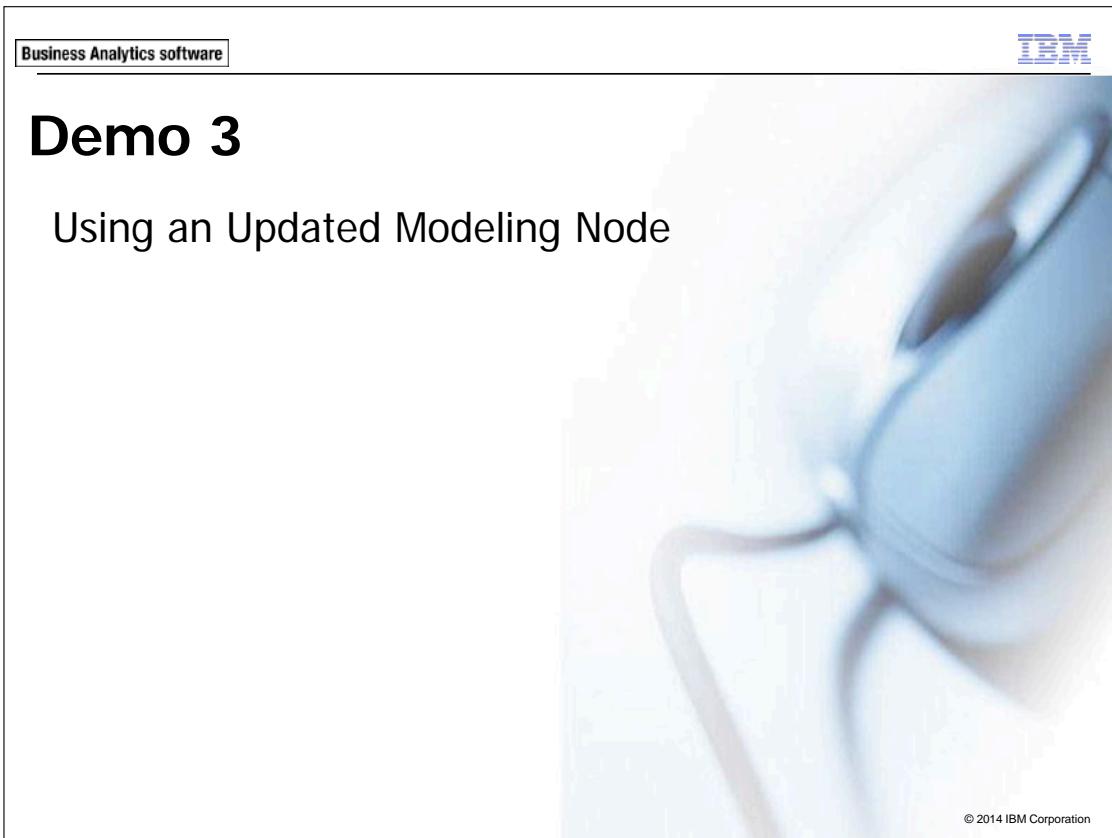
- There are two choices in the Interactive Workbench related to which option you chose to update the text mining modeling node:
 - use session work (categories, TLA, resources, etc)
 - skip extraction and reuse cached data and results

© 2014 IBM Corporation



There are two choices in the Interactive Workbench area that are related to the choice made in the Save and Exit dialog.

- "Use session work": This option is enabled if you chose to keep only session work in the Save and Exit dialog. When you launch a session with this option, the extraction settings, categories, resources, and any other work from the last time you performed a node update from an Interactive Workbench session are available and will be used. Since saved session data are used with this option, certain content, such as the resources copied from the template, and other tabs are disabled and ignored.
- "Skip extraction and reuse cached data": This option is enabled if you chose to keep the session work and cache text data with extraction results for reuse in the Save and Exit dialog. This will save you time because the extraction results are reused rather than completing a new extraction when the session is launched.



The slide has a light blue background with a faint, abstract graphic of a person's head and shoulders. In the top left corner, there is a small rectangular box containing the text "Business Analytics software". In the top right corner, the "IBM" logo is displayed. The main title "Demo 3" is centered at the top in a large, bold, black font. Below it, the subtitle "Using an Updated Modeling Node" is also centered in a smaller, regular black font. At the bottom right of the slide, there is a small, fine-print copyright notice: "© 2014 IBM Corporation".

This demo is a continuation of Demo 2. If you have not completed Demo 2, a start file has been provided for you.

This demo uses the following Modeler stream.

- C:\Train\0A105\06-Reviewing_Types_and_Concepts_in_the_Interactive_Workbench\Concepts_Model_Demo3_start.str - a Modeler stream that reads a file containing call center data for March and April

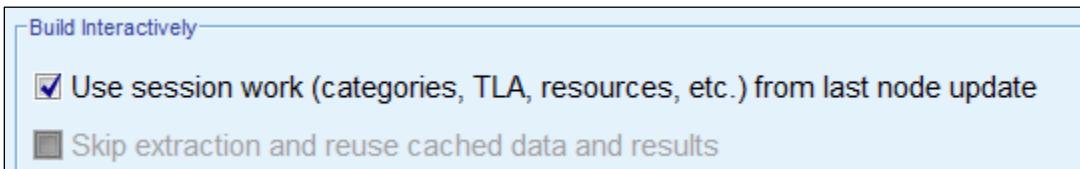
Demo 3: Using an Updated Modeling Node

Purpose:

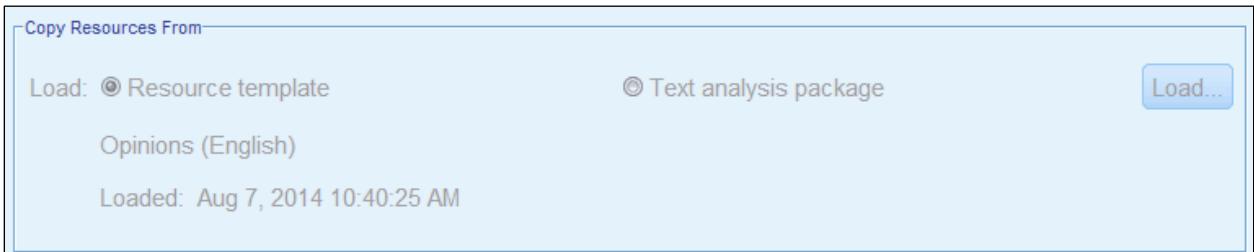
In previous the previous two demos for this module, you updated the modeling node before you exited the Interactive Workbench so you could save our work for later use. In this demo, you will get a more in depth view of what it means to update the text mining modeling node.

Task 1. Using an updated Text Mining node.

1. From the **File** menu, click **Open Stream**.
2. Navigate to **C:\Train\0A105\06-Reviewing Types and Concepts in the Interactive Workbench**, and the double-click **Concepts_Model_Demo3_start.str**.
3. Edit the **Text Mining** node, and then click the **Model** tab.



Because the modeling node was updated, the procedure will use session work (categories, TLA, resources, etc.) from the last node update.

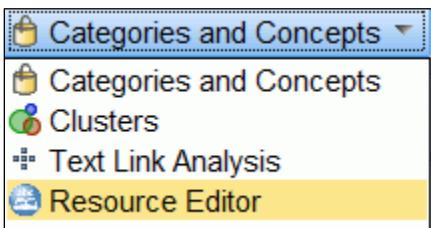


As a result, you are no longer able to select another resource template or a Text Analytics Package. The options for doing so have been grayed out.

4. Click **Run**.

After the extraction is complete, continue with the next step.

5. In the top right corner of the Interactive Workbench, click the list and select **Resource Editor** view.



The Excluded Dictionary is on the right side of the window.

6. In the **Exclude Dictionary**, sort the **Exclude List** by clicking the column header until the arrow points up.

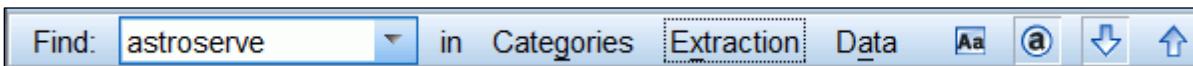
	Exclude List	Library
0	<input type="checkbox"/>	
1	<input checked="" type="checkbox"/> any kind of problem	Opinions Library (English)
2	<input checked="" type="checkbox"/> any problems i have	Opinions Library (English)
3	<input checked="" type="checkbox"/> anykinf of problem	Opinions Library (English)
4	<input checked="" type="checkbox"/> astroserve	Local Library
5	<input checked="" type="checkbox"/> can't wait	Opinions Library (English)
6	<input checked="" type="checkbox"/> copyright*	Core Library (English)
7	<input checked="" type="checkbox"/> i was out of	Opinions Library (English)
8	<input checked="" type="checkbox"/> if i ever have a problem	Opinions Library (English)
9	<input checked="" type="checkbox"/> if i ever have problems	Opinions Library (English)
10	<input checked="" type="checkbox"/> if i have a problem	Opinions Library (English)
11	<input checked="" type="checkbox"/> if i have questions	Opinions Library (English)

7. Switch to the **Categories and Concepts** view.

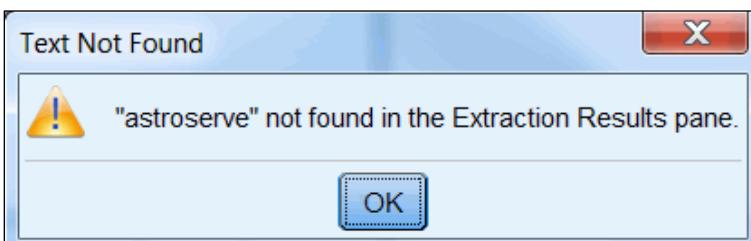
8. On the toolbar, click **Find** .

You can now initiate a search for the word "astroserve" in the extracted list to verify that it has not been extracted.

9. Beside **Find**, type **astroserve**, and then click **Extraction**.



The following message confirms that **astroserve** was not extracted:



10. From the **File** menu, click **Close**, and then click **Exit**.
11. From the **File** menu, click **Save Stream As**.
12. Name the stream **Extraction**.
13. From the **File** menu, click **Exit** to end the Modeler session.

Result:

You should now have a more in depth view of what it means to update the text mining modeling node.

Apply Your Knowledge

Purpose:**Test your knowledge the material covered in this module.**

Question 1: True or False: Filtering concepts affects the categorization of concepts since when categorization is done, it uses only concepts that are displayed.

- A. True
- B. False

Question 2: True or False: Updating a Text Mining node retains the changes that you made to the linguistic resources, text link analysis rules, or any categories that were created during the interactive session.

- A. True
- B. False

Question 3: True or False: By default, all the extracted concepts are displayed in the Extraction pane.

- A. True
- B. False

Question 4: True or False: It is important to make sure that all concepts are typed as something other than Unknown.

- A. True
- B. False

Question 5: True or False: Even though the Use session work (categories, TLA, resources, etc.) from the last node update box is checked in the Model tab of the Text Mining Modeling node, it is still possible to load a new resource template or Text Analysis Package.

- A. True
- B. False

Apply Your Knowledge - Solutions

Answer 1: B. False

Answer 2: A. True

Answer 3: B. False. By default only the top 5,000 concepts are displayed.

Answer 4: B. False. While you may want to create some new types to group together certain concepts that share something in common, it is not a concern if many of the concepts are typed as Unknown. For instance, because so few concepts were typed as Organization in the Astroserve data, you may want to fine tune the resources so Astroserve's competitors get recognized as organizations. Also it should be noted that terms that are typed as Unknown will still be used in the categorization process.

Answer 5: B. False. As long as this box is checked, the options to load another resource template or Text Analysis Package are grayed out. Of course you can override this by unchecking the box but then you no longer will be using any of your updates. This will probably not be desirable if you have made a number of changes to the dictionary resources.

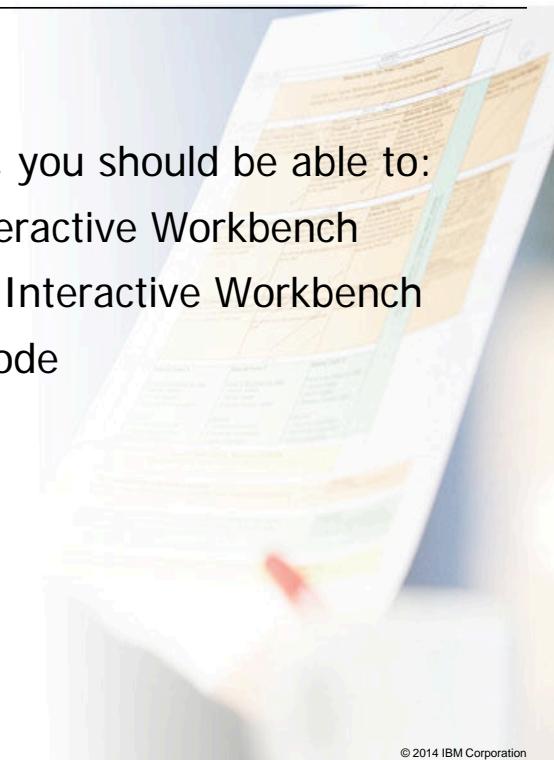
Business Analytics software

IBM

Summary

- At the end of this module, you should be able to:
 - review types in the Interactive Workbench
 - review concepts in the Interactive Workbench
 - update the Modeling node

© 2014 IBM Corporation



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

6-36

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Business Analytics software



Workshop 1

Reviewing Extracted Results in the Interactive Workbench



© 2014 IBM Corporation

The following file will be used:

- C:\Train\0A105\06-Reviewing Types and Concepts in the Interactive Workbench\Music_Survey.str - a Modeler stream that reads from a file containing customer likes and dislikes about portable music players

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Workshop 1: Reviewing Extracted Results in the Interactive Workbench

When using Text Analytics, it is important to evaluate how well the dictionary resources are performing. If not, you will need to modify them so that they do meet your expectations performing an analysis or creating a model.

- Open the **music_survey.str** stream from the C:\Train\0A105\06-Reviewing Types and Concepts in the Interactive Workbench directory.
- Launch an Interactive Workbench session.
- How many concepts were extracted? If the limits are reached, change the settings to display all concepts.
- Study the concepts closely. Look at the data records for some of the concepts.
- Look for concepts that should be synonyms and for spelling problems.
- Review the types that were extracted. Think about types that you might create for these data.
- Look for text that should be extracted but is not. Think about how you can force extraction of this text.
- Close Modeler without saving the stream file.

Workshop 1: Tasks and Results

Task 1. Reviewing results in the Interactive Workbench.

- Open the **music_survey.str** stream from C:\Train\0A105\06-Reviewing Types and Concepts in the Interactive Workbench directory.
- Edit the Text Mining modeling node, ensuring that the options Build Interactively and Using extraction results to build categories are selected.
- Launch an Interactive Workbench session.
- How many concepts were extracted? If the limits are reached, change the settings to display all concepts.

The screenshot shows the 'Extract' view in the IBM SPSS Modeler Interactive Workbench. The title bar says 'Extract'. Below it, a toolbar has icons for Extract, Import, Export, and Save. To the right are buttons for 'Map' and 'Display'. A search bar says '404 concepts'. On the far right is a 'Concept' dropdown. The main area is a table with the following data:

Concept	In	Global	Docs	Type
small		58 (5%)	58 (14%)	<Contextual>
music		54 (4%)	51 (13%)	<Features>
easy to use		45 (4%)	44 (11%)	<Positive>

A total of 404 concepts were extracted. This number is well short of the default display limit.

- Study the concepts closely. Look at the data records for some of the concepts.
One result that would be sure to get your attention: 11% of the respondents mentioned that "ease of use" is what they like most about a portable music player.
To examine their actual responses:

- Click the **easy to use** concept in the Extraction pane, and then click **Display** in the upper right corner of the pane.

For example, Respondent 8, mentioned "easy to use" twice in their response.

	Q1_What_do_you_like_most_about_this_portable_music	Categories
7	Its great look and easy to use interface Easy to use. Has a big screen. Software is easy to use, organizes folders in trees so you can open to investigate or close to save space. 20 GB hard drive.	
8		
9	Its good looking and easy to use	
10	Small and easy to use	
11	It's well designed, light and very easy to use.	
12	It is compact, tidy and easy to use	
13	Ease of use. Looks good. Compatible with car stereo.	

- Look for concepts that should be synonyms and for spelling problems.

While the dictionary resources already handle if there are misspellings, it is possible that they might miss some of them. In this case, because only 404 concepts were extracted, it would not take too much time to go through them to see if you can pick up some obvious misspellings. Also you should look for textual variants that could be considered synonymous. The best way to check for these is to sort the concepts in alphabetical order and then scroll down:

- Click the **Concept** column header in the left corner of the extraction pane to alphabetize the concepts.

- Scroll down until you see the word **amount**.

Concept	In	Global	Docs	Type
ageing ears		1 (0%)	1 (0%)	<Unknown>
album		1 (0%)	1 (0%)	<Unknown>
alternative		2 (0%)	2 (0%)	<Contextual>
amount of memory		1 (0%)	1 (0%)	<Performance>
amount of music		1 (0%)	1 (0%)	<Features>
amount of storage		1 (0%)	1 (0%)	<Characteristics>
amount of storage		1 (0%)	1 (0%)	<Characteristics>
amount of tunes		1 (0%)	1 (0%)	<Unknown>
amounts of music		1 (0%)	1 (0%)	<Features>
appearance		1 (0%)	1 (0%)	<Characteristics>
appropriate		7 (1%)	7 (2%)	<Positive>
artist		1 (0%)	1 (0%)	<Unknown>

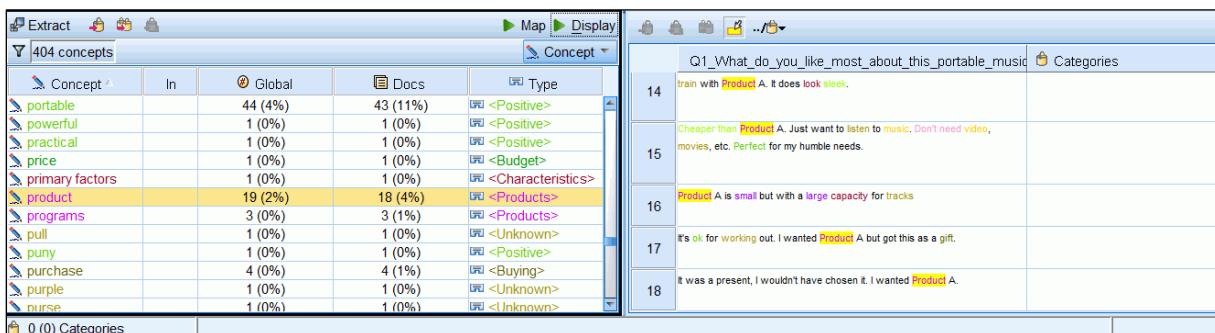
Note that there are several terms that pertain to the capacity of the portable music player: amount of memory, amount of music, etc. You want to create a synonym or a type to indicate that these concepts have something in common, which is that they all have something to do with the capacity of the player.

- Review the types that were extracted. Think about types that you might create for these data.

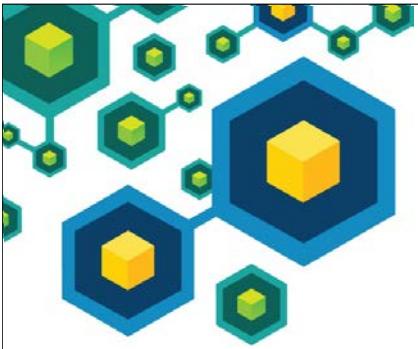
This is not an easy question. Should you group all the concepts pertaining to "capacity" in a type, or create a synonym to do the equivalent? It may not be immediately apparent how best to proceed. For example, should you create a new type to group all the terms that mention a feature, or should you create a synonym definition to do the equivalent? One key difference between these two approaches is that a synonym definition effectively modifies the terms so that in the list of concepts, only the target would appear. In other words, in synonym definitions, the goal is to replace terms with a target so that you only see a single concept. Conversely, a type definition will keep each of the concepts distinct, and so they would appear in the list of concepts separately. In the Type view in the Extraction Results pane, all the terms assigned to a type will appear with it if the type is expanded.

- Look for text that should be extracted but is not. Think about how you can force extraction of this text.

There are many different types that you could create; here is just one of them. When you examine the data, some of the respondents made comparisons between their player and another brand, Product A. In some instances, they considered their player better, and in others they did not. What did respondents like about Product A and what didn't they like? In order to perform that kind of analysis, you need to make sure that "Product A" was extracted as a concept. As it turns out, it was not. The term "Product" was extracted from the data but not "Product A". The remedy is to create a Type to force the extraction of "Product A". This will be covered in this course.



- Close Modeler without saving the steam.



Editing Linguistic Resources

IBM SPSS Modeler Text Analytics (v16)

Business Analytics software



© 2014 IBM Corporation

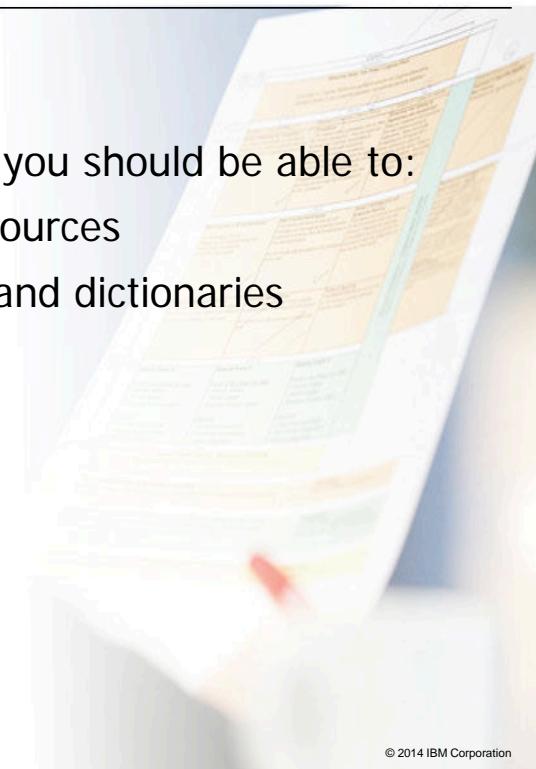
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Objectives

- At the end of this module, you should be able to:
 - modify the linguistic resources
 - access project libraries and dictionaries
 - create synonyms
 - create types
 - create exclusions



© 2014 IBM Corporation

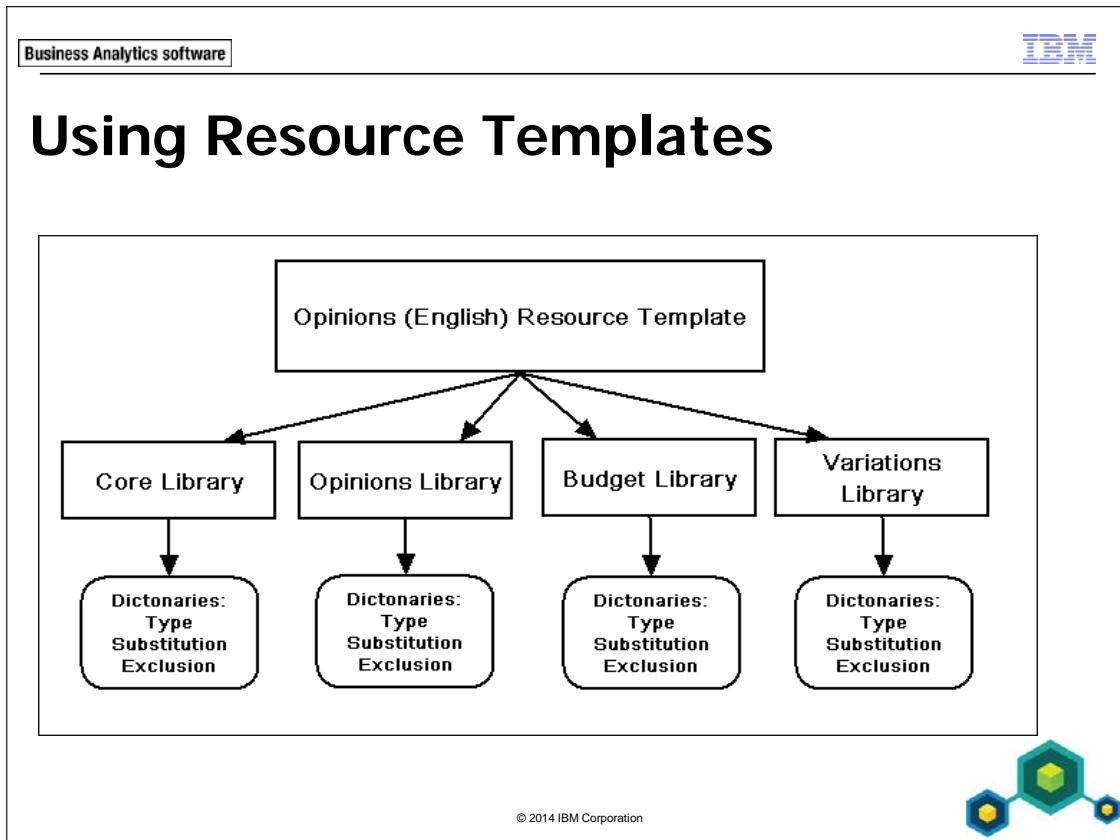
In previous modules, you learned about how text data are extracted in the Text Mining node and the options available for extraction. In Module 6, Reviewing Types and Concepts in the Interactive Workbench, you learned about reviewing the extracted concepts and types for possible modifications to the linguistic resources. Before editing any of the dictionaries, it is best to provide a complete understanding of how the linguistic resources are structured and organized. The various linguistic resources and their relationship to each other will be presented in the module.

Linguistic resources are, for the most part, edited in the Interactive Workbench. This is also the environment in which you manage the resources by saving to an existing resource template or creating a new one.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.



IBM SPSS Modeler Text Analytics is shipped with a set of specialized resource templates, each of which is comprised of a set of libraries, compiled resources, and some advanced resources. Libraries are, in turn, comprised of dictionaries, including those that control types and terms/ concepts, synonyms, optional elements, and excluded terms. Several libraries are included in each resource template. It is very important to note that some of the libraries are shared by multiple resource templates.

Thus, there is a hierarchy of structure to the linguistic resources, with resource templates at the upper level, and individual dictionaries within a library occupying the lowest level. Contained within a dictionary are the actual linguistic definitions for types, synonyms, etc.

The compiled resources cannot be edited by the user, and they contain many definitions supplementing the types in the Core library.

Many resource templates are shipped with Modeler. In English language versions, the default template is Basic Resources (English). There are Basic Resources templates in several other languages. There are also specialized templates for a variety of specific application areas, such as gene ontology, genomics, CRM, and security intelligence.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Most of the specialized templates are currently available only in English. These templates have been fine-tuned for key concepts and language in these application areas.

The date on which each template was created is listed, along with ownership information and the language used in that template. Annotations can be added to a template and are displayed here. If an icon appears in the TLA column, it means that the template contains TLA (Text Link Analysis) patterns. Text link analysis can only be done with templates that include such patterns, whether included with Modeler or added by the user.

Whenever you load a template, a copy of the template's resources at that moment is loaded and stored in the Text Mining node. Only the contents of the template are copied while the template itself is not linked to the node. This means that if this template is later updated (from another stream and node), these updates are not automatically available in the node. In short, the resources loaded into the node from the template are always used unless you either load a different template's contents or unless you make changes to the resources in a session, update the node, and select the Use session work option.

You can edit and add to the resources during an Interactive Workbench session. If you want to reuse these changes later in another node or another stream, you must save the resources as a template during the session. When you do so, you can save with a new template name or overwrite an existing template.

Using Libraries

- Each resource template contains several libraries but the most commonly used are as follows:
 - Opinions library
 - Local library
 - Budget library
 - Core library
 - Variations library

© 2014 IBM Corporation



The libraries that come standard with any new project are the Core, Opinions, Budget, and Variations. There is also a Local Library which initially is empty.

- Local Library: An initially empty custom library specific to your project. Any changes that you make to the linguistic resources directly from the Text Analysis window will be automatically stored in the first library listed in the library tree in the Resource Editor window. By default, this is the Local Library.
- Core Library: A library comprised of five built-in type dictionaries. Four of them represent People, Locations, Organizations, and Products. The fifth groups all the Unknown terms. The library also includes optional elements and one exclude term to help group terms and exclude others. Note that many terms in the compiled resources are also assigned to these types, so you may find a term assigned to a type (such as for common geographical locations), but not find it explicitly listed in the type dictionary. You can add to and modify the terms, synonyms, and exclusions in this library. However, you cannot add, delete or rename the types in this library.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

- Budget Library: A library used to extract terms referring to the cost or price of something. The library includes the Budget type plus synonyms.
- Opinions Library: A library that contains resources to extract opinions and patterns from the responses. This library includes thousands of words that represent attitudes, qualifiers, or preferences that indicate an opinion about something. The library includes fourteen built-in types, as well as synonyms and excluded terms. The patterns that you see in the text analysis window are identified using the types defined in this library. As with the Core Library, you can add to and modify the terms, synonyms, and exclusions in this library. You cannot add, delete or rename the types in this library.
- Variations Library: This library contains synonyms covering the majority of variations in English spelling, principally among American and British variations.

Note that library precedence is determined by the order of the libraries (and this can be changed).

Using Library Dictionaries

- Each library is composed of three dictionaries:
 - Type dictionaries
 - Substitution dictionary
 - Exclusion dictionary

© 2014 IBM Corporation



The resources used by the extraction engine to extract and group terms from the text data always contain one or more libraries. You can see the set of libraries in the library tree located in the upper left part of the View pane. The libraries are composed of three kinds of dictionaries:

- Type dictionary: Located to the right of the library tree, this pane displays the contents of the type dictionaries for the libraries selected in the library tree. The dictionary contains the types, with each type being defined by a collection of words to be grouped under that type name/label. When the extractor engine reads the text data, it compares words found in the text to the terms in the type dictionaries. If an extracted concept appears as a term in a type dictionary, then that type name is assigned. You can think of a type as a distinct dictionary of terms that have something in common.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Somewhat confusingly, each type, such as Positive, is also labeled a dictionary (hence the Positive type dictionary). To keep things clear, this training guide will differentiate between type dictionaries which contain all the types defined for a particular library, and the types themselves. Thus, a type dictionary, for the course usage, is comprised of one or more types (for example, Positive, Negative) and their associated terms (for instance, good, bad).

- Substitution dictionary: This dictionary contains two components:
 - Synonym dictionary: Located in the bottom portion of the Resource Editor window, this dictionary is a collection of words defined as synonyms or as optional elements used to group similar terms under one target term, called a concept in the final extracted results. In essence, synonyms are terms that have the same meaning. This dictionary can contain known synonyms, user-defined synonyms and elements, as well as common misspellings paired with the correct spelling. A library can contain only one substitution dictionary.
 - Optional elements dictionary: A collection of terms used to group variants of terms together. Optional elements are single words that, if removed from an extracted compound term, could create a match with another extracted term. These single words can appear anywhere within the compound term. Examples are inc. or corp. The optional elements mostly refer to business terms.

Since this pane manages both synonyms and optional elements, this information is organized into two tabs.

- Exclusion dictionary: Located in the right side of the Resource Editor, the exclude dictionary is a collection of terms and types that will be removed from the final extracted results. A library can contain only one exclude dictionary.

Using Type Dictionaries

- Each type dictionary consists of the following:
 - one or more types
 - a list of the terms contained in the type
 - parameters that control how the type is applied

© 2014 IBM Corporation



A type dictionary contains one or more types and their list of terms, and various parameters that control how the type is applied. Each type has three parameters that control how it is applied: Match, Inflect and Force, the column with the pushpin symbol :

- Inflect: This option instructs the extractor engine to use grammatical morphology to capture similar forms of the terms that you add to this dictionary during the extraction process, such as singular or plural forms of the term. This option is particularly useful when the type contains mostly nouns, since it is unlikely you would want inflected forms of verbs or adjectives.
- Force: This option instructs the extractor to ignore any other occurrences of this same term in other dictionaries and libraries. Double-clicking this cell turns forcing on. When forcing is used for a term that occurs with more than one type, a Resolve Conflicts dialog box appears and prompts you to select which type should be used for the term. When a term is forced, other occurrences of the term will have a black X icon in the Force column.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

- Match: This option specifies the manner in which extracted words are matched to a term associated with a type. There are 10 alternatives match options:

	Match Option	Description
1	Entire Term	The entire extracted text has to match the exact term in the dictionary. However, for the Persons type, it will also extract entire names using a first name only.
2	Start	The beginning word of the extract text must match the term in the dictionary. For example, if you enter apple , apple tart will be matched.
3	End	The end of a term extracted from the text must match the term in the dictionary. For example, if you enter room, dining room will be matched.
4	Any	Any portion of the term extracted from the text must match the term in the dictionary. For example, if you enter food, then snack food, seasonal food, wine and food, will all be matched. Any is used to match multi-terms, not single word terms.
5	Start or End	Either the beginning or end of the extracted text must match the term in the dictionary.
6	Entire and Start	Either the entire extracted text, or the beginning word, must match the term in the dictionary.
7	Entire and End	Either the entire extracted text, or the ending word, must match the term in the dictionary.
8	Entire and Any	This option is used to match single-word terms to a type, no matter where they occur in the extracted text.
9	Entire and (Start or End)	Either the entire extracted text, the starting word or the entire word, must match the term in the dictionary.
10	Entire (no compounds)	If the entire concept extracted from the text matches the extracted term in the dictionary, this type is assigned and the extraction is stopped to prohibit the extraction from matching the term to a longer compound.

To further explain how text is matched for multiple word terms, assume there is a survey with four respondents. The first column shows text extracted from each response for a question about items used for cleaning in the house.

The word "soap" is the term that was entered in the dictionary assigned to the type called "Cleaning". Thus, when an extracted term is set to "soap", it will be typed as "Cleaning". The question then becomes, depending on the matching option, which of the extracted terms will be set to the "Cleaning" type? The answer is displayed in the columns in the table.

Extracted terms	Entire term	Start	End	Any
soap powder		Cleaning		Cleaning
laundry soap			Cleaning	Cleaning
dish soap towel				Cleaning
soap	Cleaning	Cleaning	Cleaning	

The Entire term option sets only the last response to "Cleaning". The first three do not match because the entire term contains more words than just "soap". The Start option assigns "soap powder" to Cleaning because "soap" starts the term, but not "laundry soap" because "soap" ends the term. And for the End option, the reverse is true. The Any option is straightforward but it only works with multi-terms, but not single word terms.

Here is another example that illustrates another typical situation that occurs with multiple word terms. Assume that the defined term is "priority member" and that the text extracted from a response is "priority member service". Of course, the matching option of Entire term will not work. But neither will the Start option, even though the words "priority member" begin the extracted term. The reason is simple: the Start of the term extracted from the text is "priority" not "priority member". In other words, it is only the first (or last) word that is critical. In this case, you can create a synonym definition for "priority member service", or you can add that multiterm to the type definition.

Using Substitutions Dictionaries

- Collection of term substitutions that help to group similar terms under one target term
- You can define two forms of substitutions in this dictionary:
 - synonyms
 - optional elements

© 2014 IBM Corporation



A Substitution dictionary is a collection of term substitutions that help to group similar terms under one target term. You can define two forms of substitutions in this dictionary:

- Synonyms: These associate two or more words that have the same meaning. For example, synonyms for no interest include don't have any interest, nothing of interest to me, and not much interested. Synonyms are also used to group misspelled words with their correct form (the fuzzy grouping algorithm cannot find all these problems).
- Optional elements: These identify optional words in a compound term that can be ignored during extraction in order to keep like terms together even if they appear slightly different in the text. Examples are "inc." and "ltd" added to a company name. You can use the tabs at the bottom of the Substitution pane to switch between views of one or the other form of substitution. By identifying optional elements and synonyms, you can force the extractor to map these terms to one single target term. This reduces the number of terms in the final list and thus creates a more compact set of terms with higher frequencies.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

7-13

In synonym substitution, the term you want to replace is called the synonym, and the term you want to use in the extraction is called the target. For example, if you want "cost too much" to be replaced by "costly", then "cost too much" is the synonym and "costly" is the target. If a term is a synonym for more than one target, it will be matched to the first one encountered in the dictionary. It is important to know that a term must first be extracted to be a synonym. However, the target term does not need to be extracted. Thus, if "cost too much" is extracted, you can substitute "costly" for it, even if the latter term is never extracted, or even included in the response.

By default, synonym substitution is also applied to the inflected forms (such as the plural form) of the synonym. Depending on the context, you may want to impose constraints on how terms are substituted.

Synonyms can also be defined from the Categories and Concepts view by right-clicking on a term of interest and using the Context menu to add a synonym from that location.

Certain characters can be used to place limits on how far the synonym processing should go. When using these characters, the exclamation mark must immediately precede the term. A space should be left between the carat and dollar sign and the term.

	Character(s)	Description
1	Exclamation mark (!):	<p>When the exclamation mark directly precedes the synonym this indicates that no inflected forms of the synonym will be substituted by the target term. For example, in the Synonym pane, if a target "affix" has the synonym "!affixes", this implies that no variants of this synonym will be used to make the substitution.</p> <p>An exclamation mark directly preceding the target term (!target-term) means that you do not want any part of the compound target term or variants to receive any further substitutions. For example, the Opinions Library includes the target term "!doesn't meet expectation", so if a synonym is encountered and "doesn't meet expectation" is substituted, no further substitutions should be made (for example, for "expectation").</p>

	Character(s)	Description
2	Carat (^):	A carat preceding the synonym means that the synonym applies only when the extracted term begins with the synonym. Thus if "^ wage" is the synonym and "income" is the target, and "wage earner" is extracted, then "income" will be substituted. If, instead, "minimum wage" is extracted, the substitution will not be made.
3	Dollar sign (\$):	A dollar sign following the synonym means that the synonym applies only when the extracted term ends with the synonym. Thus if "soap \$" is the synonym and cleaning materials is the target, and "dish soap" is extracted, "cleaning materials" will be substituted. If, instead, "soap bubbles" is extracted, the substitution will not be made.
4	Carat(^) & Dollar sign (\$):	If the carat and dollar sign are used together, a term must be an exact match to the synonym for substitution to occur. Thus if "^ price \$" is defined as a synonym for the target "cost", the extracted term must be simply "price" for the substitution to occur. If "price fixing" is extracted, no substitution will occur.
5	Asterisk (*):	An asterisk placed directly after a synonym, such as synonym*, means that you want this word to be replaced by the target term. For example, if you defined manage* as the synonym and management as the target, then associate managers will be replaced by the target term associate management. You can also add a space and an asterisk after the word (synonym *) such as internet *. If you defined the target as internet and the synonyms as internet ** and Web *, then internet access card and Web portal would be replaced with internet. You cannot begin a word or string with the asterisk wildcard in this dictionary.

Using Exclusion Dictionaries

- Terms you do not want extracted
- Generally, fill-in words or phrases that add little to a response and tend to clutter the extraction results
- For example, "usual", "feel that", and "for instance"

© 2014 IBM Corporation



An Exclude dictionary is a list of terms that you want to ignore during extraction. The terms are characteristically fill-in words or phrases used for continuity that do not really add anything to a response and tend to clutter the extraction results. Examples are as "usual", "feel that" and "for instance". You can also exclude any terms that you do not want used, even if they have a substantive meaning. Thus, if you had asked about people's pets and wanted to focus only on dogs, you could add typical pets, cats, ferrets, fish, monkeys, to the exclude dictionary, and those pet terms would never be extracted. If an entry is also declared somewhere else in the interface, such as in a type dictionary, it is shown with a strike-through in the other dictionaries, indicating that it is currently excluded. This string does not have to appear in the text data or be declared as part of any type dictionary to be applied.

You can add character strings to the exclude dictionary as one or more words or even partial words using the asterisk as a wildcard. Partial strings can be entered with the use of an asterisk as a type of wildcard, so any text that includes the string and following text will be excluded.

If an excluded term is also a target in the substitution dictionary, then the target and its synonyms will be excluded. This is because substitutions occur before exclusions.

Modifying the Dictionaries

- There are various actions you can take when modifying the dictionaries for a project:
 - define synonyms for words/terms that have the same meaning,
 - define to fix misspellings and word variants
 - add terms to existing types to group terms
 - create new types to group terms
 - force extraction of terms
 - exclude certain words/terms

© 2014 IBM Corporation



Now that you know more about the dictionary resources, you will examine the methods to modify them:

- Define synonyms for words/terms that have the same meaning: Surveys deal with specific topics which can have their own vocabulary. While the default resources define a lot of common synonyms, there are others that arise from openended text in a survey that need to be modified, corrected or added.
- Define synonyms to fix misspellings and word variants: The Opinions library uses its Synonym dictionary to correct common misspellings, and the Customer Satisfaction library includes several misspellings in the type definitions. Although you might think about fixing spelling mistakes before importing the data into the program, these errors can be caught with additions to the synonym dictionaries (and other errors are caught with the fuzzy grouping algorithm).
- Add terms to existing types to group terms: If there are organizations, locations, products, or opinion-related terms which are not recognized by the program, you may want to add them to the appropriate existing type.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

7-17

- Create new types to group terms: You often want to create new types that are specific to the topic, and then associate several terms with those types. If these terms are already extracted, they are probably typed as Unknown and thus may or may not be grouped together when the responses are categorized. Putting the terms under one type makes certain that they will be grouped.
- Force extraction of terms: If you discover terms that have not been extracted that you feel are important, you can force their extraction principally by adding them to an existing type or by creating a new type.
- Exclude certain words/terms: The default Exclude dictionary contains many common terms that are by default excluded from extraction, but you may find others in the responses that should be added to that dictionary.

In many instances, you may need to edit the existing definitions, rather than add new ones. That is, you may find that you need to modify or delete an existing synonym definition, one or more excluded terms, or disable a term, type, synonym, or excluded term (with the check box). It is good practice to see whether a term already exists in a dictionary before editing.

Extracting Unextracted Text

- The software will not always extract a word or phrase that you would like to use in a category
- Often these words are acronyms, abbreviations, or slang

© 2014 IBM Corporation



The software will not always extract a word or phrase that you would like to use in categorization. If a term is not extracted, it will not be used in the dictionaries (with the exception of the target for a synonym). And, of course, the term cannot be used in categorization. It is highly recommended scanning through the responses in the Data pane in the Text Analysis view to see which words were not extracted (they will be black in color). If you see one or more words that you deem important, but that were not extracted, then you need to take some action.

To force extraction, you can try adding a term to an existing type or a new type. When extraction is rerun, check to see whether it has been extracted. If this does not work, it is probably because the term is already defined, perhaps as part of the compiled resources. In that case, make sure that you are adding the term to a new type in the Local library. This should take precedence over other occurrences (you may be asked to select which type should be used to force the term into that type).

Preparing for Linguistic Editing

- What concepts were not extracted?
- What concepts were extracted incorrectly?
- What concepts should have been extracted together?
- What concepts you would like to ultimately make into categories?
- What concepts should not have been extracted?
- Which concepts are important?
- Which concepts require synonyms?

© 2014 IBM Corporation



After the first extraction, it can be difficult to decide where to begin modifying the dictionaries and to know exactly what to modify. Often, a good strategy is to categorize the extracted terms and types before making any changes. Since categorization is the endpoint of text mining, if you discover that some of the categories created with the default extraction results are useful and complete, the user may not need to modify the dictionaries as extensively as first suspected. That is, you could discover that the categorization methods have grouped terms and solved other potential analysis problems that you recognized in the extraction results.

When you do begin to edit the linguistic resources, you should consider in which library you will make changes. This depends in part on whether the data are similar to data you expect to encounter in the future, and also how the user plans to save the linguistic resources. Generally, you should make changes in the Local library, which is specific to a particular project. This will allow you to publish this library under a new name so it will be available for other projects.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Preparing for Linguistic Editing (cont'd)

- Which misspellings are not corrected by the software?
- Which concepts are incorrectly grouped together because of the fuzzy grouping algorithm?
- Which concepts should be grouped under a type?
- Which terms should be excluded because they appear too often but are uninformative?

© 2014 IBM Corporation



You may have to modify the Core, Opinions or other libraries to change an existing synonym definition or add concepts to an existing type. There is no reason to avoid making such changes, because the version of these libraries you are editing is local to this project and no longer connected to the template from where they came. However, if you save the resources as a template under an existing template name, the changes you made to these libraries will overwrite the libraries in that template. As a consequence, those changes will be used for any project that loads the template after it has been saved. This may or may not be appropriate depending on specific circumstances.

An alternative is to save the resources as a template with a new name. All libraries can be published, or shared, with other projects (or exported for other users on different computers). Publishing allows you to work with a library not part of a particular template. However, if you do not publish the libraries, they remain local (to this project). There are several changes that will be made in this module. These changes will allow you to see how to make changes to the dictionaries and how create new type and synonym definitions.

Sampling Text Data

- Text data files can be quite large.
- In most cases a sample will be sufficient to find interesting comments.

© 2014 IBM Corporation



Text data files, as with other files for data mining, can be quite large. Although eventually much of, if not all, the data will be used when scoring records with a model, in most situations only a sample of the data is required to develop a successful predictive model, and this same principle applies to text mining. A valuable text-mining model can be developed with a smaller random sample of records from a larger file.

There can be some worry about sampling because of concern that using a sample will lead to missing some interesting, but less frequent, comments. This can happen, but in most applications, text mining is used not like a laser beam but instead like a broad searchlight that looks for interesting, reasonably common comments and patterns in the text. If only a small fraction of records (less than 1%) have text about a specific topic, this is often not useful in CRM, customer survey, call center, or churn applications. Data that are confined to only a small group of records are usually not widespread enough to provide insight or increase model accuracy.

In some areas, such as health care, fraud, or security intelligence, a small number of records with unusual text can, indeed, be very important. But even in these areas, while the samples may be larger, rarely would all the data be used for editing the linguistic resources when the original files include more than 25,000 or so records.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

7-22

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Setting Text Mining Goals

- It is extremely important to set goals prior to text mining the data.
- What are hoping to learn from the data?
- Is your goal to predict an outcome such as churn, or is it to summarize customer attitudes about your organization?
- Having specific goals will provide you with a general strategy for editing the dictionaries.

© 2014 IBM Corporation



It is extremely important to set goals for your text mining project. In the case of the Astroserve data you have been using in this course, in order to be successful in text mining their data, you would need to know what Astroserve is hoping to learn from the call center data. In general, as with many text-mining projects, you can presume that the company had two key goals:

- Summarize the text into higher-level categories that will reveal the range of topics that cause customers to call. As a secondary goal, they would like to understand relationships between terms and between categories for additional insight. All of this is in support of improving their service to customers by increasing their understanding of what problems customers encounter. They are less interested, for this analysis, in studying how well the call center phone reps handle the calls each receives. That is a worthwhile study, but would require additional supplementary information about call resolution that is not in these data.

- Create a model with the categories from text-mining to increase the accuracy of predicting which customers will churn, or change providers. It is one thing to summarize customer calls, but it is another to learn which problems/issues are especially critical at causing churn. Those are the areas that need more immediate focus, and Astroserve management hopes to use text mining to identify those areas.

Knowing these goals is helpful, and they do supply a general strategy for editing. You do not need to focus on all the infrequent comments made by customers, and you should certainly begin with the more numerous concepts. Because the long-term goal is to develop a model with the text-mining results, you may also want to create some broad categories (such as, problems with billing, access to services).

Beyond this, though, the vast bulk of editing must come from a close review of the extracted results to learn what terms are superfluous ("cust", "customer"), what terms are not grouped that should be ("internet", "computer"), and what higher level types users want to create ("competitors").

There is no special order to editing the dictionaries. If you plan to add terms to a type, but also use some of those terms as synonyms, you can create the type first and add those terms to it, then specify the synonym definitions. Or, you can do the reverse. Either should work, especially when these terms are new to the linguistic resources (at least those you can see and edit). The only caveat to this flexibility relates to how often you re-extract the data to see the results of the editing. Every time you make a change to the linguistic resources, the Extracted Results pane will turn yellow indicating that another extraction is needed to see the effect of the change. But, if you re-extract after every change, it would take much too long to do the editing. However, if you make many changes to the resources, and then extract, it may be more difficult to go back and figure out exactly what change led to what set of extracted results.

It is very important when you first use Text Analytics to understand how a certain change leads to new extractions. So the user should re-extract often, but not too often. Second, if you are going to create a type, add terms to that type, and then add some of those terms as synonyms to a target, or even make one of those terms a synonym target, you obviously could see incomplete results if you only finish half this editing and then re-extract.

Third, you can edit the resources from either the Resource Editor view or the Text Analysis view, and both options are shown in this lesson. Use whichever approach you find most effective. Be aware that if you work from the Text Analysis view, you will not be able to see the change until you switch to the Resource Editor view.

Types versus Synonyms

- Sometimes it is difficult to decide on whether to group terms into a type or a synonym
- Both group terms with share something in common
- How will it affect categorization?
- How will it affect extraction?

© 2014 IBM Corporation



When making changes, it may not be immediately apparent how best to proceed. For example, you should create a new type to group all the terms that mention a feature, or should you create a synonym definition to do the equivalent? One key difference between these two approaches is that a synonym definition effectively modifies the terms so that in the list of concepts, only the target would appear. In other words, in synonym definitions, the goal is to replace terms with a target so that you only see a single concept. Conversely, a type definition will keep each of the concepts distinct, and so they would appear in the list of concepts separately. In the Type view in the Extraction Results pane, all the terms assigned to a type will appear with it if the type is expanded.

Another difference is related to categorization. If you create synonyms, only one target is used, therefore all terms will always be grouped together. However, if the terms are retained as separate concepts by creating a type, then categorization may or may not group them together, whether as concept patterns or as concepts; this allows for little more flexibility if creating hierarchical categories. Also, the total number of responses will be lower for a term than if they were grouped in one concept. However, you can place all terms in one category by simply using the type itself as a category, or using the type in a rule.

A third way in which types and synonyms differ is with regard to extraction. Identifying a concept as a term within a type will force the term to be extracted. Whereas identifying a term as part of a synonym does not force extraction. That is, synonyms only work once a term has been extracted; they are only used post extraction.

These are just some of the distinctions that you need to keep in mind when making changes. No matter what you do, though, a bit of manual coding during categorization can always create appropriate categories from the concepts or types.

Business Analytics software

IBM

Demo 1

Modifying the Dictionaries



© 2014 IBM Corporation

This demo uses the following datasets coming from a (fictitious) telecommunications firm.

- C:\Train\0A105\07-Editing Linguistic Resources\Editing Linguistic Resources_Demo1_start.str - a Modeler stream that reads a file containing call center data for March and April

Demo 1: Modifying the Dictionaries

Purpose:

Before you can analyze the Astroserve call center data, you need to modify the linguistic resources to better capture participant responses. This will involve adding new types, synonyms, and exclusions to the dictionaries.

Task 1. Searching extractions for competitor names.

Astroserve's major competitors are:

Aaphone

Horizon

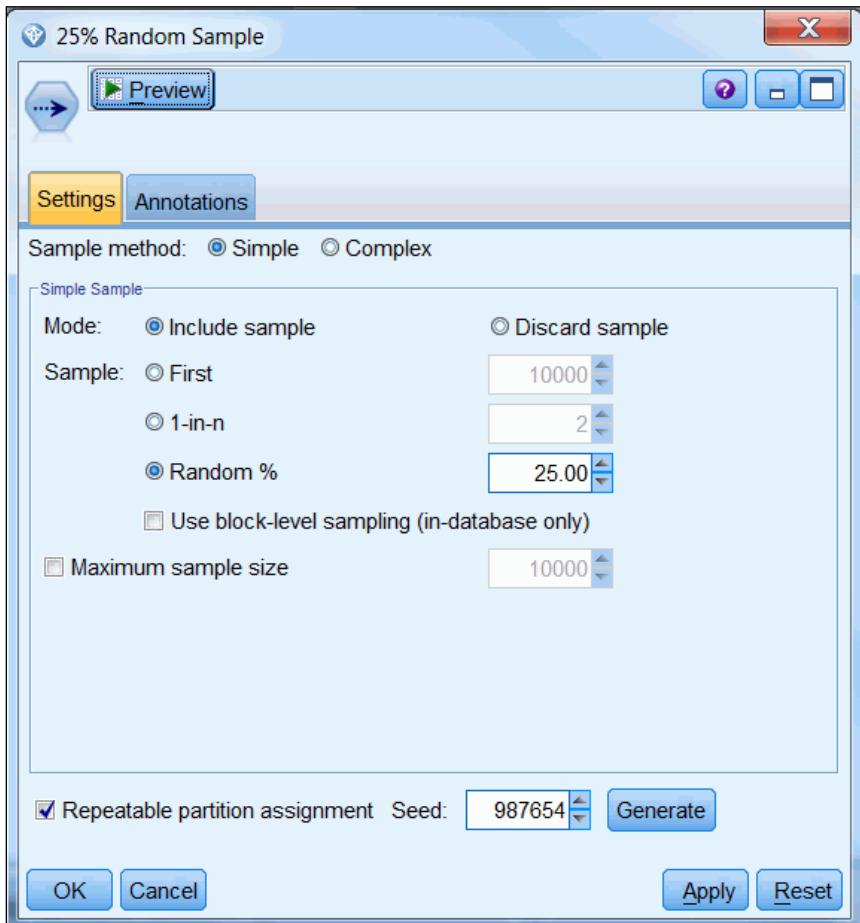
Optitel

Vodatel

In this task, you will search the extracted terms to examine whether these terms were extracted and if so, were they extracted as uniterms or parts of compound terms.

1. From the **File** menu, click **Open Stream**.
2. Navigate to **C:\Train\0A105\07-Editing Linguistic Resources**, and then double-click **Editing Linguistic Resources_Demo1_start.str**.

3. Edit the **25% Random Sample** node.



In this stream a 25% random sample is used. This will return about 1,000 records (calls) with which to work, which are sufficient to find interesting comments. A random seed was set so that the same random sample is obtained each time the stream is run.

4. Click **OK**.

5. Right-click the **Sample** node, point to **Cache** and if necessary, click **Enable**.

6. Right-click the **month** node, and then click **Run**.

This will fill the cache so it will not be necessary to draw a 25% random sample each time you run the stream.

7. Click **OK**, and then run the **Text Mining** node.

Now you are ready to search for competitor names in the list of extracted terms. You will begin with Vodafone.

8. Click **Find Toolbar**

9. If necessary, disable **Match any string containing the search string** .
10. Type **vodate** in the **Find** box.



11. Click **Extraction**.

Concept	In	# Global	Docs	Type
locked		8 (0%)	6 (0%)	<NegativeFunctioning>
worse		7 (0%)	6 (0%)	<Negative>
call astrocomm		6 (0%)	6 (0%)	<Unknown>
alpha		6 (0%)	6 (0%)	<Unknown>
register		6 (0%)	6 (0%)	<Unknown>
password		7 (0%)	6 (0%)	<Unknown>
enquiry		6 (0%)	6 (0%)	<Unknown>
3 weeks		6 (0%)	6 (0%)	<Period>
discount		8 (0%)	6 (0%)	<Unknown>
book		6 (0%)	6 (0%)	<Unknown>
reconnection fee		6 (0%)	6 (0%)	<Budget>
vodate		6 (0%)	6 (0%)	<Unknown>

The word "vodate" was extracted.

12. Repeat steps 10 and 11 for **aaphone**, **horizon**, and **optitel**.

Sometimes the results can be a little deceiving if the word you are searching for was not only extracted as a uniterm but as a compound term as well. For example, "optitel" occurred as a uniterm 20 times and in 19 documents.

Concept	In	# Global	Docs	Type
optitel		20 (0%)	19 (1%)	<Unknown>
listing		25 (0%)	19 (1%)	<Unknown>

However, it may also have been included in compound terms as well. To see this, you should search for any extraction that contains the string "optitel".

13. Enable **Match any string containing the search string**, ensure that **optitel** is in the **Find** box, and then click the **Extraction** button.

Concept	In	# Global	Docs	Type
email business		2 (0%)	2 (0%)	<Unknown>
optitel service provider		2 (0%)	2 (0%)	<Unknown>

Based on the results, the word "optitel" is included in several compound terms including this one: "optitel service provider".

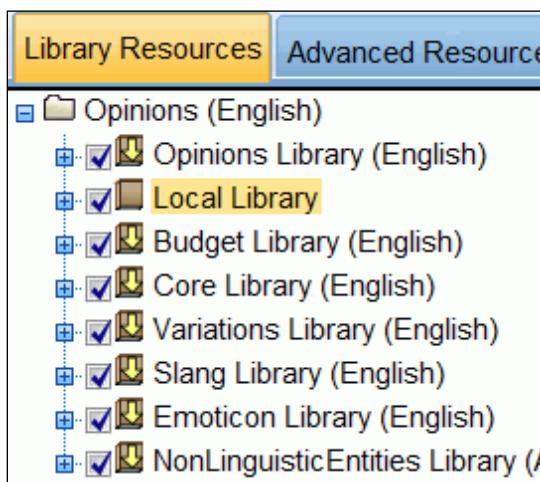
14. Repeat step 13 for **aaphone**, **horizon**, and **vodate**.

It turns out that all four competitor names were extracted as uniterms and also as part of compound terms. You will need to take this into consideration when you create a type in the next task.

Task 2. Creating a type to force extraction of competitor names.

In this task, you will create a type called Competitors.

1. Switch to the **Resource Editor** view, and ensure that **Local Library** is selected.

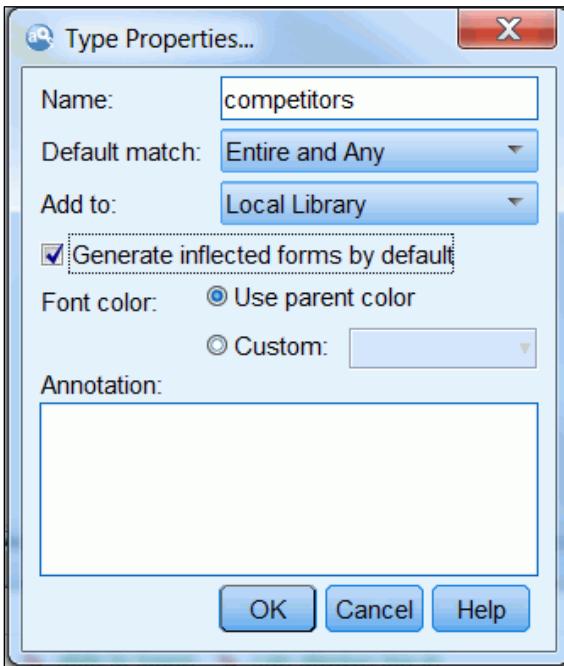


2. From the **Tools** menu, click **New Type**.
3. Beside **Name**, type **competitors**.
4. Select **Generate Inflected forms by default**.

5. Beside **Default match**, select **Entire and Any**.

This will ensure that both single-word and compound-word terms will match to the type that contain the competitor's name.

The results appear as follows:

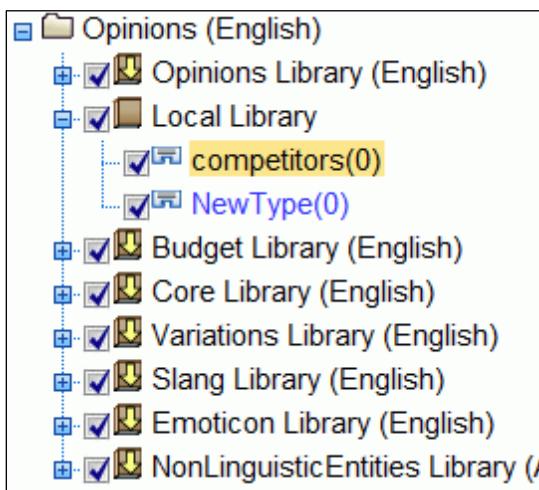


The Font color options enable you to distinguish the terms in this type from others by their text color. If you select the Use parent color option, the default type color is used for this type dictionary, as set in the Options dialog box. No colors will be changed, but you can do that as well.

6. Click **OK**.

The new type is added to the Local library. The number in parentheses is the number of terms currently assigned to the type.

The results appear as follows:

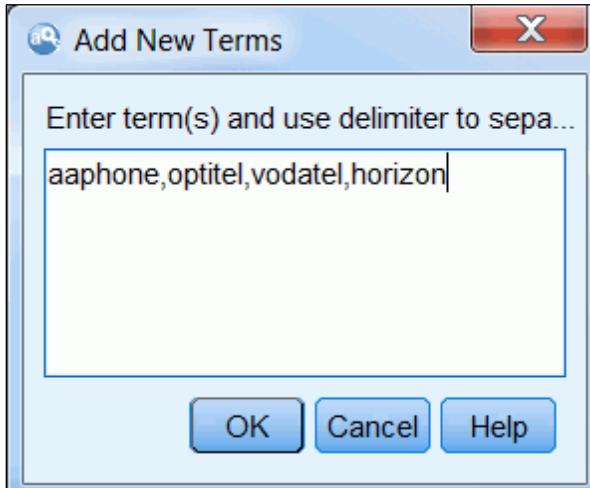


Now you will add terms to this new type.

7. From the **Tools** menu, click **New Terms**.
8. In the **Add New Terms** box, type **aaphone,optitel,vodatel,horizon**.

The terms should be lower case and separated by commas.

The results appear as follows:



9. Click **OK**.

Term	Match	Inflect	Type	Library
aaphone	Entire and Any	<input type="checkbox"/>		
optitel	Entire and Any	<input checked="" type="checkbox"/>	competitors	Local Library
vodatel	Entire and Any	<input checked="" type="checkbox"/>	competitors	Local Library
horizon	Entire and Any	<input checked="" type="checkbox"/>	competitors	Local Library

All four terms were assigned to the "competitors" type.

10. Switch to the **Categories and Concepts** view.

11. Click **Extract**.

To see how well the typing worked, use the filter facility to display all the concepts with the text "optitel".

12. From the **Tools** menu, click **Filter**.

13. Beside **Match text** type **optitel**, and then click **Filter**.

The results appear as follows:

Concept	In	Global	Docs	Type
optitel		20 (0%)	19 (1%)	<competitors>
optitel service		2 (0%)	2 (0%)	<competitors>
optitel service		2 (0%)	2 (0%)	<competitors>
churned optitel		1 (0%)	1 (0%)	<competitors>
optitelvision		1 (0%)	1 (0%)	<Unknown>
yellow pages listing		1 (0%)	1 (0%)	<competitors>
handset to join		1 (0%)	1 (0%)	<competitors>
moving to optitel		1 (0%)	1 (0%)	<competitors>
optitel.com		1 (0%)	1 (0%)	<Organization>
cust carrier optitel		1 (0%)	1 (0%)	<competitors>
astrocomm to		1 (0%)	1 (0%)	<competitors>
real estate agent to		1 (0%)	1 (0%)	<competitors>
phones to optitel		1 (0%)	1 (0%)	<competitors>
business to optitel		1 (0%)	1 (0%)	<competitors>

Based on the results, 14 concepts were extracted that contained the string "optitel". However, in two cases ("optitelvision" and "optitel.com"), the terms were not typed as competitors. You can manually add these terms to the type from this window.

14. Right-click the term **optitelvision**, point to **Add to Type** and then click **competitors**.

If you do not see the competitors on the list, click on **More** at the bottom to get a complete list of available types.

15. Repeat step 14 for the term **optitel.com**.
16. Click **Extract**.

Concept	In	Global	Docs	Type
optitel		20 (0%)	19 (1%)	<competitors>
optitel service provider		2 (0%)	2 (0%)	<competitors>
optitel service		2 (0%)	2 (0%)	<competitors>
churned optitel		1 (0%)	1 (0%)	<competitors>
optitelvision		1 (0%)	1 (0%)	<competitors>
yellow pages listing optitel		1 (0%)	1 (0%)	<competitors>
handset to join optitel		1 (0%)	1 (0%)	<competitors>
moving to optitel		1 (0%)	1 (0%)	<competitors>
cust carrier optitel		1 (0%)	1 (0%)	<competitors>
astrocomm to compensation		1 (0%)	1 (0%)	<competitors>
real estate agent to approach		1 (0%)	1 (0%)	<competitors>
Ibob@ optitel.com		1 (0%)	1 (0%)	<competitors>
phones to optitel		1 (0%)	1 (0%)	<competitors>
business to optitel service		1 (0%)	1 (0%)	<competitors>

All the terms are now correctly typed

17. Repeat steps 12 through 16 for the other three competitors: **aaphone**, **horizon**, **vodatei**.

You will undo the filtering.

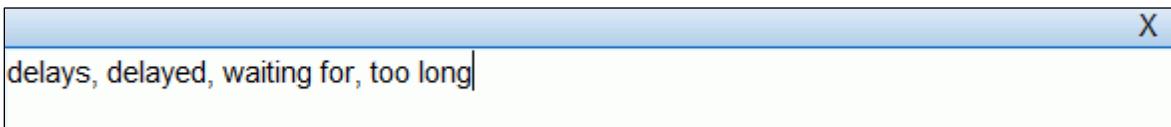
18. From the **Tools** menu, click **Filter**.
19. Delete the text in the **Match text** box, and then click **Filter**.

Task 3. Adding synonym definitions for "delay".

Synonyms can also be defined from the Resource Editor or from the Categories and Concepts view. You will begin by adding the synonym definitions for "delay". This word is common, so it would be best to first search for it in the linguistic resources to see if it has already been defined as a target.

1. Switch to the **Resource Editor** view.
2. Click the **Opinions Library (English)** resource template in the Library resources pane to select it.
3. Click **Find Toolbar** , and then in the search box, type **delay**.

4. Enable **Match any string containing the search string** .
 5. Click **Find** several times to cycle through all the instances of the string "delay" in the dictionary resources.
- Based on the results, delay has not already been defined as a target in the substitution dictionary, so there does not seem to be any complications to adding a synonym definition for delay to the dictionary.
6. Select **Local Library** in the Library pane.
 7. Scroll to the top of the **Synonyms** pane (if necessary).
 8. Double-click the blank **Target** cell, type **delay**, and then press **Tab**.
 9. Double-click the highlighted **Synonyms** cell.
 10. Type **delays, delayed, waiting for, too long** exactly as shown, with the commas and spaces.



11. Press **Enter**.

	Target	Synonyms	Library
1	<input checked="" type="checkbox"/> delay	delays, delayed, waiting for, too long	Local Library
2	<input checked="" type="checkbox"/> able to log-on	able to log-in, able to login, able to logon, can always log-in, can always log-on, can always login, can always logon, easy to log-in, easy to log-on, easy to login, easy to logon	Opinions Library (English)

12. Switch to the **Categories and Concepts** view.

13. Click **Extract**.

The synonym definition for delay is now complete.

Task 4. Adding terms to the exclusions list.

To exclude three terms plus add a wildcard again, you can add these directly in the Exclude pane or by right-clicking on a term elsewhere.

1. Switch to the **Resource Editor** view.
2. Double-click the blank **Exclude List** cell in the **Exclude** pane.
3. Type **cust**, and then press **Enter**.
4. Repeat steps 2 and 3 for **customer** and **call**.
5. Click the **Library** column header to sort alphabetically by Library.

	Exclude List	Library
0		
1	<input checked="" type="checkbox"/> copyright*	Core Library (English)
2	<input checked="" type="checkbox"/> cust	Local Library
3	<input checked="" type="checkbox"/> customer	Local Library
4	<input checked="" type="checkbox"/> call	Local Library
5	<input checked="" type="checkbox"/> any kind of problem	Opinions Library (Engl)
6	<input checked="" type="checkbox"/> any problems i have	Opinions Library (Engl)
7	<input checked="" type="checkbox"/> anykinf of problem	Opinions Library (Engl)
8	<input checked="" type="checkbox"/> can't wait	Opinions Library (Engl)

All three of these terms have been added to the Local Library. You could have excluded Astroserve and other similar terms with a wildcard character (astro*) but you chose not to because you will need them to be extracted later on in the course.

6. Switch to the **Categories and Concepts** view, and then click **Extract**.

Task 5. Fine tuning the dictionary.

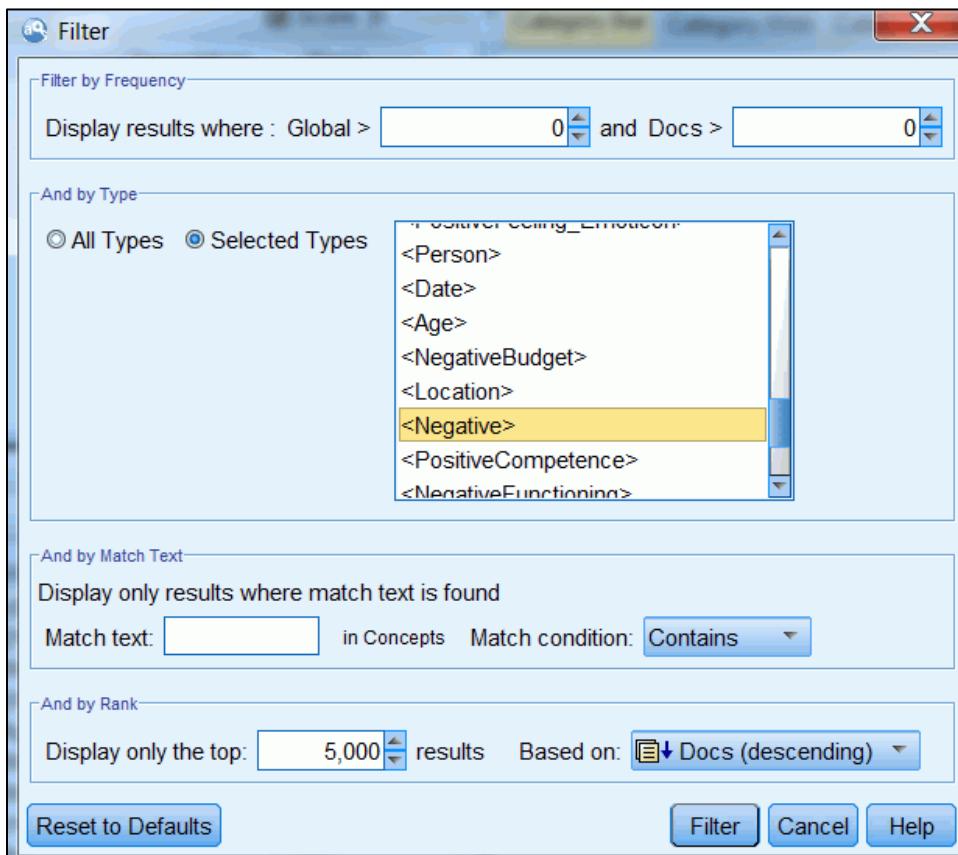
Earlier you observed that when a respondent said they were "not happy", "not too happy", "not very happy", etc. with Astroserve, only the word "happy" was extracted and it was typed as Positive. As a result, many of the responses were incorrectly typed as Positive when they should have been Negative. In this task, you will modify the dictionary so that this string is typed as Negative.

To set the stage, you should find out the number of documents that were typed as Positive and Negative.

1. From the **Tools** menu, click **Filter**.
2. In the **And by Type** section, click **Selected Types**.

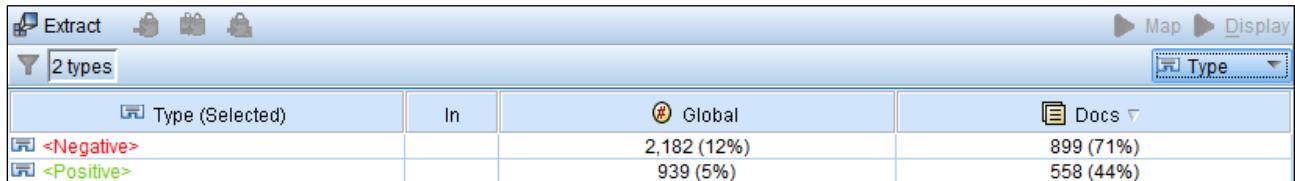
3. Ctrl+click <Positive> and <Negative>.

The results appear as follows:



4. Click **Filter**, and then click **Extract**.

5. In the **Concept** list, select **Type**.



In this window, you can observe that 44% of the documents had comments that were typed as Positive, and 71% had a Negative comment.

6. From the **Tools** menu, click **Filter**.

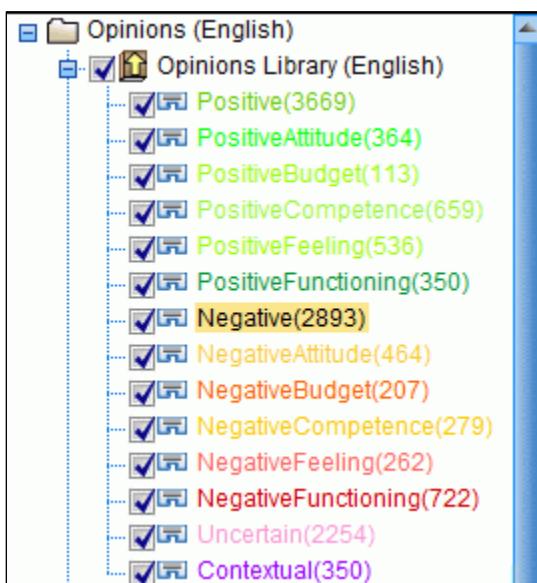
7. In the **And by Type** section, select **All Types**, and then click **Filter**.

Now you will add the term "not happy" to the Opinions Library (English) and type it as Negative.

8. Switch to the **Resource Editor** view.

9. In the **Opinions Library (English)**, select **Negative**.

The results appear as follows:



10. Click the empty **Term** cell at the top of the Type Dictionary.

11. Type **not happy**, and then press **Enter**.

Term =	Match	Inflect	Type	Library
not happy	Entire (no compounds)	<input type="checkbox"/>	Negative	Opinions Library (Engli
a bit less than expected	Entire (no compounds)	<input type="checkbox"/>	Negative	Opinions Library (Engli

12. Scroll to the top of the **Synonyms** pane, and then double-click the blank Target cell.

13. Type **not happy**, and then press **Tab**.

14. Click the highlighted **Synonyms** cell.

15. Enter the text **not happy, not very happy, not at all happy, not so happy** exactly as shown, with the commas and spaces.

The results appear as follows:

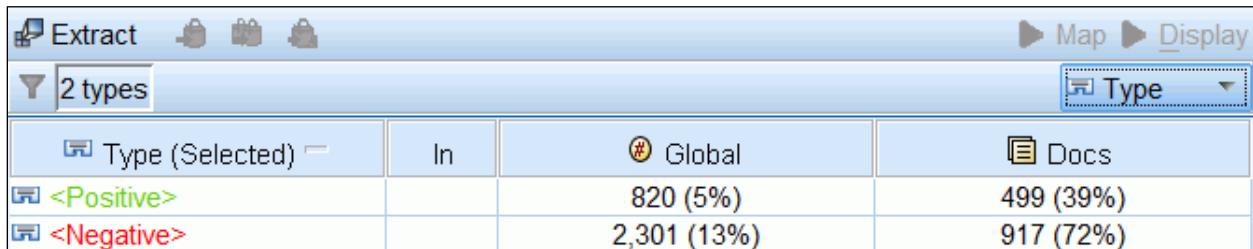
16. Press **Enter**.

Target =	Synonyms	Library
not happy	not happy, not very happy, not at all happy, not so happy	Opinions Library (Engli

17. Switch to the **Categories and Concepts** view.

18. Click **Extract**.
19. From the **Tools** menu, click **Filter**.
20. In the **And by Type** section, click **Selected Types**.
21. Ctrl+Click <Positive> and <Negative>, and then click **Filter**.

The results appear as follows:

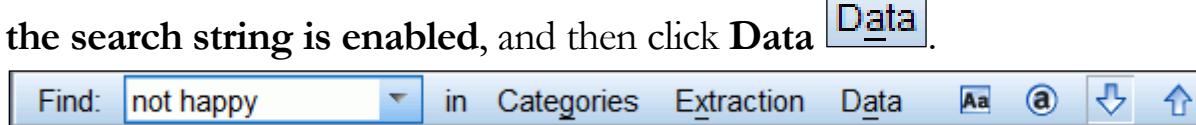


A screenshot of the Extract interface. At the top, there are icons for Extract, Map, and Display. Below that is a toolbar with a filter icon labeled '2 types' and a 'Type' dropdown. The main area shows a table with four columns: 'Type (Selected)', 'In', 'Global', and 'Docs'. There are three rows: one for <Positive> (820, 5%) and one for <Negative> (2,301, 13%), plus a total row for Global (499, 39%) and Docs (917, 72%).

Type (Selected)	In	Global	Docs
<Positive>		820 (5%)	499 (39%)
<Negative>		2,301 (13%)	917 (72%)

The relative number of comments that were typed Positive & Negative numbers have changed considerably from what they were before you modified the dictionary. The Global number of Negative types increased from 2,182 to 2,301 and the number of Positive types decreased from 939 to 820.

22. Click <Negative>, and then click **Display**.
23. In the **Find** box, type **not happy**, ensure that **Match any string containing** the search string is enabled, and then click **Data**.



24. Scroll to case **38** (if necessary).

This comment says that the "Customer is generally not happy with service..."

Now, "not happy" is correctly typed as Negative.

The results appear as follows:

query (917) ▾		Categories
37	Cust is very dissatisfied as he alleges there has been a delay in fixing a no dial tone fault he states he reported on his service number . Cust advised he rejoined the cable under his house by stripp...	
38	Customer is generally not happy with service they have received from ASTROCOMM. Customer advised to wait for half an hour before answered . They feel this is not good enough service .	not happy - <Negative>

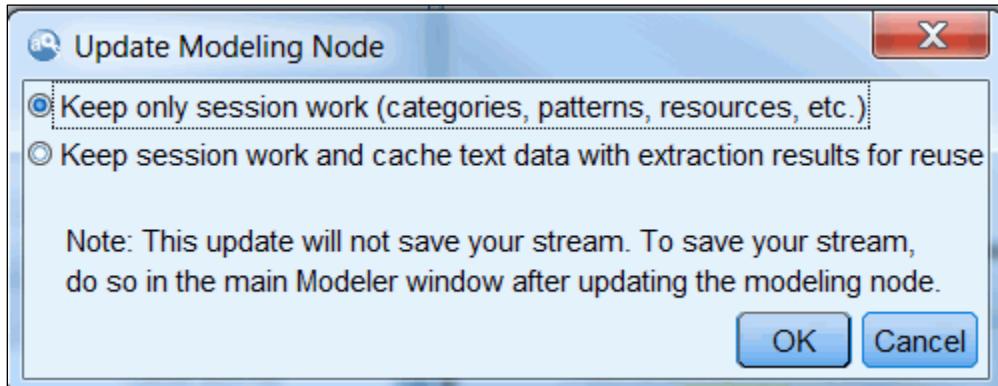
At this point, declare the editing done, at least for now. Save the work so you can easily return to it. You could also publish the Local Library with its modifications, but you will not do this yet. You could even publish the Opinions Library (English) with the change you made if you wanted. However, you should be aware that this is one of the shipped libraries that is loaded by several different templates so may not want to publish it, unless you are certain that the changes you made to how the term "not happy" is treated should be carried on into future projects. As an alternative, you may want to just keep the change local to this project by not publishing the library.

25. From the **Tools** menu, click **Filter**.
 26. Select **All Types**, and then click **Filter**.

Task 6. Saving the changes you made to the dictionary resources.

- From the **File** menu, click **Update Modeling Node**.

The results appear as follows:



- Click **OK**, and then click **OK** to the message that the modeling node has been updated.
- From the **File** menu, click **Close**, and then click **Exit**.
You will save the stream.
- From the **File** menu, click **Save Stream As**.
- Name the stream **Editing Linguistic Resources_Demo1_end.str**, and then click **Save**.
- From the **File** menu, click **Exit** and then click **Exit**.

Results:

You have successfully added a new Type to force the extraction of Astroserve's competitor names. In addition, you corrected a problem in the dictionary resources with is mistakenly recognizing the term "happy" in "not happy" as positive rather than negative sentiment.

Apply your Knowledge

Purpose:

Test your knowledge of the material covered in this module

Question 1: True or false: When editing dictionaries, you must first create a type, then add terms to that type, and then add some of those terms as synonyms to a target.

- A. True
- B. False

Question 2: True or False: You can edit the resources from either the Resource Editor view or the Categories and Concepts view.

- A. True
- B. False

Question 3: True or False: Synonyms allow you to extract terms that were previously not extracted.

- A. True
- B. False

Question 4: True or False: Every concept must have a type.

- A. True
- B. False

Question 5: True or False: Templates contain which of the following:

- A. Categories
- B. Libraries
- C. Text Analysis Packages
- D. Dictionaries

Apply Your Knowledge - Solutions

- Answer 1: B. False. Terms within a type should not be added as synonyms to the target.
- Answer 2: A. True.
- Answer 3: B. False. Types, not synonyms, can be used to force the extraction of terms that was not extracted.
- Answer 4: A. True. Although you do not have to assign a type yourself to every term, the software will automatically type terms. All terms that are not contained in the dictionary are typed as Unknown.
- Answer 5: B, D.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Summary

- At the end of this module, you should be able to:
 - modify the linguistic resources
 - access project libraries and dictionaries
 - create synonyms
 - create types
 - create exclusions

© 2014 IBM Corporation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

7-45

Business Analytics software

IBM

Workshop 1

Editing Dictionaries



© 2014 IBM Corporation

The following file will be used:

- Music_Survey.str - a Modeler stream that reads from a file containing customer likes and dislikes about portable music players

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Workshop 1: Editing Dictionaries

Now that you have finished reviewing the extracted results, at this point, your goal is to modify the linguistic resources prior to categorizing the music survey data.

Changes to be made to the Resources for Music Survey Data #1

	New Type	Terms to Add	Comments
1	storage	lot, all, many, holds, ton	These terms refer to the amount of storage. They could be grouped as synonym definitions, but a type is used because some of these terms were not extracted. Use inflection and the entire and any option because there are several forms for some of these terms
2	little	little	This term was not extracted so a new type needs to be created. Do not use inflection and use the entire term option.

Changes to be made to the Resources for Music Survey Data #2

	New Synonym Target	Synonyms to Add	Comments
1	little	small, palm of hand	Group these terms together since they mean the same thing.

Changes to be made to the Resources for Music Survey Data #3

	Terms to Exclude	Comments
1	little	These terms are redundant and do not add any information to the text because all of the responses refer to music players.

Workshop 1: Tasks and Results

Task 1. Editing dictionaries.

- Create the new types listed above along with their corresponding terms. Think carefully about the appropriate matching option. Re-extract the text to make certain that the changes work as desired.

Resources					
Term	Match	Inflect	Type	Library	
little	Entire Term	<input type="checkbox"/>	little	Local Library	
lot	Entire and Any	<input type="checkbox"/>	storage	Local Library	
all	Entire and Any	<input type="checkbox"/>	storage	Local Library	
many	Entire and Any	<input type="checkbox"/>	storage	Local Library	
holds	Entire and Any	<input type="checkbox"/>	storage	Local Library	
ton	Entire and Any	<input type="checkbox"/>	storage	Local Library	

- Create the new synonym definitions listed above. Re-extract the text to make certain that the changes work as desired.

	Target	Synonyms	Library
1	<input checked="" type="checkbox"/> little	small, palm of hand	Local Library
2	<input checked="" type="checkbox"/> able to log-on	able to log-in, able to login, able to logon, can always log-in, can always log-on, can always login, can always logon, easy to log-in, easy to log-on, easy to login, easy to logon	Opinions Library (English)

- Try excluding the various terms for music player.

	Exclude L...	Library
0	<input type="checkbox"/>	
1	<input checked="" type="checkbox"/> player	Local Library
2	<input checked="" type="checkbox"/> music player	Local Library
3	<input checked="" type="checkbox"/> any kind of prot	Opinions Library
4	<input checked="" type="checkbox"/> any problems i	Opinions Library
5	<input checked="" type="checkbox"/> anykinf of probl	Opinions Library
6	<input checked="" type="checkbox"/> can't wait	Opinions Library
7	<input checked="" type="checkbox"/> i was out of	Opinions Library

- When done, update the modeling node and save the stream as **Music_Survey with Dictionary Edits.str**.
- Close the interactive session and then save the stream as **Music_Survey with Dictionary Edits.str**.
- Exit the Modeler session.

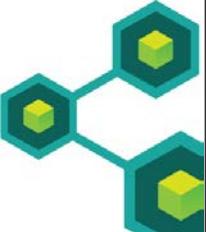
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE



Fine Tuning Resources

IBM SPSS Modeler Text Analytics (v16)

Business Analytics software



© 2014 IBM Corporation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Objectives

- At the end of this module, you should be able to:
 - edit the advanced resources
 - review the Advanced Resources tab
 - add fuzzy grouping exceptions for the Astroserve data
 - create a non-linguistic entity

© 2014 IBM Corporation

The Resource Editor in the Interactive Workbench provides an environment in which you can make the most common changes to the linguistic resources. There are many other changes that can be made to affect how Modeler extracts text, uses types, links concepts together in categories, and identifies text patterns that can be used in categorization. To modify these resources, use the Advanced Resources dialog box.

Many hours of work have gone into providing advanced resources that are appropriate for a variety of resource templates and libraries, so often you will need to modify few, if any, of the advanced resources. Also, modifying some of the resources, such as those defining text link analysis patterns, requires knowledge of the regular expression language and syntax used in those definitions. The syntax can be complex, and many users will not want to spend time learning its intricacies. Fortunately, numerous definitions and specifications are already included in the advanced resources for many types of text.

One change almost all users are likely to make is to the fuzzy grouping that controls how words are grouped to fix spelling errors. Adding words to the fuzzy grouping exceptions is a simple modification.

Fuzzy Grouping

- Fuzzy grouping is a very powerful algorithm to group terms together.
- But it may be too powerful sometimes and group terms that have different meanings.
- It is recommended to use this algorithm only when text is of poor quality or contains many shorthands.

© 2014 IBM Corporation



As has been shown, fuzzy grouping helps to group commonly misspelled words or closely spelled words by temporarily stripping vowels and double or triple consonants from extracted words and then comparing them to see if they are the same. During the extraction process, the fuzzy grouping feature is applied to the extracted terms and the results are compared to determine whether any matches are found. If so, the original words are grouped together in the final extraction list.

Fuzzy grouping is turned off by default. It is controlled by the Accommodate spelling for a minimum root character limit of check box in the Text Mining modeling node Expert tab. There are many pairs of words currently defined as exceptions to prevent them from being grouped. For example, the first pair is "actin" and "action". The word pairs are in alphabetical order, making it easy to search for a word, but there is a search feature, as in the standard Resource Editor view.

Please note that if each term is assigned to a different type, excluding the Unknown type, the fuzzy grouping technique will not be applied.

Non-linguistic Entities

- Non-linguistic entities are based on regular expressions.
- Although they are mainly used to extract «non-words», they can also be useful to detect regular sequences, like postal addresses, amino acids, proteins, .com companies or anything that presents a regular structure.

© 2014 IBM Corporation



Quite often text includes non-words, such as currency, dates, e-mail addresses, or phone numbers. The advanced resources contain regular expressions to allow for the extraction of this type of text, even though these entities are not words.

By regular expression, it is meant a type of language that is used to describe or match a certain set of text, following specific syntax rules. The regular expression language is used regularly in computational linguistics. Although the language is complex, it is not difficult to get the basic gist of an expression.

For example, the line

#@# anything@anything.whatever.etc

is a generic example of an e-mail address. The regular expression:

regexp1=[a-zA-Z0-9._-]+@[a-zA-Z0-9_-]+\.[a-zA-Z0-9]+

is designed to find e-mail addresses and assign them to the e-mail type.

Special Characters: All characters match themselves except for the following special characters, which are used for a specific purpose in expressions: .[{}()*\+?|^\$ To use these characters as such, they must be preceded by a backslash (\) in the definition.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

For example, if you were trying to extract Web addresses, the full stop character is very important to the entity, therefore, you must backslash it such as:

www\.[a-z]+\.[a-z]+

Repetition Operators and Quantifiers: To enable the definitions to be more flexible, you can use several wildcards that are standard to regular expressions. They are * ? +

- * An Asterisk indicates that there are zero or more of the preceding string. For example, ab*c matches "ac", "abc", "abbbc", and so on.
- + A plus sign indicates that there is one or more of the preceding string. For example, ab+c matches "abc", "abbc", "abbbc", but not "ac".
- ? A question mark indicates that there is zero or one of the preceding string. In this example,

```
#@# v12.1, 12.1
regexp1=(v\.|v)?()?[0-9]{1,2}\.[0-9]{1,2}
```

the first ? mark indicates that there could be zero or 1 instance of the letter "v".

The second ? mark indicates that there could be zero or 1 space before the version number.

- {} A limiting repetition with brackets indicates the bounds of the repetition. For example, [0-9]{n} matches a digit repeated exactly n times. For example, [0-9]{4} will match “1998”, but neither “33” nor “19983”.

Optional Spaces and Hyphens: In some cases, you want to include an optional space in a definition. For example, if you wanted to extract currencies such as "canadian dollars", "canadian dollar", "canada dollars", you would need to deal with the fact that there may be two words separated by a space. In this case, this definition should be written as (canadian | canada)?dollars?. Since canadian or canada are followed by a space when used with dollars/dollar, the optional space must be defined within the optional sequence (canadian | canada). If the space was not in the sequence such as (canadian | canada)? dollars?, it would not match on “dollars” or “dollar” since the space would be required.

If you are looking for a series of things including a hyphen characters (-) in a list, then the hyphen must be defined last. For example, if you are looking for a comma (,) or a hyphen (-), use [,-] and never [-,].

Order of Strings in Lists and Macros: You should always define the longest sequence before a shorter one or else the longest will never be read since the match will occur on the shorter one. For example, if you were looking for strings "billion" or "bill". So for instance (billion | bill) and not (bill | billion). This also applies to macros, since macros are lists of strings.

Order of Rules in the Definition Section: Define one rule per line. Within each section, rules are numbered (regexp1, regexp2, and so on). These rules must be numbered sequentially from 1–n. Any break in numbering will cause the processing of this file to be suspended altogether. To disable an entry, place a # symbol at the beginning of each line used to define the regular expression. To enable an entry, remove the # character before that line.

Using Macros in Rules: Whenever you use a specific sequence in several rules, you can use a macro. Then, if you need to change the definition of this sequence, you will need to change it only once, and not in all the rules referring to it. For example, assuming you had the following macro:

```
MONTH=((january|february|march|april|june|july|august|september|october|november|december)|(jan|feb|mar|apr|may|jun|jul|aug|sep|oct|nov|dec)(\.)?)
```

Whenever you refer to the name of the macro, it must be enclosed in \$(), such as:
regexp1=\$(MONTH)

All macros must be defined in the [macros] section.

Normalizing Non-Linguistic Entities

- Non-linguistic entities are normalized according to a predefined format.
- For example, currency symbols and their equivalents are treated the same (for example, united states dollars = u.s. dollars = us dollars = usd = \$ us, etc.).

© 2014 IBM Corporation



When extracting nonlinguistic entities, the entities encountered are normalized to group like entities according to predefined formats. For example, currency symbols and their equivalent in words are treated as the same. The Normalization file is broken up into distinct sections based on the entity, including type of currency, numbers, and weights and measures.

By default dates in an English template are recognized in the American style date format; that is: month, date, year. If you need to change that to the day, month, year format, disable the "format:US" line (by adding # at the beginning of the line) and enable "format:UK" (by removing the # from that line).

You can add new information to these definitions, such as other ways that Australian dollars are referenced, but you should still be careful when doing so.

Configuration

- Enable nonlinguistic extraction if you want Text Analytics to extract nonlinguistic entities, such as phone numbers, social security numbers, times, dates, currencies, digits, percentages, e-mail addresses, and HTTP addresses.
- You can include or exclude certain types of nonlinguistic entities in the Nonlinguistic Entities: Configuration section of Advanced Resources.

© 2014 IBM Corporation



You can enable and disable the nonlinguistic entity types that you want to extract in the nonlinguistic entity configuration file. By disabling the entities that you do not need, you can decrease the processing time required.

If nonlinguistic extraction is enabled, the extraction engine reads this configuration file during the extraction process to determine which nonlinguistic entity types should be extracted. You can enable and disable the nonlinguistic entity types that you want to extract in this file. By disabling the entities that you do not need, you can decrease the processing time required.

File entries in the file are written in the following form:

#name<TAB>Language<TAB>Code

For example:

Name	Language	Code
IP	0	s

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

In this instance, the non-linguistic name is IP, the language is, and the "s" means that IP is a stop word. Sometimes, some extremely common words which would appear to be of little value and can be safely filtered out. In English, some of these ignorable components might include a, and, as, by, for, from, in, of, on, or, the, to, and with. For example, the term examination of the data has the component set {data, examination}, and both "of" and "the" are considered ignorable.

In small files, disabling unnecessary non-linguistic entities will not make much difference in processing time, but in larger files, you should experiment to see what time can be gained.

Here is a complete description of the syntax for the configuration file:

Column Label	Description
Name	The name by which nonlinguistic entities will be referenced for nonlinguistic entity extraction. The names used here are case sensitive
Language	The language of the documents. Possible options are: 0 = Any which is used whenever a regular expression is not specific to a language and could be used in several templates with different languages, for instance an IP/URL/email addresses; 1 = French; 2 = English; 4 = German; 5 = Spanish; 6 = Dutch; 8 = Portuguese; 10 = Italian.
Code	Part-of-speech code. Most entities take a value of "s" except in a few exceptions. Possible values are: s = stop word; a = adjective; n = noun. For example, percentages are given a value of "a." Suppose that 30% is extracted as a nonlinguistic entity. It would be identified as an adjective. Then if your text contained "30% salary increase," the "30%" nonlinguistic entity fits the part-of-speech pattern "ann" (adjective noun noun).

Language Handling

- Used to declare the special ways of structuring sentences
 - extraction patterns
 - forced definitions
 - using abbreviations for the selected language

© 2014 IBM Corporation



All languages have unique sentence structure and abbreviations. The advanced resources include parts-of-speech (POS) patterns for each language. Parts of speech include grammatical elements, such as nouns, adjectives, past participles, determiners, prepositions, coordinators, first names, initials, and particles. A series of these elements makes up an extraction pattern.

These patterns are used to identify candidate words for extraction based on their position in text. The resources also include common abbreviations for a language to help extract all important information. You can make modifications to both of these, although it is more common to modify the abbreviations.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Extraction Patterns

- The extraction engine applies a set of parts-of-speech extraction patterns to a "stack" of words in the text to identify candidate terms (words and phrases) for extraction.
- You can add or modify the extraction patterns.

© 2014 IBM Corporation



In text mining, each part of speech is represented by a single character to make it easier to define the patterns. Parts of speech include grammatical elements, such as nouns, adjectives, past participles, determiners, prepositions, coordinators, first names, initials, and particles. A series of these elements makes up a part-of-speech extraction pattern. Each part of speech is represented by a single character to make it easier to define your patterns. For instance, an adjective is represented by the lowercase letter a, while a v references a verb. The set of supported codes appears by default at the top of each default extraction patterns section along with a set of patterns and examples of each pattern to help you understand each code that is used.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

8-12

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

POS Macro	Part of Speech
a	adjective
b	adverb (not used in the English opinions template)
c	preposition ("of") (not used in the English Opinions template - see "r")
C	misspelling (not used in the English Opinions template)
d	determiner ("the") (not used in the English Opinions template - see "e")
e	extended determiner (for example "the", "my", "those", "certain" ...)
f	first name
G	geography ("American", ...) (not used in the English Opinions template)
i	middle initial in person name
m	noun(n) or unknown(u)
n	noun
o	coordination ("and" and "&")
p	past participle
r	extended preposition (for example "of" "for", "to", "provided by" ...)
s	stop word (not used in the English Opinions template)
t	title
u	unknown
v	verb (any verb, infinitive/gerund/3rd person singular/preterite)
V	verb (infinitive)

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

8-13

POS Macro	Part of Speech
x	auxiliary verb ("be", "is", ...) (not used in English Opinions template)
y	particle ("von", "de", ...)
0	opinion adverb (perfectly", "badly"...) - see list under Forced Definitions
1	"to"- see list under Forced Definitions
2	introduction qualifier ("easy", "uneasy", ...) - see list under Forced Definitions

The main body of the definitions for POS extraction patterns contains an example of a part of speech pattern ("maintain contact"), followed by the definition of the part of speech pattern itself using the appropriate characters (Vm). Although it is easy to add POS patterns, users probably won't have to do so unless you will be using them for text link analysis.

The order in which you list the extraction patterns is very important because a given sequence of words is read only once by the extraction engine and is assigned to the first extraction patterns for which the engine finds a match.

The formatting rules for Extraction Patterns are as follows:

- One pattern per line.
- Use # at the beginning of a line to disable a pattern.

Here is an example of one of the extraction patterns in the file:

```
# message received from tech support
mprmm
```

The first line, is a sample sentence that matches the pattern **mprrmm**, which translates to noun, past participle, extended preposition, noun, noun. Whenever the extraction engine finds a match, the term is extracted.

Forced Definitions

- A word can fit several different roles depending on the context.
- In the Forced Definitions portion of Advanced Resources, you can force a word to take a particular part-of-speech or to exclude the word completely from processing.

© 2014 IBM Corporation



If you want to force a word to take a particular POS or to exclude the word completely from POS processing, you can do so in the Forced Definition section.

It is also possible to prevent words from being extracted in the Forced Definitions section by using the lowercase s as a part-of-speech code to stop a word from being extracted altogether.

The format for a forced definition is:

Term:code

Entry	Description
term	A term name
code	A single-character code representing the part-of-speech role. You can list up to six different part-of-speech codes per uniterm. Additionally, you can stop a word from being extracted into compound words/phrases by using the lowercase code s, such as additional:s.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

You can use the asterisk character (*) as a wildcard at the end of a string for partial matches. For example, if you enter add*:s, words such as add, additional, additionally, addendum, and additive will never be extracted as a term or as part of a compound word term. However, if a word match is explicitly declared as a term in a compiled dictionary or in the forced definitions, it will still be extracted. For example, if you enter both add*:s and addendum:n, addendum will still be extracted if found in the text.

When you want to exclude a word, it is better to use Advanced Resources /Forced Definitions with PoS ":s" than to use the Exclude dictionary. This is because with the Exclude dictionary you can lose terms, synonyms, and are discarding simple and compound terms.

Abbreviations

- The extraction engine generally considers any period it finds as an indication that a sentence has ended.
- This is typically correct but not when abbreviations are contained in the text.
- If you find that certain abbreviations were mishandled, you should explicitly declare that abbreviation in this section.

© 2014 IBM Corporation

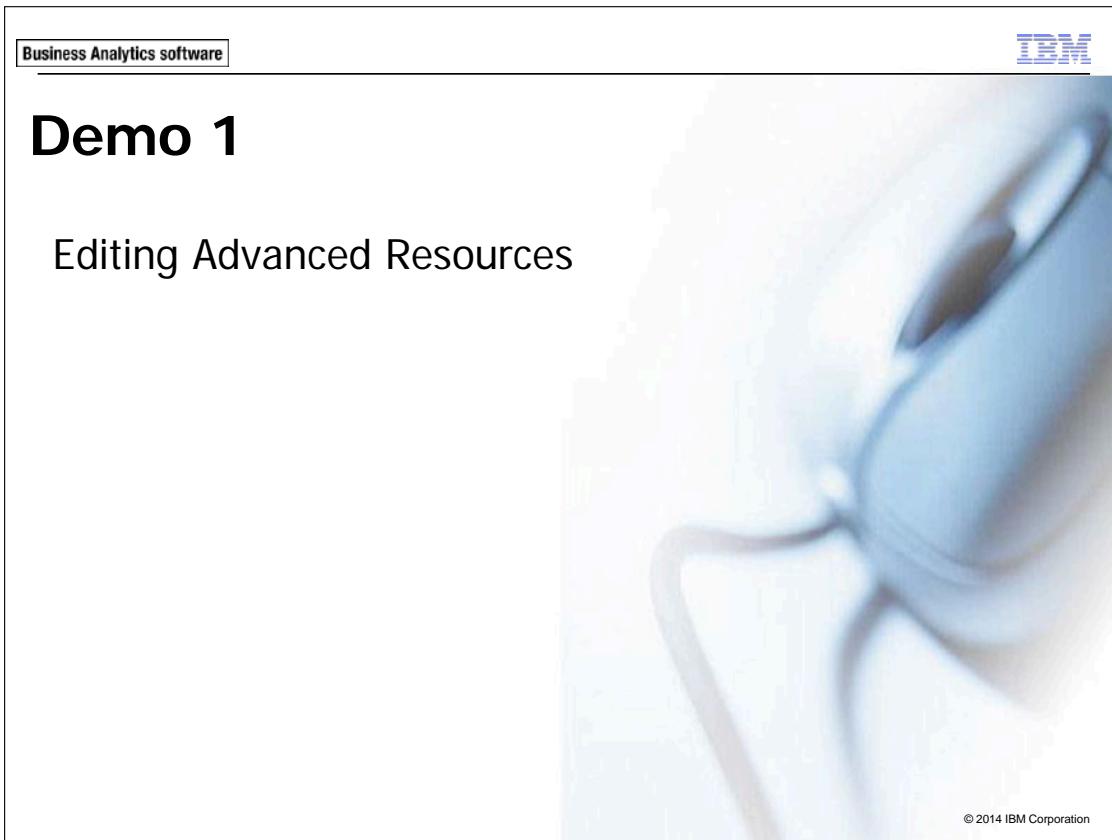


When the extractor engine is processing text, it will generally consider any period it finds as an indication that a sentence has ended. This is typically correct but not when the period follows an abbreviation.

Many common abbreviations are included with the default resources for each language. For English, the abbreviations include "mr.", "dept.", and all the months (for example, "sept."). An abbreviation that does not end with a period is not affected by these definitions.

If the abbreviation already appears in a synonym definition or is defined as a term in a type dictionary, there is no need to add the abbreviation entry here.

Abbreviations are added one per line, and they are not case sensitive.



The slide has a light blue background with a faint, abstract graphic of a person's head and shoulders. In the top left corner, there is a small rectangular box containing the text "Business Analytics software". In the top right corner, the "IBM" logo is displayed. The main title "Demo 1" is centered at the top in a large, bold, black font. Below it, the subtitle "Editing Advanced Resources" is also centered in a smaller, regular black font. At the bottom right of the slide, there is a small, horizontal text "© 2014 IBM Corporation".

This demo uses the following files:

- C:\Train\0A105\08-Fine_Tuning_Resources\Fine Tuning Resources_Module 8_start.str - a Modeler stream that reads a file containing call center data for March and April
- C:\Train\0A105\08-Fine_Tuning_Resources\Nonlingistic Entity Definition.txt

Demo 1: Editing Advanced Resources

Purpose:

While the resources libraries that were shipped with the software may be sufficient for most purposes, often you need to modify the advanced resources so that the software extracts concepts that would not otherwise be extracted. In the case of the Astroserve data, you need to force it to extract phone types along with model numbers.

Task 1. Adding to the list of fuzzy group exceptions.

While analyzing the Astroserve data, while examining the extracted results, you noticed there are several pairs of words that the fuzzy grouping algorithm paired together that should not be considered equivalent:

Moist most

Stall stale

Fail fall

These words are important not to confuse. If you search for these, you will see they are not currently used in any fuzzy grouping exception pair.

1. From the **File** menu, click **Open Stream**.
2. Navigate to **C:\Train\0A105\08-Fine_Tuning_Resources** and then double-click **Fine Tuning Resources_Module 8_start.str**.
3. Run the **Text Mining** node.
4. Switch to the **Resource Editor** view.
5. Click the **Advanced Resources** tab.

- Click **Fuzzy Grouping \ Exceptions** to view the exceptions pairs that are already defined.

The results appear as follows:

actin	action
active site	activist
albania	albany
alberta	alberti
amine	amino
analog	analogy
andersen	anderson
antarctic	antarctica
appellation	appleton
arm	army
attention	attenuation
bacillus	baycol
bacteremia	bacterium
bakery	bakeractin
active site	activist
albania	albany
alberta	alberti
amine	amino

- Put the cursor before the word **actin**, and then press **Enter**.
- Type **waist**, press **Tab**, and then type **waste**.
- Repeat step 8 to add the word pairs as follows.

waiter	waitress
waiter	water
waiver	waver
water	watery
weakened	week-end
weakened	week-ends
weakened	weekends
weakened	weekend
winery	winner
yemen	yeoman
yugoslav	yugoslavia
moist	most
stall	stale
fail	fall

- Switch to the **Categories and Concepts** view.
- Click **Extract**.
- From the **File** menu, click **Update Modeling Node**.
- Click **OK**, and then click **OK** to the message that the modeling node has been updated.

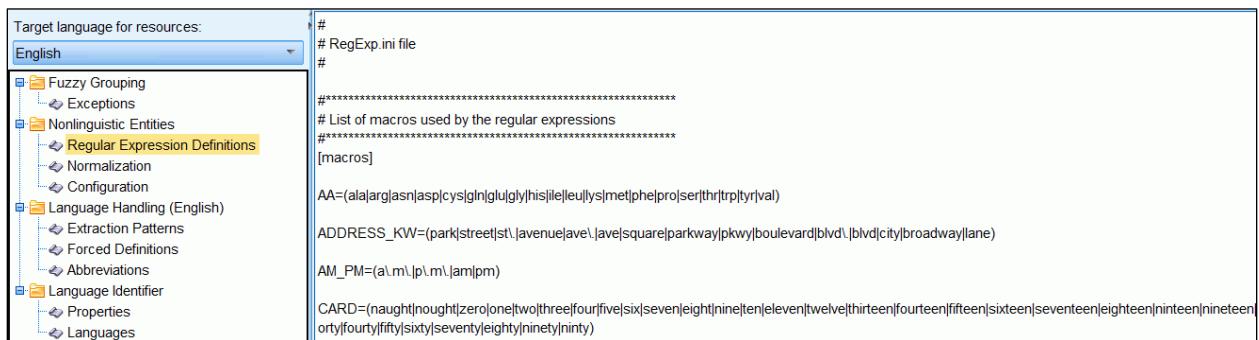
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Task 2. Adding non-linguistic entities.

A few of the Astroserve customers referred to specific model numbers for their mobile phone. For example, several customers mentioned specific models of Nokia phones in their complaints (Nokia 9950, Nokia 6650, etc.) but just the term "nokia" was extracted. Because it would be of interest later on to examine caller sentiment toward specific models of Nokia phones, you will need to create a non-linguistic entity so both the brand name and model number are extracted.

1. Open the file **C:\Train\0A105\08-Fine_Tuning_Resources\Nonlinguistic Entity Definition.txt** in **Notepad**.
2. In the **Interactive Workbench**, switch to the **Resource Editor** view.
3. Ensure that the **Advanced Resources** tab is selected, and then below **Nonlinguistic Entities** click **Regular Expression Definitions**.

At the top, you will see a list of macros, starting with AA, ADDRESS, etc.



```

Target language for resources: English
# RegExp.ini file
#
#####
# List of macros used by the regular expressions
#####
[macros]
AA=(ala|arg|asn|asp|cys|gln|glu|gly|his|iie|leu|lys|mef|phe|pro|ser|thr|trp|tyr|val)
ADDRESS_KW=(park|street|st|avenue|ave|ave|square|parkway|pkwy|boulevard|blvd|blvd|city|broadway|lane)
AM_PM=(a|m|pm|am|pm)
CARD=(naught|nought|zero|one|two|three|four|five|six|seven|eight|nine|ten|eleven|twelve|thirteen|fourteen|fifteen|sixteen|seventeen|eighteen|nineteen|nineteen|forty|fifty|sixty|seventy|eighty|ninety|ninety)


```

4. Scroll down until you see the macro named **YEAR**.
5. Copy and paste the **PHONES** macro at the top of the file **Nonlinguistic Entity Definition.txt** to just below the **YEAR** macro.

<pre> W_A_M_SYMB=(mhz mm ml mb cm kg oz mg km ft k -bps tb gb kb hz k m² m³ m b) YEAR=([12][0-9]{3} [0-9]{2}) PHONES=(nokia hyundai iphone kyocera motorola ericsson) </pre>
--

6. Scroll to the bottom of the Regular Expressions Definitions.

7. Copy and paste the definitions for the PHONES nonlinguistic entity from Nonlinguistic Entity Definition.txt to the bottom.

The results appear as follows:

```
*****
#          PROTEINS
*****
[Gene]

#@# p110
regexp1=p[0-9]{2,3}

regexp2=[a-zA-Z]{2,4}-?[0-9]{1,3}-?[rR]

regexp3=[a-zA-Z]{2,4}-?[0-9]{1,3}p?

caseSensitive=1
accentSensitive=0

*****
#          Phones
*****
[Phones]

#@# iPhone 5C
regexp1=$(PHONES)( )?$(NUM)[c|s]

#@# nokia 9950
regexp2=$(PHONES)( )?$(NUM)

#@# motorola T91, motorola e668
regexp3=$(PHONES)( )?[t|e]( )?$(NUM)

caseSensitive=0
accentSensitive=0
```

8. Below **Nonlinguistic Entities** click **Configuration**.
 9. Copy and Paste the configuration definition for **Phones** at the bottom of the file nonlinguistic entity definitions.txt to just below the first line.
- The results appear as follows:

#name	Language	PoS
Phones	0	s
IP	0	s
url	0	s
email	0	s
PhoneNumber	2	s
SocialSecurityNumber	2	s

10. Close the **Nonlinguistic Entity Definition.txt** file.
11. Switch to the **Categories and Concepts** view.
12. Click **Extract**.
13. Click **Find Toolbar**, in the **Find** box, type **Nokia**, and then click **Extraction**.
14. Click the **Type** column header to sort the concepts by type.

Concept	In	Global	Docs	Type
mr ang		5 (0%)	1 (0%)	<Person>
nokia 8310		2 (0%)	2 (0%)	<Phones>
nokia 6100		1 (0%)	1 (0%)	<Phones>
nokia 9950		1 (0%)	1 (0%)	<Phones>
nokia 7650		1 (0%)	1 (0%)	<Phones>
nokia 7250		2 (0%)	2 (0%)	<Phones>
nokia 6610		1 (0%)	1 (0%)	<Phones>
nokia 6385		1 (0%)	1 (0%)	<Phones>
nokia 3650		1 (0%)	1 (0%)	<Phones>
genuine		1 (0%)	1 (0%)	<Positive>
accurate		43 (0%)	41 (3%)	<Positive>

It correctly extracted the term Nokia along with model number.

Task 3. Saving the non-linguistic entities.

1. From the **File** menu, click **Update Modeling Node**.
2. Click **OK**, and then click **OK** to the message that the modeling node has been updated.
3. From the **File** menu, click **Close**, and then click **Exit** to end the interactive session.
4. From the **File** menu, click **Save Stream As**.
5. Name the stream **Fine Tuning Resources.str**.
6. Click **Save**.
7. From the **File** menu, click **Exit** to close the Modeler session.

Result:

You have successfully modified the advanced resources by adding to the fuzzy group exceptions, and by adding a non-linguistic entity which you designed to extract phone types that customers are complaining about along with their model numbers.

Apply Your Knowledge

Purpose:

Test your knowledge of the material covered in this module.

- Question 1: True or false: The fuzzy grouping technique applies to all concepts unless they are of the Unknown type.
- A. True
 - B. False
- Question 2: True or False: You can enable and disable the nonlinguistic entity types that you want to extract in the nonlinguistic Entities Normalization file.
- A. True
 - B. False
- Question 3: True or False: When you want to exclude a word, it is better to use Advanced Resources /Forced Definitions with PoS ":s" than to use the Exclude dictionary.
- A. True
 - B. False
- Question 4: True or False: The extraction engine applies a set of parts-of-speech extraction patterns to a "stack" of words in the text to identify candidate terms (words and phrases) for extraction.
- A. True
 - B. False
- Question 5: True or False: You cannot add or modify the extraction patterns file.
- A. True
 - B. False

Apply Your Knowledge - Solutions

Answer 1: B. False.

Answer 2: B. False. This can be done in the Configuration file

Answer 3: A. True

Answer 4: A. True

Answer 5: B. False

Summary

- At the end of this module, you should be able to:
 - edit Advanced Resources
 - review the advanced resources tab
 - add fuzzy grouping exceptions for the Astroserve data
 - create a non-linguistic entity

© 2014 IBM Corporation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

8-27

Business Analytics software

IBM

Workshop 1

Editing Advanced Resources



© 2014 IBM Corporation

No supporting materials are needed for this workshop.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

8-28

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Workshop 1: Editing Advanced Resources

Question 1: True or False: If the terms "arm" and "army" were not given the same type (and neither was of the Unknown type) fuzzy groupings could group these terms together.

- A. True
- B. False

Question 2: True or False: The order in which POS patterns are listed is very important.

- A. True
- B. False

Question 3: True or False: In order for any abbreviation to be extracted correctly, it must be specified in the abbreviations section.

- A. True
- B. False

Question 4: True or False: Which of the following options in the Expert tab is used to turn the Fuzzy Grouping algorithm on or off?

- A. Accommodate punctuation errors
- B. Extract uniterms
- C. Uppercase algorithm
- D. Accommodate spelling for a minimum root character limit of

Question 5: True or False: Which option in Advanced Resources would you use to select which nonlinguistic entities you want extracted? Of course, this assumes that you checked the Extract nonlinguistic entities box in the Expert tab.

- A. Fuzzy Grouping \ Exceptions
- B. Language Handling (English) \ Forced Definitions
- C. Nonlinguistic Entities \ Regular Expression Definitions
- D. Nonlinguistic Entities \ Configuration

Workshop 1: Tasks and Results

- Answer 1: B. False. The Fuzzy Grouping Algorithm will not group terms together if they have different types and neither was typed Unknown
- Answer 2: A. True
- Answer 3: B. False
- Answer 4: A. Accommodate spelling for a minimum root character limit of
- Answer 5: D. Nonlinguistic Entities \ Configuration



Performing Text Link Analysis

IBM SPSS Modeler Text Analytics (v16)

Business Analytics software



© 2014 IBM Corporation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Objectives

- At the end of this module, you should be able to:
 - perform Text Link Analysis
 - review interactive Text Link Analysis
 - review the Text Link Analysis node
 - create text link rules and macros

© 2014 IBM Corporation

After extracting concepts from text, and creating types to group related concepts under a higher level topic, the next step in a text mining project is often to look for patterns, or relationships, between the concepts and\or between the types. Text Link Analysis (TLA) provides a way to identify and extract patterns from the text and then use the pattern results. Patterns are most useful when your are attempting to discover relationships between concepts or opinions about a particular subject. An example of a pattern would be discovering that the concepts of "fast" and "service" are often reported together in the text and are referencing each other.

Text Link analysis can be accessed from either the Interactive Workbench or from its own Text Link Analysis node. From the Interactive Workbench, once you have extracted some TLA patterns, you can explore them in the Data or Visualization panes and even add them as categories in the Categories and Concepts view. The Text Link Analysis node restructures the data and outputs concept patterns as new records that you can use in various analyses, including modeling. Both types of text link analysis will be tried in this lesson, but the focus here will be on its use through the Interactive Workbench.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

9-3

There must be some TLA pattern rules defined in the resource template you load in order to extract TLA results, and only some of the templates include these patterns. For example the Opinions template does. You can always add TLA pattern rules, but this requires understanding the use of regular expressions and the syntax by which TLA rules are defined in Modeler.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

9-4

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

The screenshot shows the IBM Business Analytics software interface for Text Link Analysis. The interface is organized into four main sections:

- Type Pattern Pane:** Located in the top-left quadrant, it contains a navigation bar with links: Global | In | Type 1 | Type 2 | Type 3 | Type 4.
- Visualization Pane:** Located in the top-right quadrant, it contains a navigation bar with links: Global | Docs | In | Concept 1 | Concept 2.
- Concept Pattern Pane:** Located in the bottom-left quadrant.
- Data Pane:** Located in the bottom-right quadrant.

At the bottom right of the interface, there is a decorative graphic consisting of several interconnected hexagons in shades of blue, green, and yellow.

The Type Pattern pane and Concept Pattern pane on the left side, are interconnected panes in which you can explore and select the TLA pattern results. Patterns are made up of a series of types or concepts. Pattern results are first grouped at the type level and then divided into concept patterns. For this reason, there are two different result panes: Type Patterns and Concept Patterns. The Visualization pane, enables you to visually explore how the concepts and types in your patterns interact in the pane. The Data pane lets you explore and review text contained within documents and records that correspond to selections in another pane.

Business Analytics software

IBM

Type Patterns Pane

Global	In	Type 1	Type 2
7095		<UNKNOWN>	
1050		<NEGATIVE>	
963		<UNKNOWN>	<NEGATIVE>
881		<BUDGET>	
69		<BUDGET>	<NEGATIVE>
2		<PHONES>	<NEGATIVEFUNCTIONING>

© 2014 IBM Corporation



The Type pane presents pattern results consisting of one or more related types matching a TLA pattern rule. Type patterns are shown as <Phones> + <Negative>, or <Phones> + <Positive>, which might provide positive or negative feedback about customer attitudes toward specific kinds of phones. In this example, two customers contacted the call center to report that their phones were not functioning properly. To get more specific, you need to switch to the Concept Pattern pane.

The syntax is as follows:

<Type1> + <Type2> + <Type3> + <Type4> + <Type5> + <Type6>

Patterns can consist of up to six types. For this reason, the rows in both patterns panes contain up to six slots, or positions. Each slot corresponds to an element's specific position in the TLA pattern rule as it is defined in the linguistic resources. In the Interactive Workbench, if a slot contains no values it is not shown in the table.

Concept Patterns Pane

Global	Docs	In	Concept 1	Concept 2	Concept 3
1	1		nokia 7250	not working	
1	1		nokia 7250	replaced	

© 2014 IBM Corporation



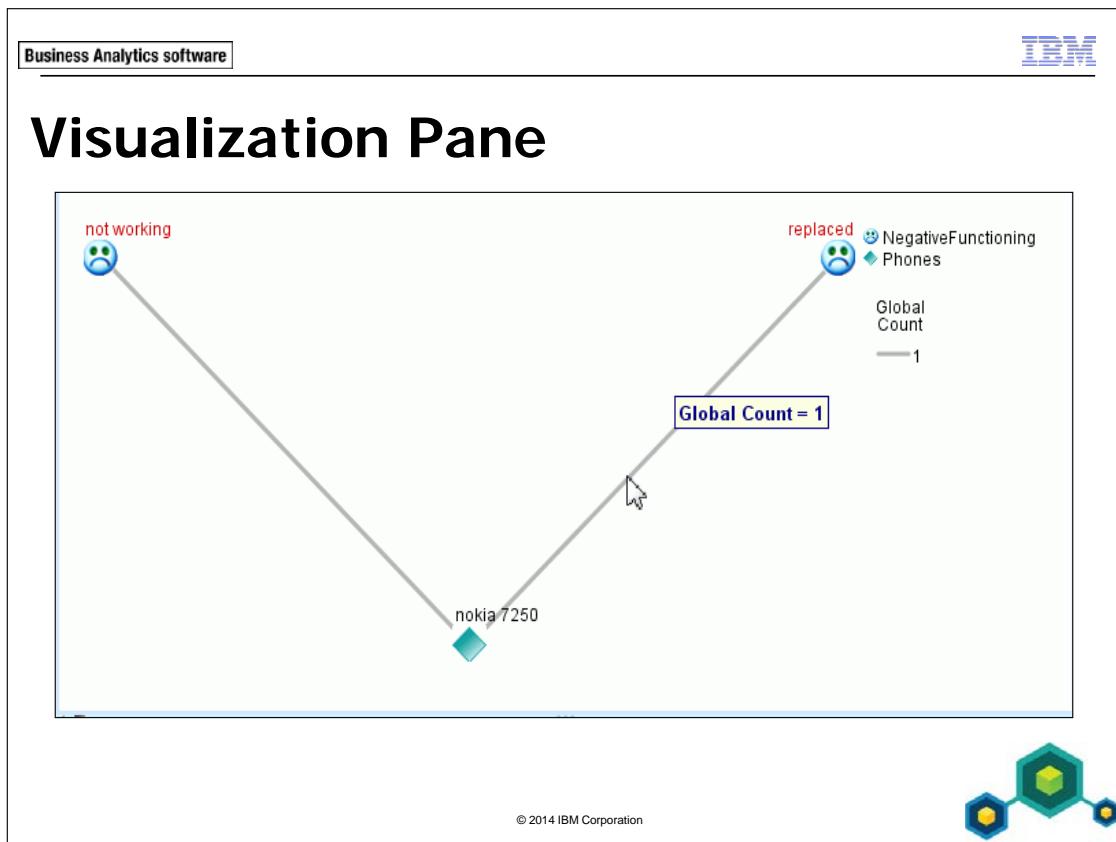
The Concept Pattern pane presents the pattern results at the concept level for all of the type pattern(s) currently selected in the Type Patterns pane above it. Concept patterns follow a structure such as "nokia 7250" + "not working".

The syntax is as follows:

concept1 + concept2 + concept3 + concept4 + concept5 + concept6

Patterns can consist of up to six types, which is why the pane contains up to six slots, or positions. Each slot corresponds to an element's specific position in the TLA pattern rule as it is defined in the linguistic resources. When pattern uses less than the six maximum slots, only the necessary number of slots (or columns) are displayed.

Now you can get a better idea of exactly what these two customers were experiencing: one customer's Nokia 7250 was not working and another said theirs had to be replaced.



The Visualization Pane displays a concept Web graph of the patterns. This Web graph presents all of the concepts represented in the current selection. In this case, you selected the <Phones> + <NegativeFunctioning> pattern in the Type Pattern pane. If you selected a type pattern that had three matching concept patterns, this graph would show three sets of linked concepts. The line width represents the global frequency counts. The nodes are represented by an icon that characterizes the type of that concept. A frowning face is used for a negative functioning type. A smiley face is used to represent Positive sentiment. The type of each concept can be represented by the type color or an icon depending on what you select on the graph toolbar.

In this example, the two frowning faces clearly indicate that two customers were dissatisfied with their Nokia 7250 phones.

TLA Graph Layouts

- Text Link Analysis graphs can be displayed in four different layouts:
 - grid layout
 - circle layout
 - network layout
 - DAG layout

© 2014 IBM Corporation



The default layout of the graph is a grid layout; other layouts can be selected by clicking a button on the toolbar. The layouts include:

- Grid Layout A layout that assumes that links are undirected and treats all nodes the same. Nodes are only placed at grid points within the space.
- Circle Layout A layout that assumes that links are undirected and treats all nodes the same. Nodes are placed around the perimeter of a circle as much as possible.
- Network Layout A layout that assumes that links are undirected and treats all nodes the same. Nodes are placed freely within the layout to maximize viewability. This is often a good choice, at least for initially viewing graphs.
- DAG Layout A layout that is normally used for directed graphs (as with the Web node in Modeler). This layout produces treelike structures from root nodes down to leaf nodes and organizes by colors. Hierarchical data tends to display well with this layout.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Business Analytics software

IBM

Data Pane

query (2) ▾	
1	Aaron has a nokia 7250 which has been replaced 3 times due to faults within the handset. The customer does not want this phone anymore due to the amount of times it has become faulty and wants to know what his options are.
2	Customer has had multiple handset failure with nokia 7250, faults have been that handset will have excessive call drop out and screen freezing , customer wants to upgrade with \$0 early termination charges

© 2014 IBM Corporation



As you extract and explore text link analysis patterns, you may want to review some of the data you are working with. For example, you may want to see the actual records in which a group of patterns were discovered. You can review records or documents in the Data pane, which is located in the lower right. The Data pane presents one row per document or record corresponding to a selection in the view, up to a certain display limit. By default, the number of documents or records shown in the Data pane is limited in order to make it faster for you to see your data. However, you can adjust this in the Options dialog box.

In the top example, the customer reported that he replaced his Nokia 7250 three times due to faults in the handset. The second customer, called to report that their Nokia 7250 handset had failed multiple times and requested an upgrade with zero termination charges.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Demo 1

Exploring Text Link Patterns in the Interactive Workbench

© 2014 IBM Corporation

The following file(s) are used in this demo:

- Performing_Text_Link_Analysis_demo1_start.str - a Modeler stream that reads a file containing call center data for March and April

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

9-11

Demo 1: Exploring Text Link Patterns in the Interactive Workbench

Purpose:

Your goal is to use Text Link Analysis to identify patterns in the data. In particular, you will identify the concepts or types that are linked with negative sentiment.

Task 1. Starting the Text Link Analysis session.

1. From the **File** menu, click **Open Stream**.
2. Navigate to **C:\Train\0A105\09-Performing_Text_Link_Analysis**, and then double-click **Performing Text Link Analysis_demo1_start.str**.
3. Edit the **Text Mining** node labeled **query**.
4. Click the **Model** tab, and ensure that **Exploring text link analysis (TLA) results** is selected.

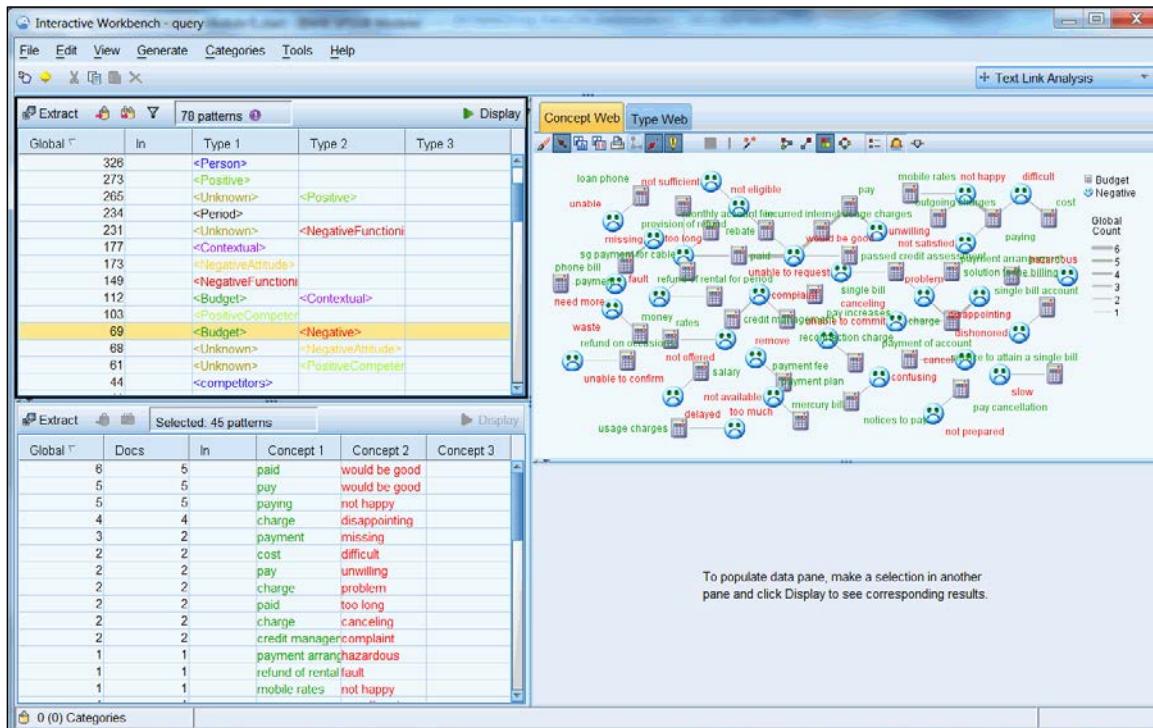
Begin session by:

- Using extraction results to build categories
- Exploring text link analysis (TLA) results
- Analyzing co-word clusters

5. Click **Run**.

Task 2. Examining the results of the Text Link Analysis.

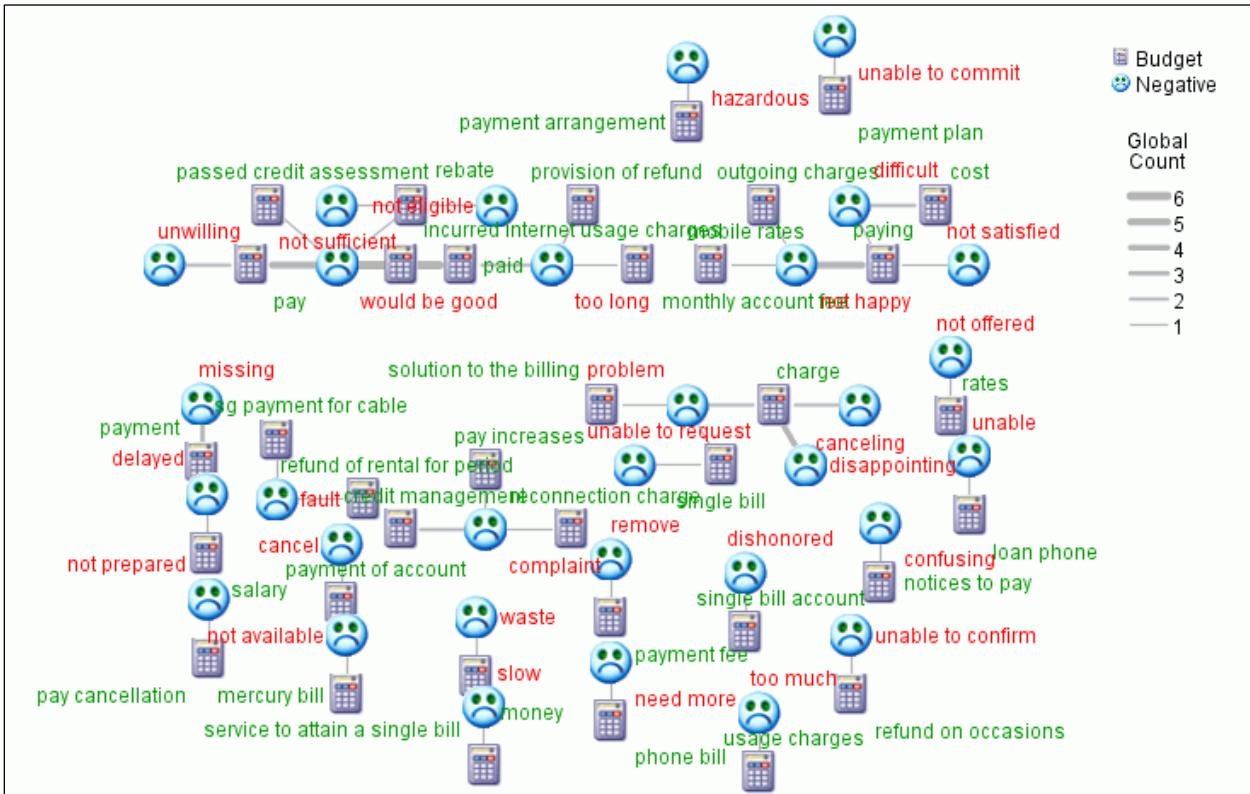
- Click the <Budget> + <Negative> pattern in the Type Patterns pane.



The negative concepts are represented by the frowning face, while the budget concepts are represented by the calculator icon.

2. Click on the **DAG layout** tool  button in the toolbar to change to a DAG layout.

Your results may differ somewhat from the following:



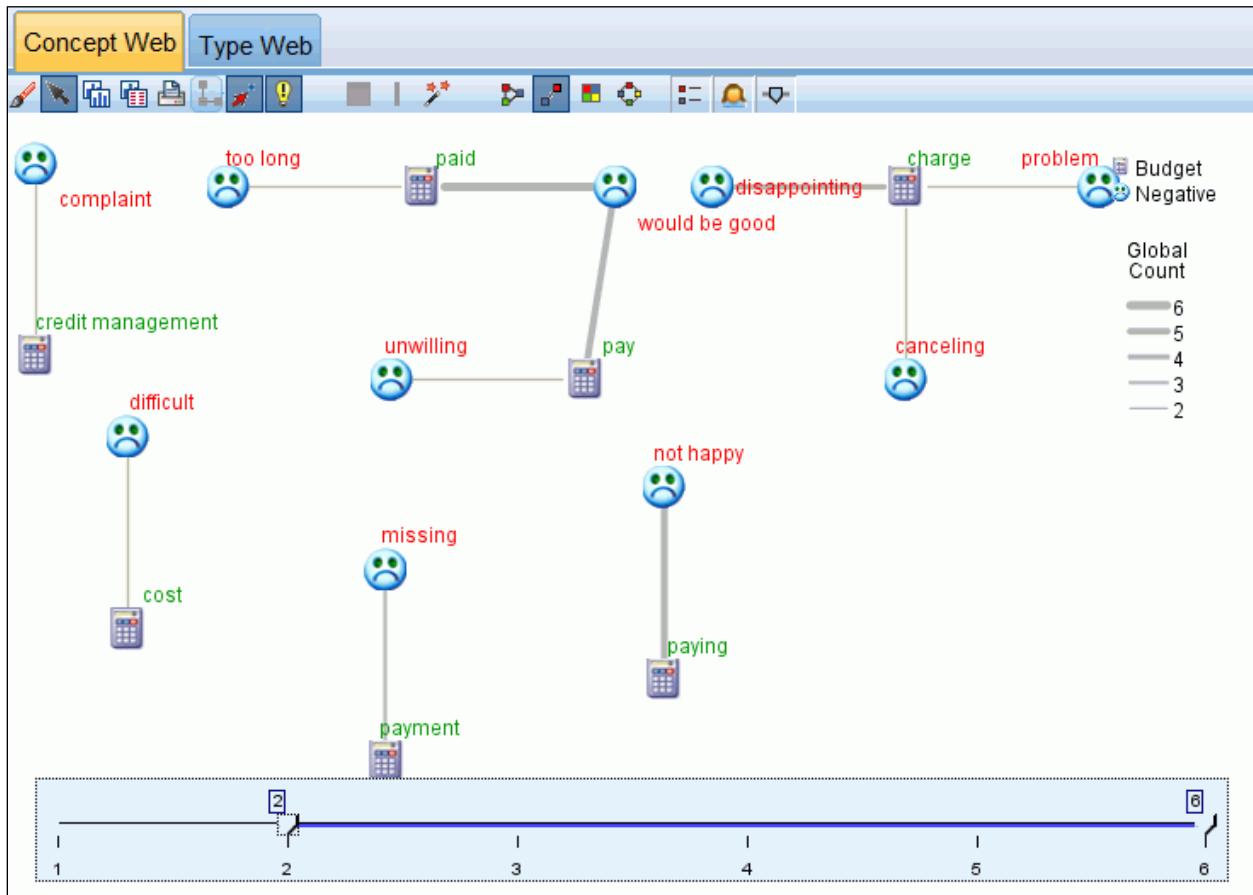
It is now straightforward to see the various patterns. Notice how the concept "not satisfied" is connected to the "paying" concept. Also, observe how some concepts are in more than one pattern. For example, the concept "problem" is connected to "billing" and "credit management". It can be very useful to see how one concept appears in multiple patterns, as "problem" does. Also notice that "pay" and "unwilling" are connected, a finding which should certainly be of interest to Astroserve. If you wished, you could create a category putting these negative comments about pay together.

You can move the icons around on the graph with the mouse to better see relationships, or to modify the graph for presentation. The graph can be copied to the clipboard with this button  to use in other applications.

Because the graphs are difficult to read when so many concepts are displayed, it is often useful to change the display to links of only a certain size.

3. Click the **Show slider**  button.

4. Move the left slider to 2.



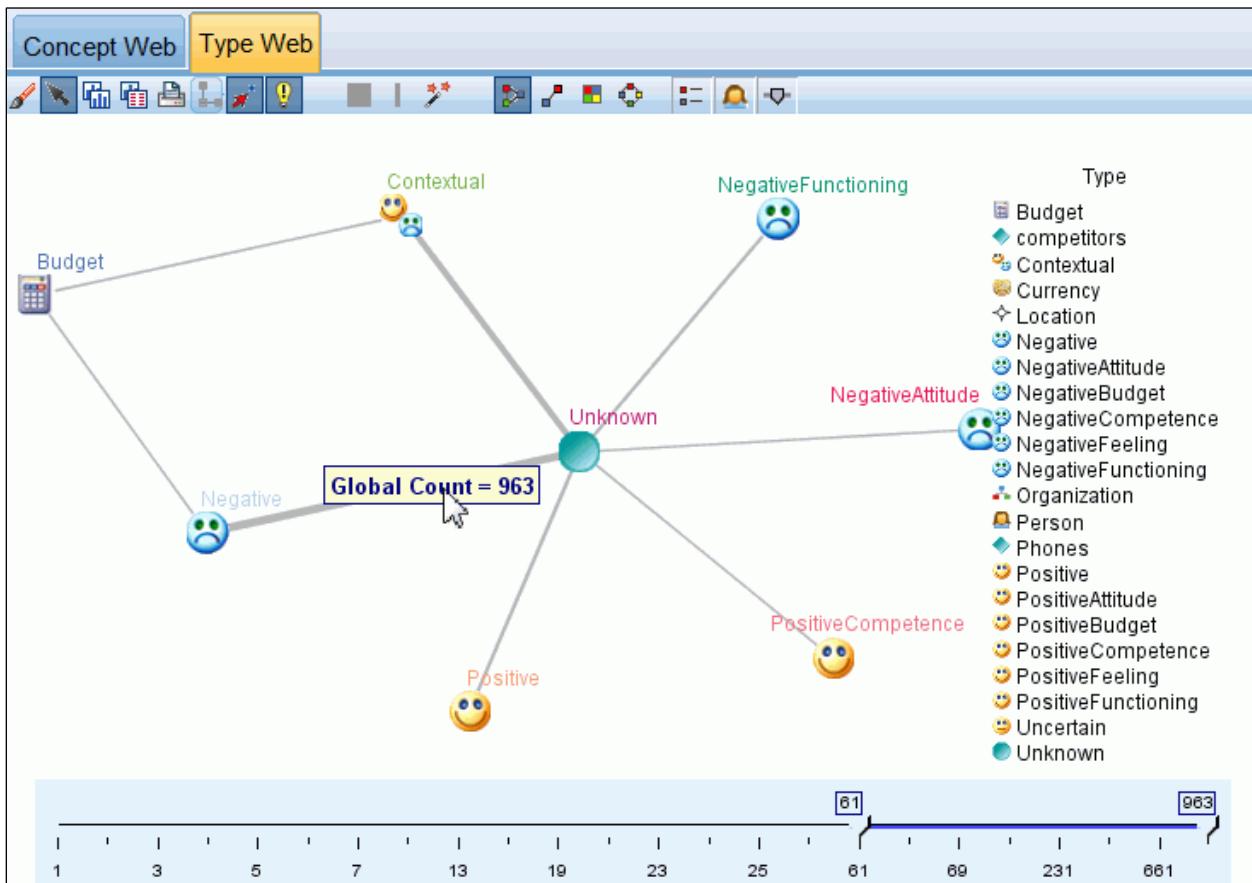
The graph is a little more readable now that you have eliminated links of less than two.

To see a Type Web graph, you need to select more than this one type pattern. A Type Web graph presents all the types in the selected pattern(s), but since you have only two types (Budget and Negative), all you will see is a line between two nodes. You will select all the type patterns.

5. Select the **first type pattern** in the Type Pattern pane, and then **Shift+click** the **last type pattern**.
6. Click the **Type Web** tab.

7. Click **Changes a web plot to a network layout** , and then click **Show slider**.
8. Move the slider to **61**.

The graph will only display the strongest connections.



The thickest lines, hence strongest connections, are between the "Unknown" type and the "Negative" type. The connection between the "Budget" type and the "Negative" type is not as strong, but at least you have identified at one class of problems that Astroserve customers are complaining about. Otherwise you still need to examine the concepts that were typed "Unknown" to find out what else they are complaining about.

9. Select the **<Unknown> + <Negative>** pattern in the **Types Pattern** pane.
10. Click **Concept 2**  in the **Concept Pattern** pane to alphabetize the concepts that were typed as Negative.

11. Scroll down until you see the term **noisy**.

Global	Docs	In	Concept 1	Concept 2	Concept 3
1	1		hearing	no interest	
1	1		service	noisy	
6	6		line	noisy	
1	1		rcf	noisy	
1	1		lines person	noisy	
1	1		support team	not able to assis	
1	1		windw	not able to close	
1	1		system	not able to dowr	
1	1		restoration	not acceptable	
1	1		response time f	not acceptable	
2	2		service	not acceptable	
1	1		information	not acceptable	
1	1		waiting	not acceptable	
1	1		turn around time	not acceptable	

It appears that nine customers complained of noisy lines, service, etc. You will take a close look at their actual complaints so that you can better understand what they are complaining about.

12. Select the first line that contains the term **noisy**, and then **Shift+click** the last line that has the term **noisy**.

Global	Docs	In	Concept 1	Concept 2	Concept 3
1	1		hearing	no interest	
1	1		service	noisy	
6	6		line	noisy	
1	1		rcf	noisy	
1	1		lines person	noisy	
1	1		support team	not able to assis	

13. Click **Display** to view their complaints in the Data pane.

	query (9) ▾	Categories
3	Lilian rang to advised that his line was very noisy claims that she rang earlier in the week but there are no records to show that she did, fault was lodge today for linesman to be fasttracked out to...	
4	The customer needs to discust her ongoing noisy line on her communic8 service . The customer advised she is losing money as cannot work from home with this service the customer needs to discuss compen...	
5	Cust upset that she has been given the run around for years. Cust has been complaining about her noisy line and has been told it was her equipment . Cust has picked up many phones and noise still on t...	

All three of these customers complained about noisy lines. This is certainly the type of complaint that would be of interest to Astroserve.

When there are so many patterns to view, and you want to simplify the results in the patterns panes, you can use the Filter choice from the contexts menu. Or you can open this dialog box using the Filter button from the toolbar. For example, you can look for which concepts the term "not working" is linked to in the call center data.

14. Click the **<Unknown> + <NegativeFunctioning>** pattern in the Type Pattern pane.
15. From the **Tools** menu, click **Filter**.
16. Beside the first **Match text** box, type **not working**, and then click **Filter**.

17. If no concepts are listed in the Concept Pattern pane, select <Unknown> + <NegativeFunctioning> pattern again in the Type Pattern.

Global	Docs	In	Concept 1	Concept 2
14	13		service	not working
10	9		line	not working
8	7		phone	not working
4	4		equipment	not working
2	2		telephone	not working
2	2		dsl	not working
2	2		cable	not working
1	1		system	not working
1	1		area	not working
1	1		email service	not working
1	1		amount of times	not working
1	1		nsy	not working
1	1		handset	not working
1	1		workmen on proper	not working
.

When you look at these results, remember that not all instances of not working are listed here. What is displayed instead are those patterns that contain this concept. The concept was not always included in a pattern (rarely would all instances of a concept be in at least one pattern), and so those records where it was not associated with another concept would not appear in the Text Link Analysis view.

18. From the **Tools** menu, click **Filter**.
 19. Beside **Match text**, delete **not working**, and then click **Filter**.
 20. From the **File** menu, click **Close**, and then click **Exit** to close the Interactive Workbench.
 21. From the **File** menu, click **Close Stream** without saving the changes.
 22. From the **File** menu, click **New Stream**.

Do not close Modeler. Leave it open for the next demo.

Result:

You now have identified concepts and types that are linked with positive and negative sentiment.

Using the Text Link Rules Editor

- Used to create new rules or edit existing ones.
- Can be accessed from:
 - the Text Link Rules tab in the Resource Editor within an interactive session
 - the Text Link Rules tab in the Template Editor, outside an interactive session

© 2014 IBM Corporation



You can edit and create rules directly in the Text Link Rules tab in the Template Editor or Resource Editor view. To help you see how rules might match text, you can run a simulation in this tab. During simulation, an extraction is run only on the sample simulation data and the text link rules are applied to see if any patterns match. Any rules that match the text are then shown in the simulation pane. Based on the matches, you can choose to edit rules and macros to change how the text is matched.

Unlike the other advanced resources, TLA rules are library-specific; therefore, you can only use the TLA rules from one library at a time. From within the Template Editor or Resource Editor, go to the Text Link Rules tab. In this tab, you can specify the library in your template that contains the TLA rules you want to use or edit. For this reason, it is strongly recommend that you store all your rules in one library unless there is a very specific reason not to do so.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

9-20

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Setting up Text Link Rule Values

Element	Quantity	Token
(mSupporting no)	Exactly 1	
	0 or 1	
mTopic	Exactly 1	
mToo	0 or 1	
NegativeAttitude	Exactly 1	

© 2014 IBM Corporation



This table contains the elements of the rule that are used for matching a rule to a sentence. You can add or remove rows in the table using the buttons to its right. The table consists of three columns:

- Element column: Enter values as one or a combination of types, literal strings, word gaps (<Any Token>), or macros. Double-click the element cell to enter the information directly. Alternatively, right-click in the cell to display a contextual menu offering lists of common macros, type names, and nonlinguistic type names. Keep in mind that if you enter the information into the cell by typing it in, precede the macro or type name with a '\$' character such as \$mTopic for the macro mTopic. The order in which you create your element rows is critical to how the rule will be matched to the text. When combining arguments, you must use parentheses () to group the arguments and the character | to indicate a Boolean OR. Keep in mind that values are case-sensitive.

- Quantity column: This indicates the minimum and maximum number of times the element must be found for a match to occur. For example, if you want to define a gap, or a series of words, between two other elements of anywhere from 0 to 3 words, you could choose "Between 0 and 3" from the list or enter the numbers directly into the dialog box. The default is "Exactly 1". In some cases you will want to make an element optional. If this is the case, then it will have a minimum quantity of 0 and a maximum quantity greater than 0 (such as 0 or 1, between 0 and 2). Note that the first element in a rule cannot be optional, meaning it cannot have a quantity of 0.
- Example Token column: If you click Get Tokens, the program breaks the Example text down into tokens and uses those tokens to fill this column with those that match the elements you defined. You can also see these tokens in the output table if you choose to.

Rule Output Table

Concept 1	Type 1	Concept 2	Type 2
(1) to	 (1)	(7)	 (7)

© 2014 IBM Corporation



Each row in this table defines how the TLA pattern output will appear in the results. Rule output can produce patterns of up to six Concept/Type column pairs, each representing a slot. For example, the type pattern <Location> + <Positive> is a two slot pattern meaning that it is made up of 2 Concept/Type column pairs.

To help you define the output quickly with fewer errors, you can use the context menu to choose the element you want to see in the output. Alternatively, you can also drag and drop elements from the Rule Value table into the output. For example, if you have a rule that contains a reference to the mTopic macro in row 3 of the Rule Value table, and you want that value to be in your output, you can simply drag/drop the element for mTopic to the first column pair in the Rule Output table. Doing so will automatically populate both the Concept and Type for the pair you've selected. Or if you want the output to begin with the type defined by the third element (row 3) of the rule value table, then drag that type from the Rule Value table to the Type 1 cell in the output table. The table will update to show the row reference in parenthesis (3).

Alternatively, you can enter these references manually into the table by double-clicking the cell in each Concept column you want to output and entering the \$ symbol followed by the row number, such as \$2 to refer to the element defined in row 2 of the Rule Value table. When you enter the information manually, you need to also define the Type column, enter the # symbol followed by the row number, such as #2 to refer to the element defined in row 2 of the Rule Value table.

Most rules have only one output row but there are times when more than one output is possible and desired. In this case, define one output per row in the Rule Output table.

When to Create or Edit Rules

- To capture an idea or relation that isn't being extracted with the existing rules
- To change the default behavior of a type you added to the resources
- To add new types to existing text link analysis rules and macros
- To add types to an existing text link analysis rule
- To slightly modify an existing rule, instead of creating a new one

© 2014 IBM Corporation



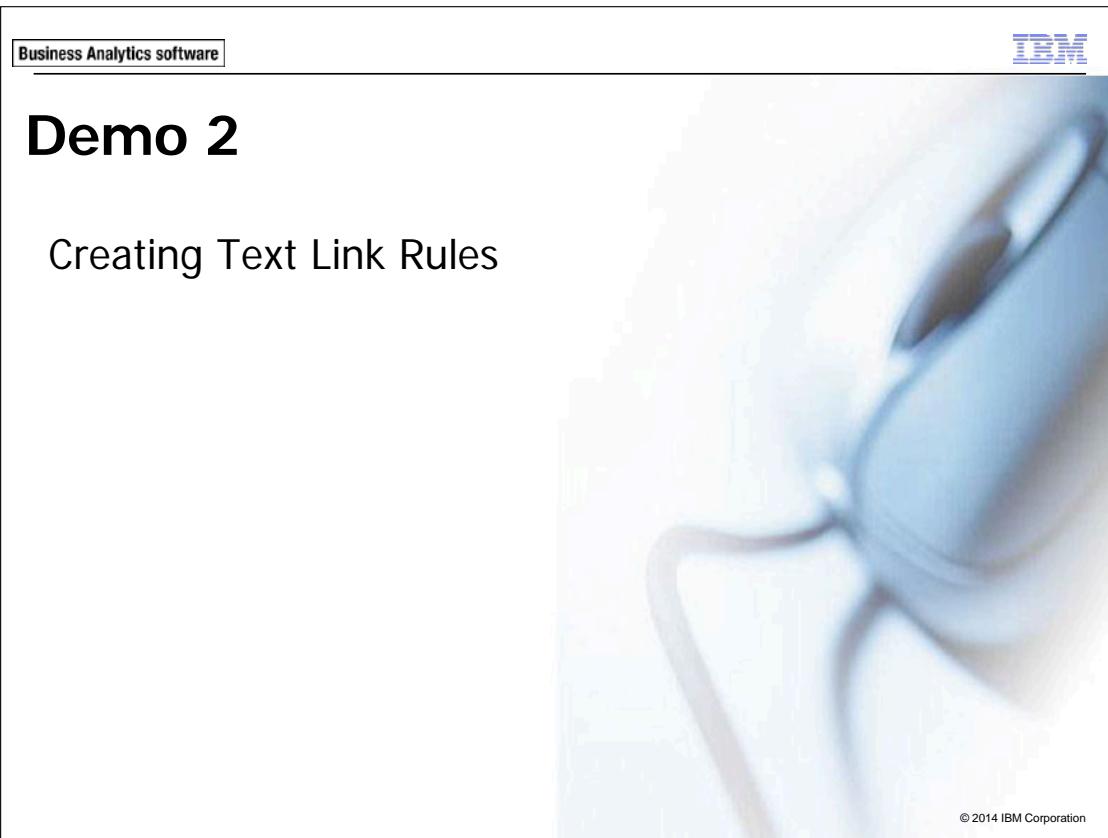
While the text link analysis rules delivered with each template are often adequate for extracting many simple or complex relationships from your text, there are times that you may want to make some changes to these rules or create some rules of your own. Examples include:

- To capture an idea or relation that is not being extracted with the existing rules by creating a new rule or macro.
- To change the default behavior of a type you added to the resources. This usually requires you to edit a macro such as mTopic or mNonLingEntities.
- To add new types to existing text link analysis rules and macros. For example, if you think the type <Organization> is too broad, you could create new types for organizations in several different business sectors such as <Pharmaceuticals>, <Car Manufacturing>, <Finance>, and so on. In this case, you must edit the text link analysis rules and/or create a macro to take these new types into account and process them accordingly.

- To add types to an existing text link analysis rule. For example, if you have a rule that captures the text john doe called jane doe but you want this rule that captures phone communications to also capture email exchanges. You could add the nonlinguistic entity type for email to the rule so it would also capture text such as: johndoe@ibm.com emailed janedoe@ibm.com.
- To slightly modify an existing rule, instead of creating a new one. For example, if you have a rule that matches the following text: xyz is very good but you want this rule to also capture: xyz is very, very good.

In order to help define new text link rules or help understand how certain sentences are matched during text link analysis, it is often useful to take a sample piece of text and run a simulation. During simulation, an extraction is run only on the sample simulation data using the current set of linguistic resources and the current extraction settings. The goal is to obtain the simulated results and use these results to improve your rules, create new ones, or better understand how matching occurs. For each piece of text (sentence, word, or clause depending on the context), a simulation output displays the collection of tokens and any TLA rules that uncovered a pattern in that text. A token is defined as any word or word phrase identified during the extraction process.

If you use a data file, it is strongly recommended that you ensure that the text it contains is short in order to minimize processing time. The goal of simulation is to see how a piece of text is interpreted and to understand how rules match this text. This information will help you write and edit your rules. Use the text link analysis node or run a stream with interactive session with TLA extraction enabled to obtain results for a more complete data set. This simulation is for testing and rule authoring purposes only.



The slide is titled "Demo 2: Creating Text Link Rules". It features the IBM logo in the top right corner and the text "Business Analytics software" in the top left corner. The background is a blurred image of a person's face. A small copyright notice "© 2014 IBM Corporation" is visible at the bottom right.

The following file(s) are used in this demo:

- Performing_Text_Link_Analysis_demo2_first step_start.str - a Modeler stream that reads a file containing call center data for March and April

Demo 2: Creating Text Link Rules

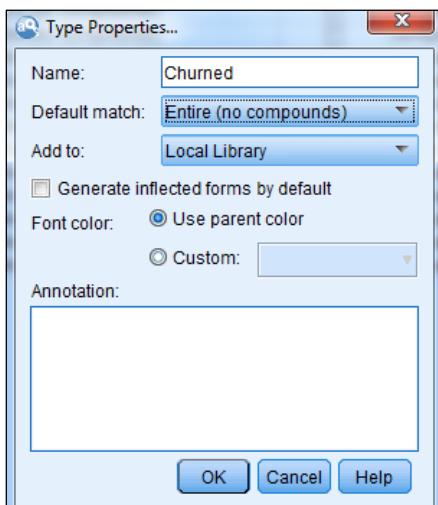
Purpose:

After examining the text data, you determined that the text link analysis is not picking up all the patterns you want it to. In particular, you noticed that some of the records indicate that the customer was either churning to a competitor or back to Astrocomm. In order to study those patterns, you need to add some Text Link Analysis rules that will capture them.

Task 1. Adding a type to use in the Text Link rules.

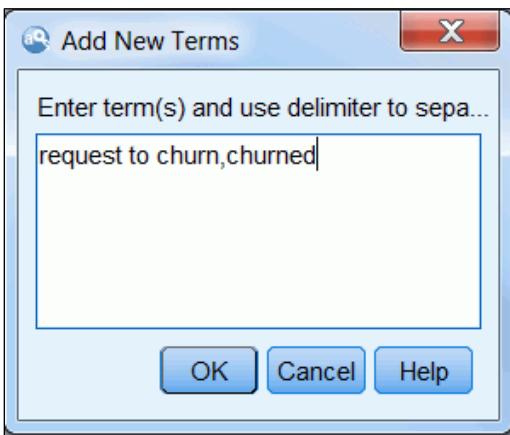
1. From the **File** menu, click **Open Stream**.
2. Navigate to **C:\Train\0A105\09-Performing_Text_Link_Analysis**, and then double-click **Performing Text Link Analysis_demo2_first_step_start.str**.
3. Run the Text Mining node named **query**.
Because, Text Link Analysis links types, you need to create a new type called "Churned".
4. Switch to the **Resource Editor**.
5. Click **Local Library** in the left pane.
6. From the **Tools** menu, click **New Type**.
7. Beside **Name**, type **Churned**.
8. Beside **Default match**, select **Entire (no compounds)**.

The results appear as follows:



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

9. Click **OK**.
10. From the **Tools** menu, click **New Terms**.
11. In the **Add New Terms** box, type **request to churn, churned**.



12. Click **OK**.
13. Switch to the **Categories and Concepts** view.
14. Click **Extract** in the **Extraction** pane, and then select **Type** from the list.
15. Select **<Churned>**, and then click **Display** to display all the customer responses that contained that type.

	query (9) ▾
1	Michael churned from optitel back to telsta in November 2002 on the promise he would be getting a total of 12.5% discount of all his calls this was to be made up of 2.5% from qtrs membership and 10% ...
2	Cust called ASTROSERVE about alleged advice received from a consultant in sales advising cust that if she churned optitel back to ASTROCOMM , she wld receive a \$50.00 credit off her first bill . This w...
3	Customer returned to churned back to ASTROCOMM (optitel) 02/10/2001. Contract form filled out correctly . Including pension concession number. Customer brought contract in ASTROCOMM office 20.03.2003....
4	Mr boboy has emailed through the following complaint about how his optitel service has been churned back to ASTROCOMM . For your reference the order number is 99777 and the kana case number is please ...

You will use this response as the type of pattern that you want to capture in the TLA rule.

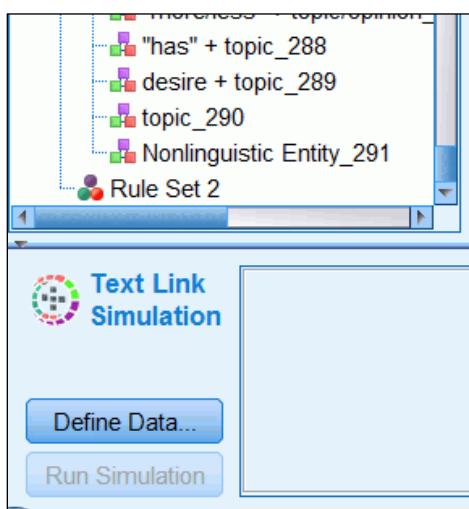
16. Right-click the first response, and then click **Copy**.
17. Switch to the **Resource Editor**.

Task 2. Creating a Text Link rule.

1. Click the **Text Link Rules** tab.
2. Scroll down until you get to the **Rules** section.

To make it easier to focus on the new rules you want to create, you will create a separate Rule set just for Churning. Although there is no technical reason for doing this, in TLA as soon as a rule fires, a same sequence cannot fire 2 rules, unless those rules are in different sets. It will also make it faster to tune the new Churning rules.

3. Right click **Rules** and select **Create Rule Set**.
4. Scroll down to view the new rule set (**Rule Set 2**) at the bottom of the window.

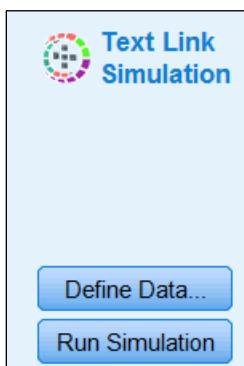


5. Minimize **Rule Set 1**, right-click **Rule Set 1**, and then click **Disable**.

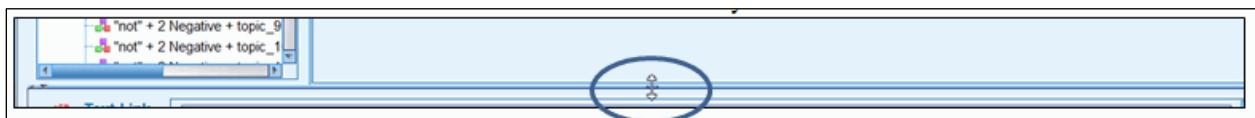
The results appear as follows:



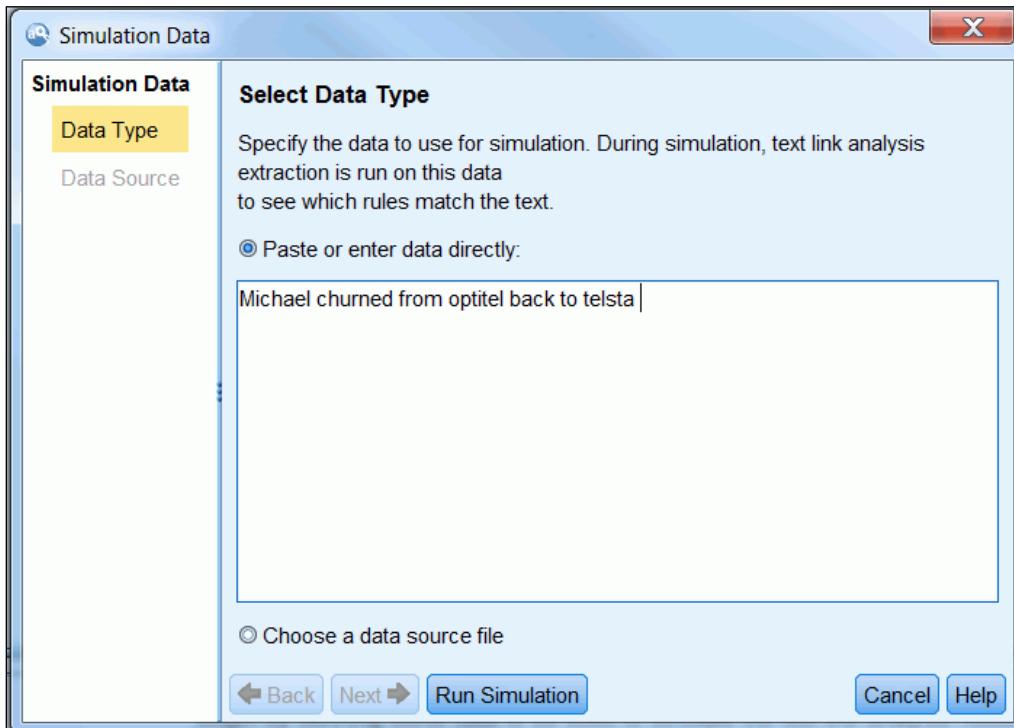
Now you are ready to create the text link rule. In the lower left corner you will see the Text Link Simulation portion of the dialog.



If you have trouble seeing this section of the dialog, pull up on the resize tool at the bottom of the blank window until you see it.



6. Click **Define Data**.
7. In the empty box, press **Ctrl-V** to paste the response into the box.
8. Remove everything after the word **telsta**.



9. Click Run Simulation.

Input Text Token	Typed As	Matching Macro
Michael	-	-
churned	Churned	-
from	-	mPrep

Define Data... Run Simulation Previous Previous Unmatched 1 of 1 results Next Unmatched Next Generate Rule

10. Click Generate Rule.

Again, you may have to resize the windows a bit in order to see this button.

Element	Quantity	Example Token
1 Churned	Exactly 1	
2 mPrep	Exactly 1	
3 competitors	Exactly 1	
4 -	0 or 1	
5 mEmpty	Exactly 1	

Name: Rule1
Example: Michael churned from optitel back to telsta
Rule Value table:
Rule Output table:

Add Remove View Source
Get Tokens Insert Row Remove Row

11. Beside **Name**, delete **Rule1**, and type **ChurnedFrom**.
12. Right-click the **Type 1** box and select **#1 --> \$Churned**.
13. Double-click the **Concept 1** box and type **\$1**.
14. Right -click the **Type 2** box and select **#3 --> \$competitors**.
15. Double-click the **Concept 2** box and type **\$3**.
16. Right -click the **Type 3** box and select **#6 --> \$competitors**.

17. Double-click the **Concept 3** box and type **\$6**.

	Element	Quantity	Example Token
1	Churned	Exactly 1	
2	mPrep	Exactly 1	
3	competitors	Exactly 1	
4		0 or 1	
5	mEmpty	Exactly 1	
6	competitors	Exactly 1	

Concept 1	Type 1	Concept 2	Type 2	Concept 3	Type 3
(1)	(1)	(3)	(3)	(6)	(6)

18. Click **View Source** in the upper right to see the code created by the dialog box. The results appear as follows:

```
#@# Michael churned from optitel back to telsta
[pattern(1)]
name=ChurnedFrom
value=$Churned $mPrep $competitors @{0,1} $mEmpty $competitors
output(1)=$1#1\$3#3\$6#6
```

19. Click **Exit Source** to return back to the rule editor.

Task 3. Applying the Text Link rule to the data.

1. Switch to the **Text Link Analysis** view.
2. Click **Extract**, in the top left pane.
3. Click the <Churned> + <competitors> + <competitors> pattern in the **Text Analysis** pattern pane.

Global	In	Type 1	Type 2	Type 3
1		<Churned>	<competitors>	<competitors>

- Click the pattern in the **Concept** pane, and then click **Display**.

The response you used in the simulation is displayed in the data pane on the lower right side. In the upper left corner, you can verify that you captured that the customer switched from "optitel" to "telsta" (<Churned> + <competitors> + <competitors>).

- Switch to the **Resource Editor**.
- Drag the **ChurnedFrom** to **Rule Set 2** (if necessary).

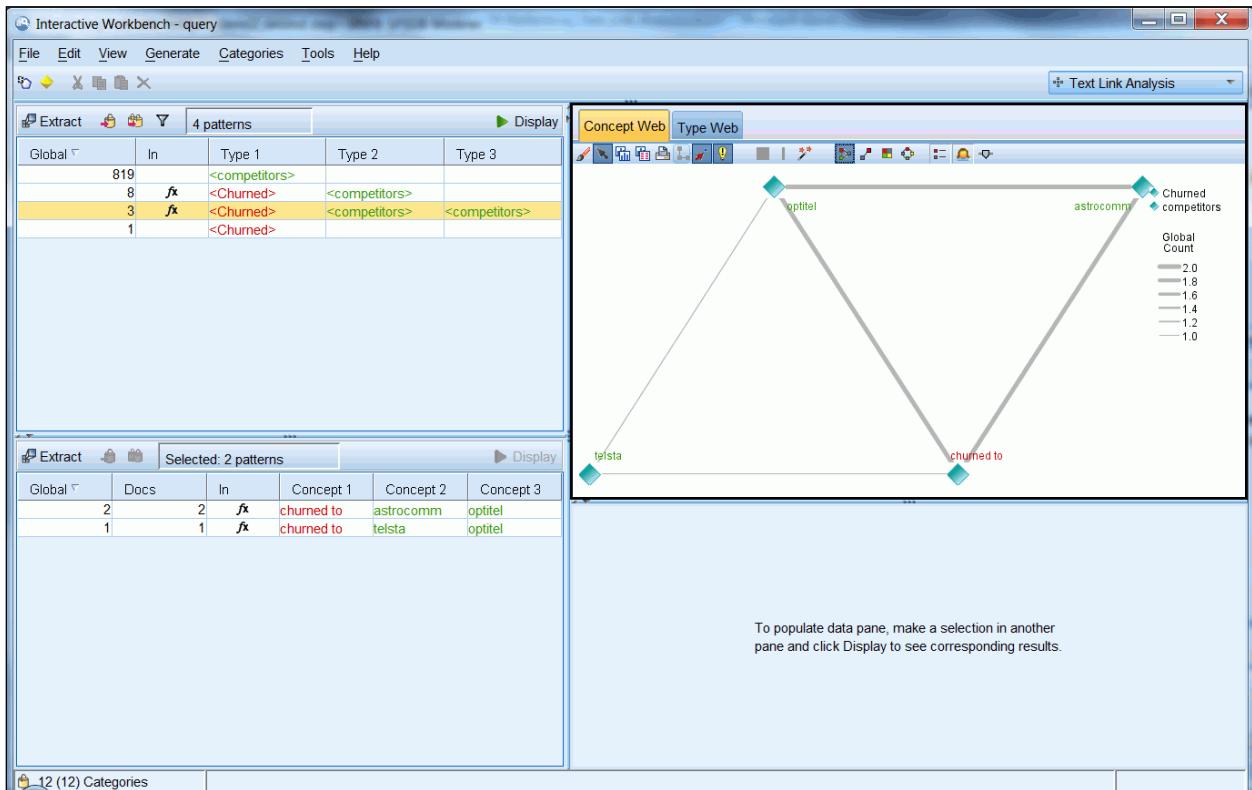
Task 4. Saving the Text Analysis rule.

- From the **File** menu, click **Update Modeling Node**.
- Click **OK**, and then click **OK** to the message that the modeling node has been updated.
- From the **File** menu, click **Close**, and then click **Exit** to end the Interactive Session.
- From the **File** menu, click **Save Stream As**.
- Name the stream **Performing Text Link Analysis_demo2_first step_end.str**, and then click **Save**.
- From the **File** menu, click **Close Stream**.
- From the **File** menu, click **New Stream**.

At this point you will switch to a different stream that contains this rule and some additional rules after that have been fine tuned. For example you added a rule that checks for whether a customer Churned To a competitor as opposed to Churned From a competitors as you did in the previous example.

- From the **File** menu, click **Open Stream**.

9. Navigate to **C:\Train\0A105\09-Performing_Text_Link_Analysis**, and then double-click **Performing Text Link Analysis_demo2_second step.str**.
10. Run the Text Mining node named **query**.
11. Click the pattern **<Churned> + <competitors> + <competitors>**.



There were customers who churned to Optitel as you observed before, and there are also customers who churned to Telsta and Astrocomm. You will examine the same rule you were working with after it has been modified.

12. Switch to the **Resource Editor**.
13. Click the **Text Link Rules** tab.

14. Click Rule Set 0 > rule1_churnedFromTo.

The screenshot shows the configuration of a rule named "rule1_churnedFromTo". The rule example is "Michael churned from optitel back to telsta". The Rule Value table lists tokens and their quantities: "Churned" (Exactly 1), "(back | over | away)" (0 or 1), "from" (0 or 1), "mDet" (Between 0 and 2), and "competitors" (Exactly 1). The Rule Output table shows the mapping of these tokens to concepts: "(1) to" maps to Type 1 (1), Concept 2 (10), Type 2 (10), Concept 3 (5), and Type 3 (5). Buttons for "Get Tokens", "Insert Row", and "Remove Row" are visible on the right. At the bottom, there are options to "Show output as" (References to row in Rule Value table or Specific token from example) and buttons for "Apply" and "Cancel".

Now, not only do you capture whether the customer Churned to a competitor but also if they Churned from a competitor.

15. From the **File** menu, click **Close**, and then click **Exit** to exit from the Text Link Rules window.
16. From the **File** menu, click **Close Stream**.
17. From the **File** menu, click **New Stream**.

Do not close Modeler. Leave it open for the next demo.

Result:

You have successfully created some custom rules that are specific to the types of calls the Astroserve call center receives from its customers.

Converting TLA Patterns to Categories

- Once you have identified patterns in the text data, you can convert the patterns into new or existing categories.
- These categories can later be used to build reports or they can be used as predictors in a model.

© 2014 IBM Corporation



While text link analysis is extremely useful for identifying patterns in the text data, until you convert the patterns into categories, you cannot make use of the patterns as variables outside the Interactive Workbench to perform further analysis. For example, you may want to create reports based on these patterns, or use them as predictors in a predictive model. In the case of the Astroserve customers, it would be desirable to use the two patterns you identified, <Budget> + <Negative> and Not Working + <NegativeFunctioning> as predictors in a model aimed at predicting which customers are likely to churn. It would also be of interest to discover if men or women were more likely to complain about these sorts of issues, or whether it depended on age group, income group or something else.

It is important to rely on your business knowledge to identify which patterns you want to convert into categories. Because Text Analytics extracts so much information, you need to rely on your business expertise to identify which types, concepts or patterns to zero in on. For example, the sorts of issues a Telco company such as Astroserve, would probably want to focus on would involve things like mobile phones, service, coverage, and billing issues.

Business Analytics software

IBM

Demo 3

Converting TLA Patterns into Categories



© 2014 IBM Corporation

The following file(s) are used in this demo:

- Performing Text Link Analysis_demo3_start.str - a Modeler stream that reads a file containing call center data for March and April

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

9-38

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Demo 3: Converting TLA Patterns into Categories

Purpose:

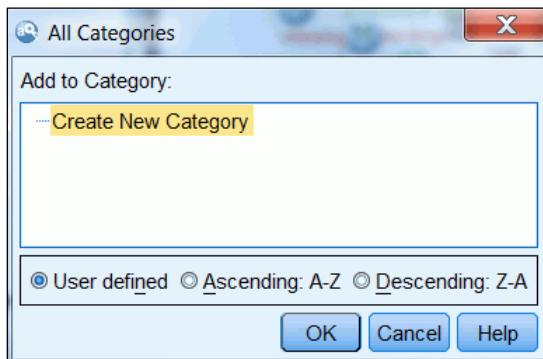
You have identified some key patterns that you believe may be useful for predicting which Astroserve customers are likely to churn. In order to use these patterns as predictors in model, you must convert them into categories.

Task 1. Converting <Budget> + <Negative> into a category.

1. From the **File** menu, click **Open Stream**.
2. Navigate to **C:\Train\0A105\09-Performing_Text_Link_Analysis**, and then double-click **Performing Text Link Analysis_demo3_start.str**.
3. Run the **Text Mining** node named **query**.
4. Select the <Budget> + <Negative> pattern in the Type Patterns pane.
5. Right-click the <Budget> + <Negative> pattern, and then click **Add to Category**.

The Add to Category selection can be used to add the pattern to an existing category or a new one. There are no categories currently in the Interactive Workbench, so there is only a choice to create a new category.

6. Click **Create a New Category**.



7. Click OK.



The screenshot shows two panes of the IBM SPSS Text Analytics interface. The top pane, titled '78 patterns', displays a list of patterns categorized into Type 1, Type 2, and Type 3. The bottom pane, titled 'Selected: 45 patterns', shows a list of terms with their associated documents (Docs) and concepts (Concept 1, Concept 2, Concept 3). Both panes include a 'Display' button at the top right.

Global ▾		In	Type 1	Type 2	Type 3
	149		<NegativeFunctioni		
	112		<Budget>	<Contextual>	
	103		<PositiveCompete		
	69	⌚	<Budget>	<Negative>	
	68		<Unknown>	<NegativeAttitude>	
	61		<Unknown>	<PositiveCompete	
	44		<competitors>		
	41		<NegativeCompete		
	38		<email>		
	38		<Uncertain>		
	34		<Organization>		
	30		<Location>		
	30		<Unknown>	<PositiveAttitude>	
	29		<Currency>		

Global ▾		Docs	In	Concept 1	Concept 2	Concept 3
	6	5	⌚	paid	would be good	
	5	5	⌚	pay	would be good	
	5	5	⌚	paying	not happy	
	4	4	⌚	charge	disappointing	
	3	2	⌚	payment	missing	
	2	2	⌚	cost	difficult	
	2	2	⌚	pay	unwilling	
	2	2	⌚	charge	problem	
	2	2	⌚	paid	too long	
	2	2	⌚	charge	canceling	
	2	2	⌚	credit manager	complaint	
	1	1	⌚	payment arran	hazardous	
	1	1	⌚	refund of rental	fault	
	1	1	⌚	mobile rates	not happy	

The  symbol in the In column next to the <Budget> + <Negative> pattern indicates that the pattern has been converted to a category. The same symbol is used in In column of the Concept Pattern pane next to each term that has been used in a category.

Task 2. Creating a new category called "Noisy Phone".

1. Click <Unknown> + <Negative> in the Type Patterns pane.
2. Click the **Concept 2** column header until these results are in ascending alphabetical order.
3. Scroll down until the concept **noisy** is visible.

Global	Docs	In	Concept 1	Concept 2	Concept 3
1	1	1	sentiment	negative	
1	1		premises	no interest	
1	1		hearing	no interest	
1	1		service	noisy	
6	6		line	noisy	
1	1		rcf	noisy	
1	1		lines person	noisy	
1	1		support team	not able to assi:	
1	1		windw	not able to close	
1	1		system	not able to down	
1	1		restoration	not acceptable	
1	1		response time	not acceptable	
2	2		service	not acceptable	
1	1		information	not acceptable	
1	1		waiting	not acceptable	

There are 4 concept patterns that include the concept noisy. None occur very frequently, although line is associated with noisy in 6 records. All of them clearly indicate a problem with a noisy phone line.

4. **Shift+click** all the patterns with **noisy**.
5. Right-click anywhere in the highlighted area, and then click **Add to Category**.

6. Select **Create New Category**, and then click **OK**.

Global	Docs	In	Concept 1	Concept 2	Concept 3
1	1		sentiment	negative	
1	1		premises	no interest	
1	1		hearing	no interest	
1	1	bag	service	noisy	
6	6	bag	line	noisy	
1	1	bag	rcf	noisy	
1	1	bag	lines person	noisy	
1	1		support team	not able to assi:	
1	1		windw	not able to close	
1	1		system	not able to down	
1	1		restoration	not acceptable	
1	1		response time	not acceptable	
2	2		service	not acceptable	
1	1		information	not acceptable	
1	1		waiting	not acceptable	

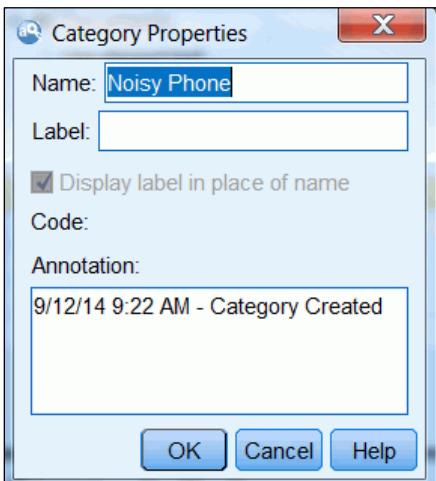
Normally, you would continue in this manner, looking at the many TLA patterns that were discovered and using them to understand the comments in more detail, or to create potentially useful categories. The review of TLA in the Interactive Workbench will conclude at this point. You should view the categories just created and update the modeling node.

7. Switch to the **Categories and Concepts** view.
 8. Click **Score** to see how many records there are in each category.

Category	Descriptors	Docs
All Documents	-	1269
Uncategorized	-	1150
No concepts extracted	-	2
<Budget> + <Negative>	1	110
<> <Budget> + <Negative>		110
New Category	4	9
• line + noisy		6
• lines person + noisy		1
• rcf + noisy		1
• service + noisy		1

There are two categories corresponding to the two patterns. The name of the category that includes budget and negative type concepts is descriptive enough, but the one containing nine instances of noisy is not, so you will change its name.

9. Right-click **New Category**, and then click **Rename Category**.
10. Beside **Name**, type **Noisy Phone**.



11. Click **OK**.

Task 3. Saving the new categories in the Text Mining node.

Now you need to update the modeling node so that you do not lose any of your changes.

1. From the **File** menu, click **Update Modeling Node**.
2. Click **OK** in the informational dialog, then click **OK** again.
3. From the **File** menu, click **Close**, and then click **Exit** to close the Interactive Workbench.
4. From the **File** menu, click **Save Stream As**.
5. Name the stream **Performing Text Link Analysis_demo3_end.str**, and then click **Save**.
6. From the **File** menu, click **Close Stream**.
7. From the **File** menu, click **New Stream**.

Do not close Modeler. Leave it open for the next demo.

Result:

You have successfully converted some Text Link Analysis patterns into categories that you can later use to predict churn.

Text Link Analysis Node

- Used to automatically find concepts patterns and output the information as new records or fields.
- Unlike the Interactive Workbench:
 - cannot modify settings interactively
 - cannot view patterns with visualization
- Saving the concept patterns as fields allows you to analyzed the relationship between these concept patterns and other variables in your file.

© 2014 IBM Corporation



The Text Link Analysis node can be used to automatically find concept patterns and output this information as new records and fields. In essence, the patterns become new records in the data, with fields representing types and concepts.

The principles are the same as in doing TLA in the Interactive Workbench, but the Text Link Analysis node does not provide any output directly. Nor does it enable you to modify the settings interactively or view the concept patterns with visualization. You can explore the concept patterns with the output from this node, but only by using other nodes in Modeler.

ID	Concept1	Type1	Concept2	Type2	Concept3	Type3	Concept4	Type4
1	cust	Unknown	upset	Negative	Null	Null	Null	Null
2	phone	Unknown	Null	Null	Null	Null	Null	Null
3	charge	Budget	Null	Null	Null	Null	Null	Null
4	outgoing mobile calls	Unknown	Null	Null	Null	Null	Null	Null
5	phone service	Unknown	Null	Null	Null	Null	Null	Null
6	wiring	Unknown	deteriorated	NegativeFunctioning	Null	Null	Null	Null
7	box	Unknown	deteriorated	NegativeFunctioning	Null	Null	Null	Null
8	wall	Unknown	Null	Null	Null	Null	Null	Null
9	cust	Unknown	Null	Null	Null	Null	Null	Null
10	wiring	Unknown	long	Negative	Null	Null	Null	Null

© 2014 IBM Corporation 

In the output from this node, each record will represent a concept pattern. This node creates 15 fields, including six fields for concepts found in the pattern, six fields representing the type of each concept, an ID field using the name of the ID field specified in the node, a Rule Name field, and a Matched Text field that represents the portion of the text data in the original record or document that was matched to this pattern.

The same results obtained with the Interactive Workbench will not be found because the modified linguistic resources are not being used. Although you saved those in the query Text Mining node, you did not save the resources in a new resource template, which you could load into this node. Normally, you would do so in a text mining project, but for this example the point was to demonstrate the capabilities of the Text Link Analysis node, so using the same resources was not necessary.

Business Analytics software

IBM

Text Link Analysis Node Output (cont'd)

Preview from Text Link Analysis Node (15 fields, 10 records)

	bt6	Type6	Rule Name	Query_ID	Matched Text
1	Null	1/topic + opinion	179	386504	<*Cust*> is <*upset*> that her phone was out for a week
2	Null	1/topic	290	386504	Cust is upset that her <*phone*> was out for a week and
3	Null	1/topic	290	386504	Cust is upset that her phone was out for a week and she
4	Null	1/topic	290	386504	Cust is upset that her phone was out for a week and she
5	Null	1/topic	290	386504	Cust is upset that her phone was out for a week and she
6	Null	1/opinion + 2 topics	190	386521	Lightening <*damaged*> <*wiring*> & <*box*> on wall 3
7	Null	1/opinion + 2 topics	190	386521	Lightening <*damaged*> <*wiring*> & <*box*> on wall 3
8	Null	1/topic	290	386521	Lightening damaged wiring & box on <*wall*> 3 weeks ac
9	Null	1/topic	290	386521	<*Cust*> is concerned that is taking so long to have wirin
10	Null	1/opinion + topic	235	386521	Cust is concerned that is taking so <*long*> to have <*wi

© 2014 IBM Corporation

Although the first record's ID appears in the first five records, a careful examination shows that there are not five concept patterns for this customer. The node also output single slot concept patterns, which are those concepts that were found in the data without being included in a pattern.

The sole pattern for ID 386504 is cust + upset. Notice how the concept "cust" was extracted even though earlier you excluded from extraction. This is because the Text Mining node does not use changes made to the linguistic resources in the Interactive Workbench where this term was excluded).

The precise words that are included in this pattern are listed in brackets in the Matched Text field. This field does not always include the full text from a record but just the relevant portion. Most of the fields representing a concept or type have a value of Null because there are no patterns to represent, just single slot concepts.

Text Link Analysis Node - Tips

- Cache on the Text Link Analysis node. Otherwise each time you run the node all the TLA patterns will have to be recreated.
- Insert a Type node to change the new concept and type fields from Typeless to Nominal. Otherwise, the fields will not be listed in variable selection lists.

© 2014 IBM Corporation



At this point, the output from the Text Link Analysis node can be used in a variety of analyses. Before proceeding, you would probably want to enable a cache on the Text Link Analysis node. Otherwise, when you run nodes using output from the Text Link Analysis node, the complete TLA patterns will need to be re-created each time the stream is run, which in large data files can increase processing time significantly.

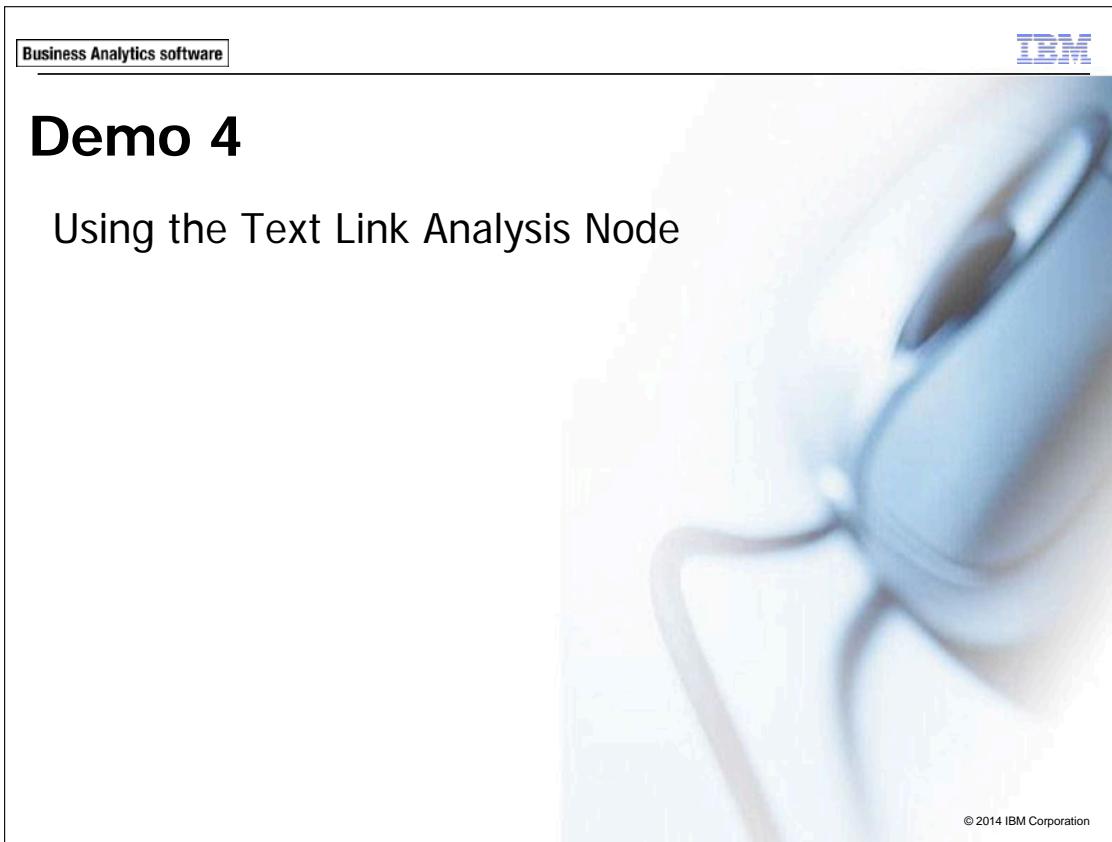
The new concept and type fields created for each slot in a pattern are Typeless by default. This means that many nodes, such as those in the Graphs palette, will not list these fields in variable selection lists. To use these fields, you will have to add a Type node to the stream and manually change the type to Nominal for these 12 fields.

The new fields can be used in Distribution or Web graph nodes to examine the relationships in more detail. You can match the output from the Text Link Analysis node back to the original data file so that the original records have all the patterns. To do this you will have to restructure these fields, perhaps using the History or Restructure nodes, so that there is one record that contains all the patterns. (First you could remove all the records that contain only single slot patterns). Or you could match the original data file to this file. This would allow you to see how the patterns are related to other fields, such as customer demographics.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.



The slide features a large, faint background image of a person wearing a white lab coat and a stethoscope around their neck. In the top left corner, there is a small rectangular box containing the text "Business Analytics software". In the top right corner, the "IBM" logo is displayed. At the bottom right of the slide, the copyright notice "© 2014 IBM Corporation" is visible.

Demo 4

Using the Text Link Analysis Node

The following file(s) are used in this demo:

- Performing Text Link Analysis_demo4_start.str - a Modeler stream that reads a file containing call center data for March and April

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

9-48

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Demo 4: Using the Text Mining Node

Purpose:

You would like to analyze the relationship between text link patterns and other variables in your file. Because this is not possible in the Interactive Workbench, you need to use the Text Link Analysis node to convert the patterns to fields.

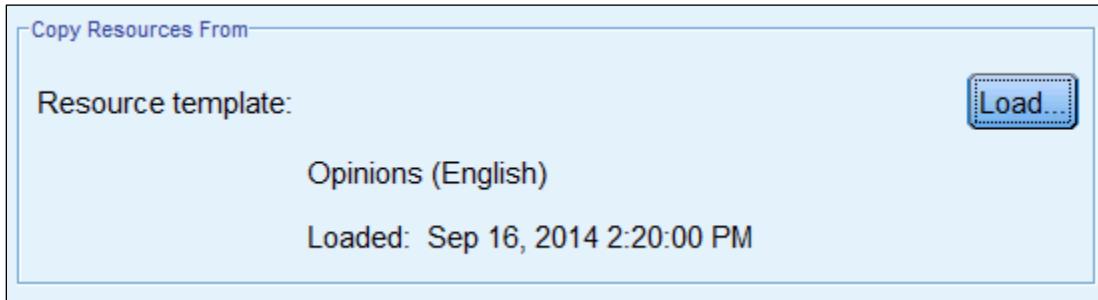
Task 1. Running the Text Link Analysis node.

1. From the **File** menu, click **Open Stream**.
2. Navigate to **C:\Train\0A105\09-Performing_Text_Link_Analysis**, and then double-click **Performing Text Link Analysis_demo4_start.str**.
3. Add a **Text Link Analysis** node from the **IBM SPSS Text Analytics** tab to the stream.
4. Connect the **Text Link Analysis** node to the **25% Random Sample** node.
5. Edit the **Text Link Analysis** node.
6. Beside **Text field**, select **query**, and beside **ID field**, select **Query_ID**.



An ID field is required because of the restructuring that will be done to the data. Otherwise, the Fields tab dialog is identical to that in the Text Mining modeling node.

7. Click **Load**, select the **Opinions (English)** template, and then click **OK**.



8. Click the **Expert** tab.
9. Click the **Accommodate spelling for a minimum root character limit of** check box and change the value to 4.

10. Click **OK**.
11. From the **Output** tab, add a **Table** node to the stream and connect it to the **Text Link Analysis** node.
12. Run the **Table** node.

Table (15 fields, 19,807 records)

	Concept1	Type1	Concept2	Type2	Concept3	Type3	Concept4	Type4
1	cust	Unknown	upset	Negative	Null	Null	Null	Null
2	phone	Unknown	Null	Null	Null	Null	Null	Null
3	charge	Budget	Null	Null	Null	Null	Null	Null
4	outgoing mobile calls	Unknown	Null	Null	Null	Null	Null	Null
5	phone service	Unknown	Null	Null	Null	Null	Null	Null
6	wiring	Unknown	deteriorated	NegativeFunctioning	Null	Null	Null	Null
7	box	Unknown	deteriorated	NegativeFunctioning	Null	Null	Null	Null
8	wall	Unknown	Null	Null	Null	Null	Null	Null
9	3 weeks	Period	Null	Null	Null	Null	Null	Null
10	cust	Unknown	Null	Null	Null	Null	Null	Null
11	wiring	Unknown	long	Negative	Null	Null	Null	Null
12	resolved	Positive...	Null	Null	Null	Null	Null	Null
13	dial tone	Unknown	Null	Null	Null	Null	Null	Null
14	service rep	Unknown	Null	Null	Null	Null	Null	Null
15	cust	Unknown	Null	Null	Null	Null	Null	Null
16	not definite	Positive	Null	Null	Null	Null	Null	Null
17	resolved	Positive...	Null	Null	Null	Null	Null	Null
18	service	Unknown	not satisfied	Negative	Null	Null	Null	Null
19	astrocomm	Unknown	Null	Null	Null	Null	Null	Null
20	exchange	Unknown	Null	Null	Null	Null	Null	Null

OK

The sample of Astroserve data contains 1,269 records. The stream now contains 19,807 records.

13. Scroll to the right until you see the **Query ID** column.

	Rule Name	Query_ID	Matched Text
1	1/topic + opinion_179	386504	<*Cust*> is <*upset*> that her phone was out for a week and she is being
2	1/topic_290	386504	Cust is upset that her <*phone*> was out for a week and she is being
3	1/topic_290	386504	Cust is upset that her phone was out for a week and she is being <*c
4	1/topic_290	386504	Cust is upset that her phone was out for a week and she is being cha
5	1/topic_290	386504	Cust is upset that her phone was out for a week and she is being cha
6	1/opinion + 2 topics_190	386521	Lightening <*damaged*> <*wiring*> & <*box*> on wall 3 weeks ago
7	1/opinion + 2 topics_190	386521	Lightening <*damaged*> <*wiring*> & <*box*> on wall 3 weeks ago
8	1/topic_290	386521	Lightening damaged wiring & box on <*wall*> 3 weeks ago
9	1/Nonlinguistic Entity_291	386521	Lightening damaged wiring & box on wall <*3 weeks*> ago
10	1/topic_290	386521	<*Cust*> is concerned that is taking so long to have wiring completem
11	1/opinion + topic_235	386521	Cust is concerned that is taking so long to have <*wiring*> comp
12	1/opinion_286	386521	Cust is concerned that is taking so long to have wiring completely <*
13	1/topic_290	386521	<*Dial tone*> has been restored by service rep
14	1/topic_290	386521	Dial tone has been restored by <*service rep*>
15	1/topic_290	386521	<*Cust*> concerned that is not definate will be fixed on 3/3 as previou
16	1/not" + Negative_258	386521	Cust concerned that is not <*definante*> will be fixed on 3/3 as previous
17	1/opinion_286	386521	Cust concerned that is not definate will be <*fixed*> on 3/3 as previous
18	1/opinion + topic_235	386526	Simon is <*unhappy*> with the <*service*> received from ASTROCO
19	1/topic/opinion + "comm...	386526	Simon is unhappy with the service received from <*ASTROCOMM*> <
20	1/topic_290	386526	Simon is unhappy with the service received from ASTROCOMM , there

Customer 386504 appears on the first five records. In the first record, the terms "Cust" and "upset" were parts of the pattern, and in the second record "phone" was contained in the pattern. Remember that although you excluded the term "Cust" from extraction, the Text Link Analysis node does not use changes made in the Interactive Workbench.

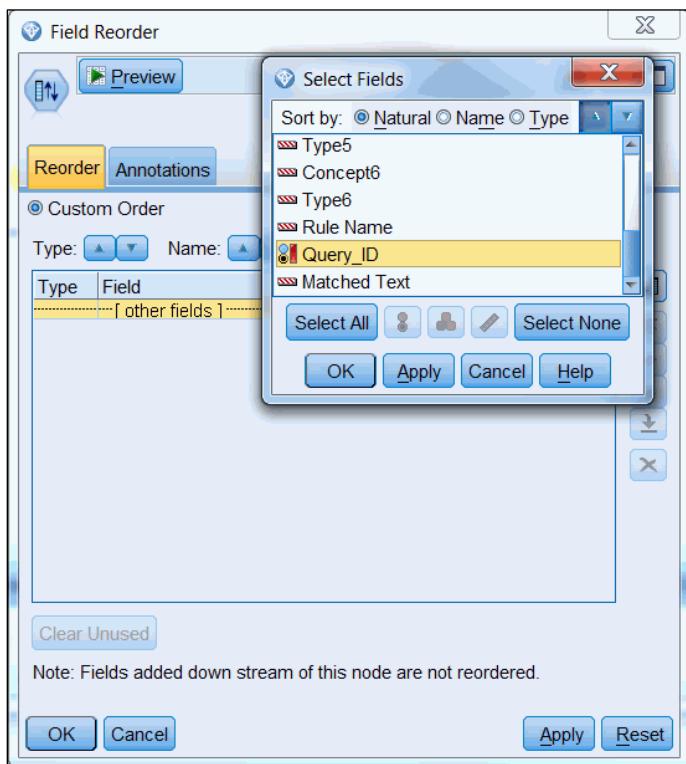
14. Close the **Table** output window

Task 2. Using the results in subsequent analyses.

In order to use the new fields in nodes such as Distribution, Matrix or Web graph to examine them in more detail, you must change the types for the new concept and type fields from Typeless to Nominal. Otherwise, these fields will not be listed in the variable selection lists in these nodes.

Before you retype the fields, you will cache the data at the Text Link Analysis node and also reorder the fields in the output so that Query ID is the first variable on the output. This will make it easier to locate a particular customer's record in the Table output.

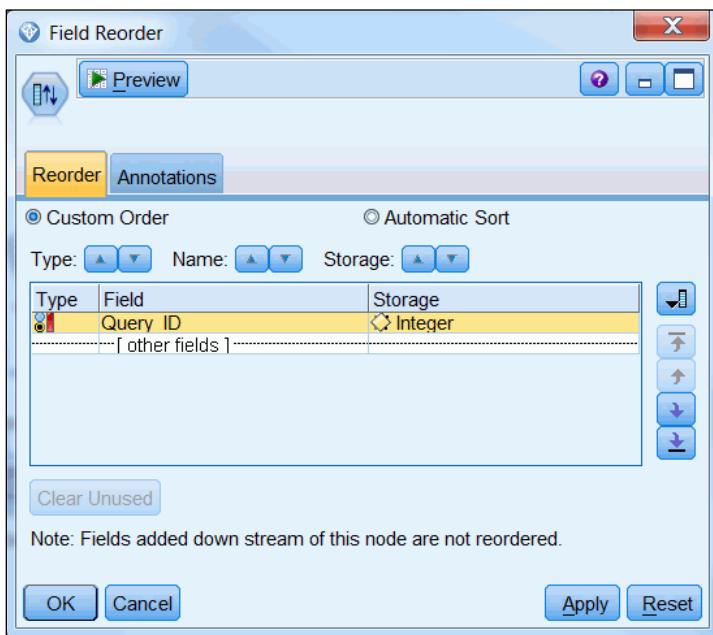
1. Right-click the **Text Link Analysis** node, point to **Cache**, and then click **Enable**.
2. Run the **Table** node to cache the data.
3. From the **Field Ops** tab, add a **Field Reorder** node to the stream.
4. Connect the **Field Reorder** node to **Text Link Analysis** node.
5. Edit the **Field Reorder** node.
6. Select **[other fields]**.
7. From the **Select Fields** list, select **Query ID**.



8. Click **OK** to close the **Select Fields** window.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

9. Select **Query ID**, and then click **Top**  to move the variable to the top of the list.



10. Click **Preview**.

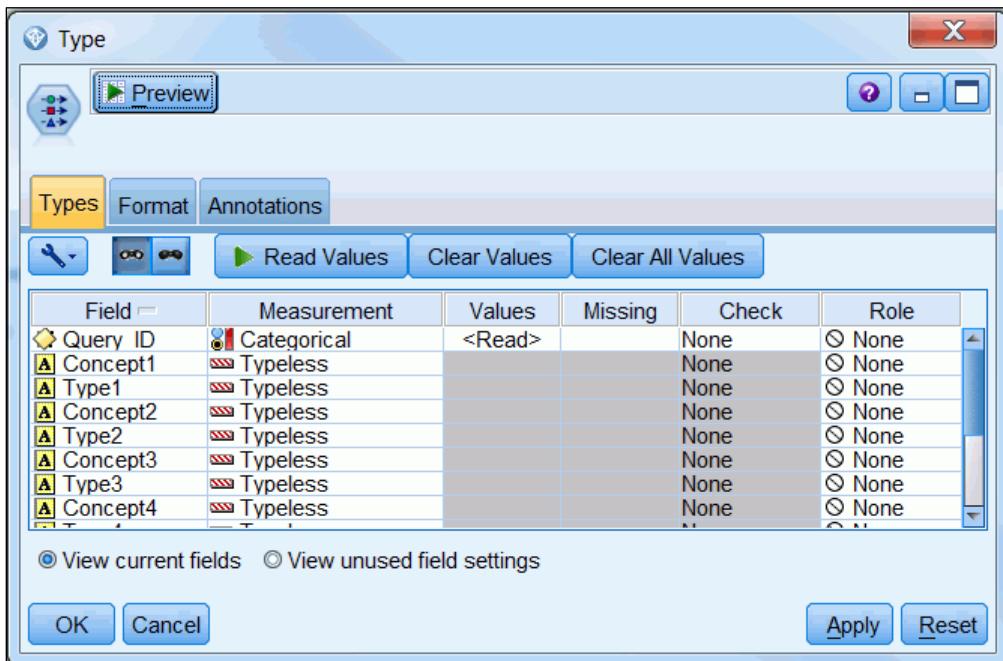
Preview from Field Reorder Node (15 fields, 10 records)									
	Query_ID	Concept1	Type1	Concept2	Type2	Concept3	Type3	Concept4	
1	386504	cust	Unknown	upset	Negative	Null	Null	Null	
2	386504	phone	Unknown	Null	Null	Null	Null	Null	
3	386504	charge	Budget	Null	Null	Null	Null	Null	
4	386504	outgoing mobile calls	Unknown	Null	Null	Null	Null	Null	
5	386504	phone service	Unknown	Null	Null	Null	Null	Null	
6	386521	wiring	Unknown	deteriorated	NegativeFunctioning	Null	Null	Null	
7	386521	box	Unknown	deteriorated	NegativeFunctioning	Null	Null	Null	
8	386521	wall	Unknown	Null	Null	Null	Null	Null	
9	386521	cust	Unknown	Null	Null	Null	Null	Null	
10	386521	wiring	Unknown	long	Negative	Null	Null	Null	

The Query ID field is now displayed on the far left.

11. Close the **Preview** window.
12. Click **OK**.

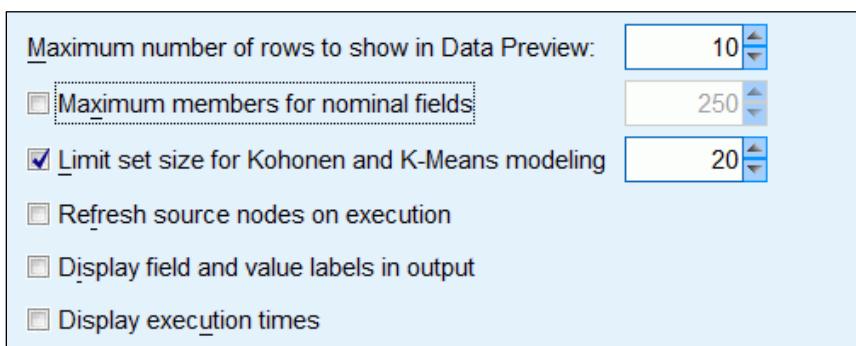
You will retype the Concept and Type fields.

13. From the **Field Ops** tab, add a **Type** node to the stream.
14. Connect the **Type** node to **Field Reorder** node, and then edit the **Type** node.



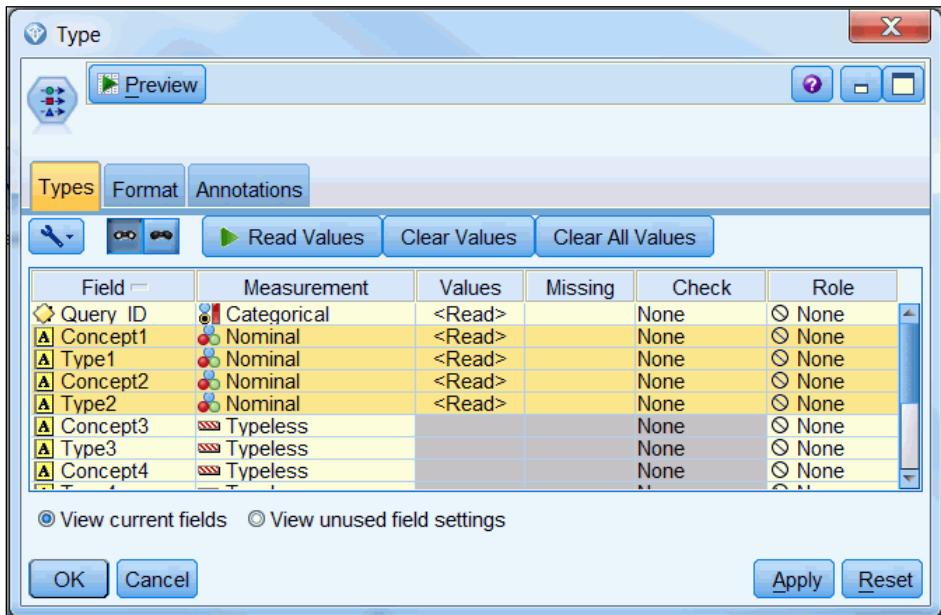
In this window you see that the Measurement value for all the new Concept and Type fields is set to Typeless. This means that none of these variables will show up in the variable list if you attempt to use them in a Distribution node, Matrix node, or Web graph node. These nodes are designed to only list Nominal and Ordinal fields. However, before you can change them to Nominal, you have to change a stream option to allow Nominal fields with more than 250 values which is the default maximum. While this might not be a problem with the Type fields, almost certainly all the Concept fields have more than 250 unique values.

15. Click **OK**.
16. From the **Tools** menu, point to **Stream Properties**, and then click **Options**.
17. Deselect **Maximum members for nominal fields**.

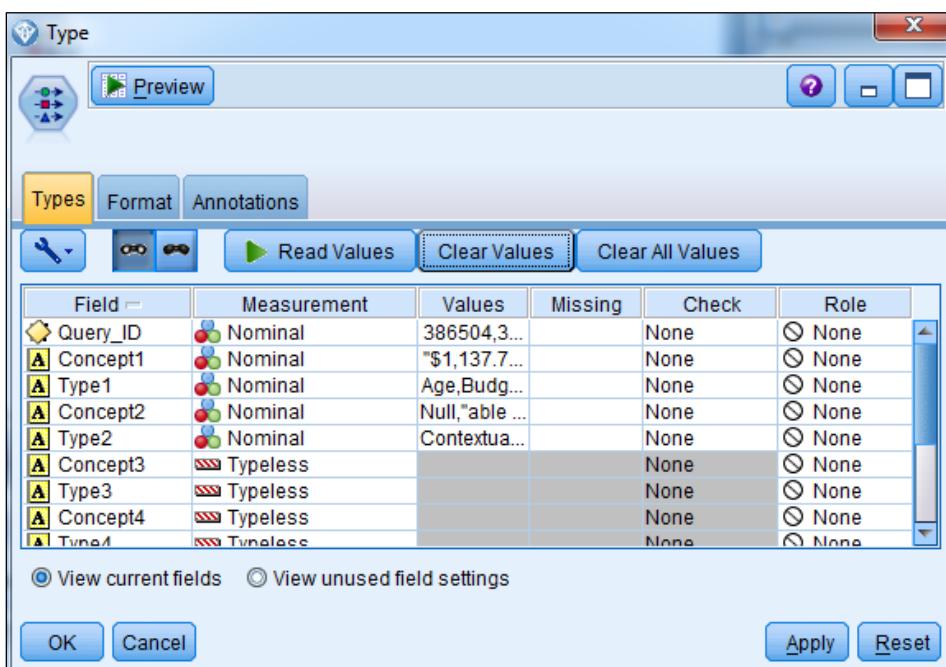


This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

18. Click **OK**, and then edit the **Type** node.
19. Select **Concept1** and Shift+click **Type 2** (none of the other slots were used so you do not have to worry about them).
20. Right click the highlighted area, point to **Set Measurement**, and then click **Nominal**.
21. Right click the highlighted area, point to **Set Values**, and then click **<Read>**.



22. Click **Read Values**.



23. Click **OK**.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

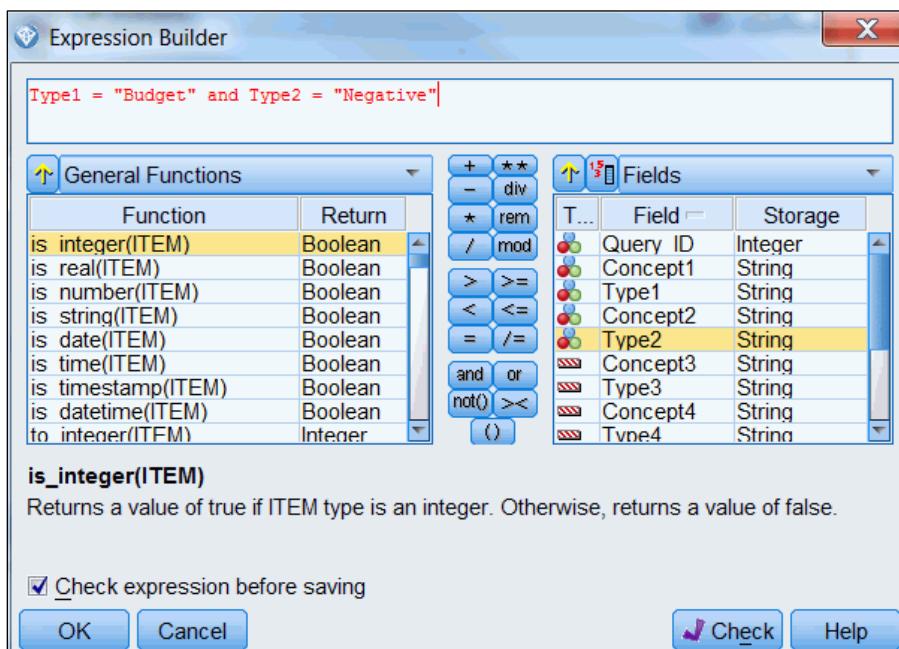
© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Task 3. Creating reports from the TLA results.

Now you can use these fields in any node that accepts Nominal or Ordinal level variables. Earlier in this module, you identified <Budget>+<Negative> as an important pattern. You will get a list of the customer IDs that had this pattern in their response.

1. From the **Record Ops** tab, add a **Select** node to the stream.
2. Connect the **Select** node to the **Type** node, and then edit the **Select** node.
3. Use the expression builder to create the expression **Type1 = "Budget"** and **Type2 = "Negative"**.



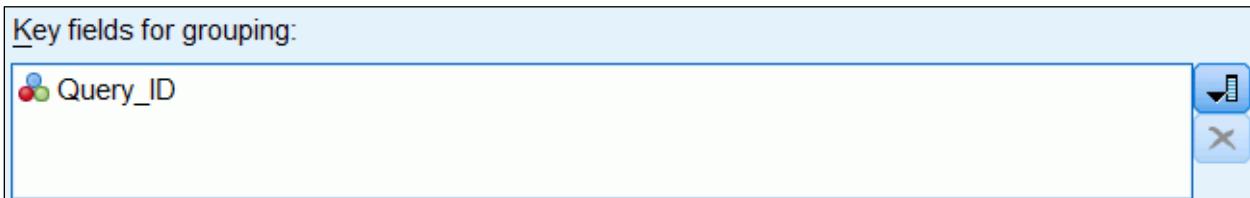
4. Click **OK** to exit the Expression Builder.

5. Click **OK** to exit the Select node.

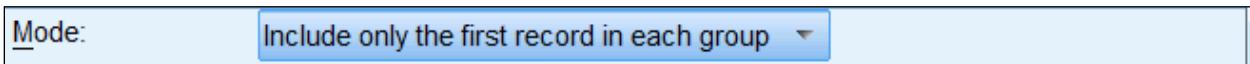
Because some of the customer responses may contain more than one occurrence of this pattern, you need to use the **Distinct** node so you do not double count them.

6. From the **Record Ops** tab, add a **Distinct** node to the stream.
7. Connect the **Distinct** node to **Select** node.
8. Edit the **Distinct** node.

- Select **Query_ID** as the **Key field for grouping**.



- Beside **Mode**, select **Include only the first record in each group**.



- Click **OK**.

- From the **Output** tab, add a **Table** node to the stream.
- Connect the **Table** node to **Distinct** node.
- Run the **Table** node.

	Query_ID	Concept1	Type1	Concept2	Type2	Concept3	Type3	Concept4	Type4	C
1	386548	paid	Budget	would be good	Negative	Null	Null	Null	Null	Null
2	386636	sg payment for cable	Budget	fault	Negative	Null	Null	Null	Null	Null
3	387264	compensation	Budget	noisy	Negative	Null	Null	Null	Null	Null
4	387445	compensation for service	Budget	difficult	Negative	Null	Null	Null	Null	Null
5	387677	pay	Budget	would be good	Negative	Null	Null	Null	Null	Null
6	387712	credit management	Budget	complaint	Negative	Null	Null	Null	Null	Null
7	387757	rebate	Budget	not eligible	Negative	Null	Null	Null	Null	Null
8	387844	astrocomm to raise	Budget	problem	Negative	Null	Null	Null	Null	Null
9	388024	paying	Budget	not satisfied	Negative	Null	Null	Null	Null	Null
10	388040	bill number q99962	Budget	problem	Negative	Null	Null	Null	Null	Null

The output contains 101 records which contained the <Budget> +<Negative> pattern. The customer ID appears at the beginning of each record.

- Click **OK** to close the **Table**.
- From the **File** menu, click **Save Stream As**.
- Name the stream **Performing Text Link Analysis_demo4_end.str**, and then click **Save**.
- From the **File** menu, click **Exit**.

Result:

You have been able to successfully use the Text Link Analysis node to in conjunction with customer data to relate Text Link Analysis patterns with customer demographics.

Apply your Knowledge

Purpose:

Test your knowledge of the material covered in this module.

- Question 1: True or False: Concept patterns from various type patterns can be combined into a single category.
- A. True
 - B. False
- Question 2: True or False: The Text Link Analysis node and interactive text link analysis will produce the exact same patterns if the same linguistic resources are used.
- A. True
 - B. False
- Question 3: True or False: Text Link Patterns that you identify in the Interactive Workbench cannot be used for subsequent analyses after you exit from the workbench.
- A. True.
 - B. False
- Question 4: True or False: Text Link Analysis is primarily a knowledge discovery tool.
- A. True
 - B. False
- Question 5: True or False: Text Link Analysis cannot be used to perform Sentiment analysis.
- A. True
 - B. False

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Apply your Knowledge - Solutions

Answer 1: A. True.

Answer 2: A. True, as long as any changes that you made in the Interactive Workbench have been saved to a template or Text Analysis Package and both analyses are using the same resources.

Answer 3: B. False, provided they have been converted into Categories. Otherwise, they cannot be used outside the Interactive Workbench.

Answer 4: B. False. Text Link Analysis is performed using rules which are designed to identify patterns you expect to find in the data. While there is an element of knowledge discovery, for example you may be unsure how much positive or negative sentiment you will find in customer response, for the most part it is not a knowledge discovery tool.

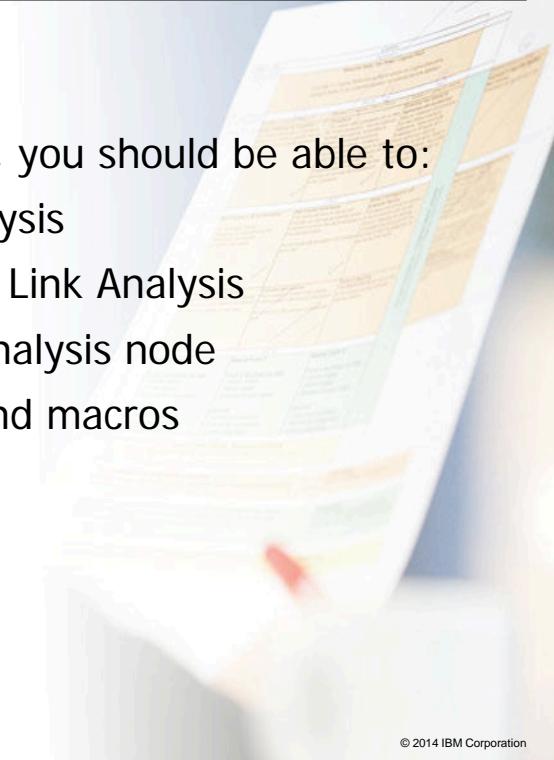
Answer 5: B. False. However, in order to perform sentiment analysis, you must be using a set of text link rules which are designed to capture sentiment. Otherwise, if your rules are not designed to identify customer sentiment, then the answer is True.

Business Analytics software

IBM

Summary

- At the end of this module, you should be able to:
 - perform Text Link Analysis
 - review interactive Text Link Analysis
 - review the Text Link Analysis node
 - create text link rules and macros



© 2014 IBM Corporation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

9-60

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Business Analytics software



Workshop 1

Text Link Analysis



© 2014 IBM Corporation

The following file(s) will be used:

- Music_Survey with Dictionary Edits.str - a Modeler stream that reads from a file containing customer likes and dislikes about portable music

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Workshop 1:Text Link Analysis

You want to use Text Link Analysis to discover which aspects of their portable music players customers are positive about and also which ones they are negative about.

- Create text link analysis patterns.
- Explore the patterns using the visualization pane.
- Select the Sound & Good and Sound & Excellent concept patterns to view the data.
- Create a Good-Excellent Sound category for the Sound & Good and Sound & Excellent concept patterns.
- Update the modeling node and save the stream as Music Survey with Text Link Analysis.str.

For more information about where to work and the workshop results, refer to the Tasks and Results section that follows. If you need more information to complete a task, refer to earlier demos for detailed steps.

Workshop 1: Tasks and Results

- Open the C:\Train\0A105\09-Performing_Text_Link_Analysis\Music_survey with Dictionary Edits.str. stream
- Edit the **Text Mining** node, and select the **Model** tab.
- Select **Exploring text link analysis (TLA) results**, and then click **Run**.
- Explore the patterns using the visualization pane.
- Within the **Features & Positive** type pattern, select the **Sound & Good** and **Sound & Excellent** concept patterns to view the data.
- Click **Add to category**  to create a new category for the **sound & good** and **sound & excellent** concept patterns.
- Switch to the **Categories and Concept** pane to verify that the new category was created.
- Rename the new category **Good-Excellent Sound**.

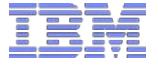
Category	Descriptors	Docs
All Documents	-	405
Uncategorized	-	
No concepts extracted	-	
Good-Excellent Sound	2	
• sound + good		
• sound + excellent		

- Update the Text Mining node and save the stream as **Music Survey with Text Link Analysis.str**.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.



Clustering Concepts

IBM SPSS Modeler Text Analytics (v16)

Business Analytics software



© 2014 IBM Corporation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Objectives

- At the end of this module, you should be able to:
 - analyze co-word clusters
 - build clusters
 - fine tune the cluster analysis
 - create categories from cluster concepts

© 2014 IBM Corporation

A cluster is a group of related concepts that occur together frequently in the same records or documents, compared to how often they occur separately. Clustering is especially useful when you are text mining documents composed of text records.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

10-3

Clusters of Concepts

- Clusters are a grouping of concepts generated by clustering algorithms based on how often concepts occur and how often they appear together.
- The goal of clusters is to group concepts that occur together.
- Clusters can help you uncover relationships among concepts that would otherwise be too time-consuming to search for.

© 2014 IBM Corporation



You may be familiar with the various types of clustering algorithms in Modeler that group records based on various measures of similarity. Clustering in the Interactive Workbench takes a different approach than any of these algorithms; its key difference is that it does not group records, but the concepts themselves. As with standard clustering routines, clustering in text mining can find dozens and dozens of clusters in a typical data file, only some of which you are likely to find interesting or useful. Your job as a data mining analyst is to sort through the cluster results and find those concept clusters that provide insight above and beyond the extracted concepts and types and any text link analysis you may have conducted.

Categories group related concepts or types, but clusters cannot directly be turned into categories. This is because, except in the simplest cluster with only 2 concepts, there is no guarantee that each concept will be linked with all other concepts in that cluster. While you cannot add entire clusters to the categories directly, you can add some or all the concepts in a cluster to a category through the Cluster Definitions dialog box.

Building Clusters

- Clusters are built from the Clusters view in the Interactive Workbench.
- To open the Interactive Workbench in this view, select Analyzing co-word clusters in the Text Mining node Model tab.

© 2014 IBM Corporation



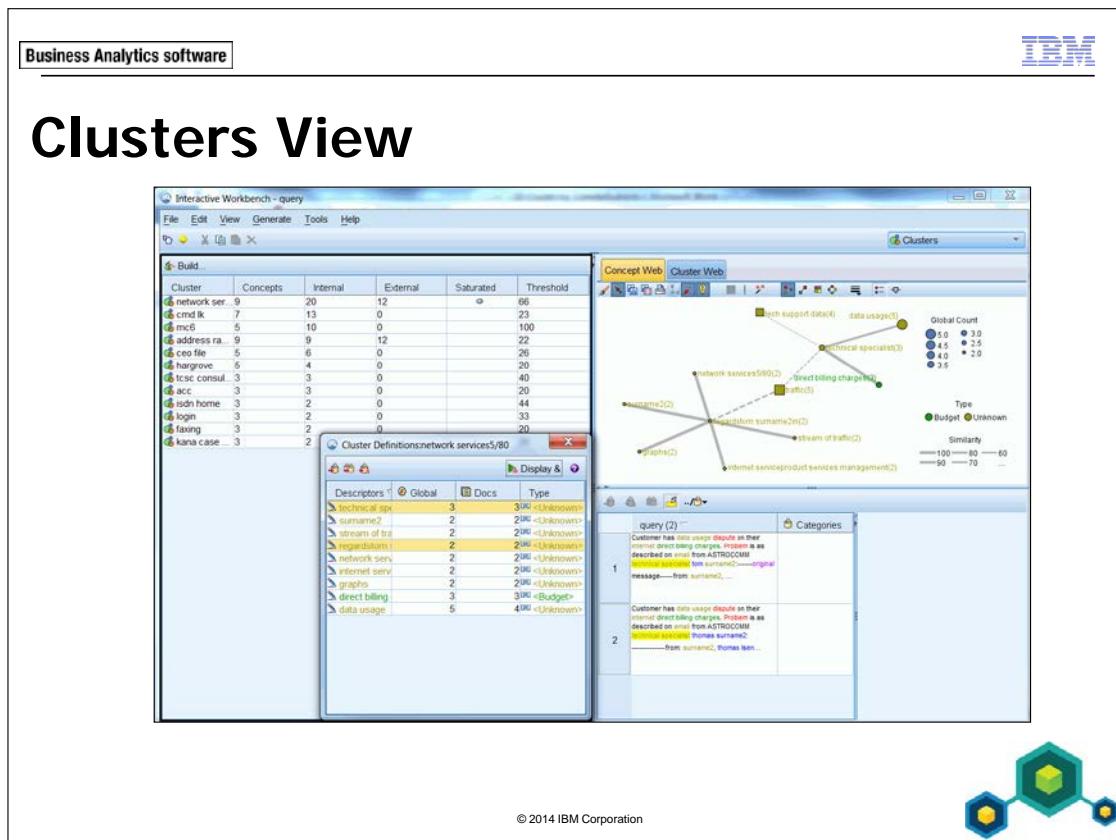
Cluster analysis can only be done within the Cluster view of the Interactive Workbench. To request a cluster analysis, you need to select the Analyzing co-word clusters option in the Text Mining node model tab. This option launches in the Clusters view and updates any outdated extraction results. In this view, you can perform co-word cluster analysis, which produces a set of clusters. Co-word clustering is a process that begins by assessing the strength of the link value between two concepts based on their co-occurrence in a given record or document and ends with the grouping of strongly linked concepts into clusters. There are various settings and limits that control the creation of clusters, which are described below.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

10-5



After the extraction occurs, the Interactive Workbench opens in the Clusters view. There are three panes in this view:

- Clusters pane: You can build and manage your clusters in this pane.
- Visualization pane: You can visually explore your clusters and their relationships in this pane.
- Data pane: This is the standard pane that enables you to view records that correspond to selections in the Cluster Definitions dialog. You cannot display any Data pane results from the Clusters pane, only the Clusters Definition dialog box.

Cluster Analysis Settings

- Cluster Analysis is very exploratory.
- There are a number of settings you need to adjust to get the optimal cluster solution.
- The defaults will probably not create the best possible solution.

© 2014 IBM Corporation



Clusters are built from descriptors derived from certain types. You can select the types to include in the building process. The types that capture the most records or documents are preselected by default.

Choose the method of selecting the concepts that you want to use for clustering. You can speed up the clustering process by reducing the number of concepts. You can cluster using a number of top concepts, a percentage of top concepts, or using all the concepts:

- Number based on doc count. When you select Top number of concepts, enter the number of concepts to be considered for clustering. The concepts are chosen based on those that have the highest doc count value. Doc count is the number of documents or records in which the concept appears.
- Percentage based on doc count. When you select Top percentage of concepts, enter the percentage of concepts to be considered for clustering. The concepts are chosen based on this percentage of concepts with the highest doc count value.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

10-7

By default, link values are calculated using the entire set of documents or records. However, in some cases, you may want to speed up the clustering process by limiting the number of documents or records used to calculate the links. Limiting documents may decrease the quality of the clusters. To use this option, select the check box to its left and enter the maximum number of documents or records to use.

By default, up to 50 clusters will be created, and you will definitely want to set this value higher in larger data files. However, clustering can require a significant amount of processing time, so be sure to sample from the data when first doing clustering, unless you are beginning with a small file, as is the case here.

Other settings include:

- Minimum concepts in a cluster: By default, a cluster need only contain 3 concepts. Do not reduce this value, and increase it only moderately. The larger the value, the fewer the clusters.
- Maximum concepts in a cluster: Typically this value is not changed until you have tried other settings.
- Maximum number of internal links: Links between concepts are what lead to the creation of a cluster, so normally the more links the better. If you are getting saturated clusters (defined below), then you can increase this value.
- Maximum number of external links: Although not important for building clusters, external links are links between concept pairs in separate clusters. Again, if you are getting saturated clusters, you could increase this value.
- Minimum link value: This value is the smallest link value accepted for a concept pair to be considered for clustering. Link value is calculated using a similarity formula.
- Prevent pairing of specific concepts: You can specify concepts that should not be linked together.

Setting the Cluster Link Value

- The similarity link value is measured using the co-occurrence document count compared to the individual document counts for each concept in the relationship.
- The algorithm reveals those relationships that are strongest, meaning that the tendency for the concepts to appear together in the text data is much higher than their tendency to occur independently.

© 2014 IBM Corporation



To explain the minimum link value setting, you should understand the calculation of the link value itself. The first key point is that the building blocks of clusters are pairs of concepts. Two concepts that co-occur together often are considered to be linked. After finding concept pairs, the clustering process groups similar concepts into clusters by aggregation, taking into account their link values and the settings defined in the Cluster Settings dialog. Aggregation consists of adding concepts to an existing cluster or merging a smaller cluster into a larger cluster until a cluster is saturated. A cluster is saturated when additional merging would exceed the limits set in the Cluster Settings dialog.

The link value is based upon the so-called similarity coefficient. The similarity coefficient is calculated using the co-occurrence document count compared to the individual document counts for each concept in a pair.

The default link value is set to 20. If you get lots of clusters (assuming you set the maximum number of clusters above 50), then you might increase the link value. If you do not get enough clusters, you could reduce it. However, the smaller the value, the less connected, and possibly less meaningful, the clusters.

The similarity coefficient is calculated using the following formula:

$$\text{Similarity Coefficient} = \frac{(C_{ij})^2}{(C_i \times C_j)}$$

Where:

- C_i is the number of documents or records in which the concept i occurs.
- C_j is the number of documents or records in which the concept j occurs.
- C_{ij} is the number of documents or records in which concept pair i and j co-occurs in the set of documents.

When calculating the similarity coefficient, the unit of measurement is the number of documents (doc count) in which a concept or concept pair is found. The algorithm reveals those relationships that are strongest, meaning that the tendency for the concepts to appear together in the text data is much higher than their tendency to occur independently. Internally, the algorithm yields a similarity coefficient ranging from 0 to 1, where a value of 1 means that the two concepts always appear together and never separately. The similarity coefficient is then multiplied by 100 and rounded to the nearest whole number to arrive at the similarity link value or, shortly, link value.

For example, consider the following table:

Concept/Pair	Scenario A	Scenario B
Concept i	Occurs in 20 docs	Occurs in 30 docs
Concept j	Occurs in 20 docs	Occurs in 60 docs
Concept Pair ij	Co-occurs in 20 docs	Co-occurs in 20 docs
Similarity coefficient	1	0.22222
Similarity link value	100	22

Initial Clustering Results

Cluster	Concepts	Internal	External	Saturated	Threshold
address	10	10	14	●	15
title	9	20	0	●	15
isdn	10	12	0		10
mc6	5	10	0		100

© 2014 IBM Corporation



After building the initial clusters, the output will display the cluster name (named after the concept with the highest number of internal links); the number of concepts in the cluster; the number of internal and external links in the cluster; the threshold value; and whether or not the cluster was saturated.

The first four column names are fairly self-explanatory, but the last two need some further explanation:

- Threshold: For all of the co-occurring concept pairs in the cluster, this is the lowest similarity link value of all in the cluster. A cluster with a high threshold value signifies that the concepts in that cluster have a higher overall similarity and are more closely related than those in a cluster whose threshold value is lower.
- Saturation: Cluster could have been larger but one or more limits have been exceeded and therefore, the clustering process ended for that cluster and is considered to be saturated. At the end of the clustering process, saturated clusters are presented before unsaturated ones and therefore, many of the resulting clusters will be saturated.

Clusters are created from pairs of concepts, and each pair forms an internal link, so the internal link value lists the number of links between pairs in a cluster. The maximum possible number of internal links occurs when every concept is linked to every other concept. For example, the cluster isdn contains 10 concepts. The maximum number of internal links is $(10^2)/2=45$, and this cluster has only 10 internal links. This cluster has no external links between its concepts and those in other clusters.

There is no information about how many records are members of a cluster because there is no one answer to that question.

To see more unsaturated clusters, you can change the Maximum number of clusters to create setting to a value greater than the number of saturated clusters or decrease the Minimum link value. If the threshold values are lower than you want, you can try increasing the Minimum link value higher to get tighter clusters in which the concepts have a higher degree of similarity.

Be careful not to attribute too much meaning to the name given a cluster. You need to review the concepts themselves rather than rely on the name generated by the software.

Also, do not be deterred by a low threshold value for a cluster. Recall that this value represents the lowest link value of all pairs in a cluster. However, within the cluster there could be links that are substantially higher.

Cluster Web Graphs

- Cluster Web graphs help you see how the concepts and clusters are linked
- The thicker the line, the stronger the connection
- Internal links are displayed with solid lines
- External links are displayed with dotted lines

© 2014 IBM Corporation



The Clusters view can display two types of Web graphs:

- Concept Web Graph: The graph shows the connections between all of the concepts within the selected cluster(s) as well as linked concepts outside the cluster. This graph can help you see how the concepts within a cluster are linked and any external links. The internal links between the concepts within a cluster are drawn and the line thickness of each link is directly related to either the doc count for each concept pair's co-occurrence or the similarity link value, depending on your choice on the graph toolbar. The external links between a cluster's concepts and those concepts outside the cluster are also shown.
- Cluster Graph: The graph displays a Web graph showing the selected cluster(s). The external links between the selected clusters as well as any links between other clusters are all shown as dotted lines. In a Cluster Web graph, each node represents an entire cluster and the thickness of lines drawn between them represents the number of external links between two clusters.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

10-13

In order to display a Cluster Web graph, you must have already built clusters with external links. External links are links between concept pairs in separate clusters (a concept within one cluster and a concept in another cluster).

You can switch between having the thickness of the lines represent the similarity value or the co-occurrence value (number of records in which both appear) with the size

links  button.

Cluster Concepts into Categories

- Clusters cannot be converted into categories.
- Concepts within Clusters can be added to existing categories or converted into new ones.

© 2014 IBM Corporation



There will be times when you want to create a category from the concepts comprising a cluster. You can do this from the Cluster Definitions dialog box. For example, if you want to create a category for the concept pair poor + coverage. There are three ways of doing it:

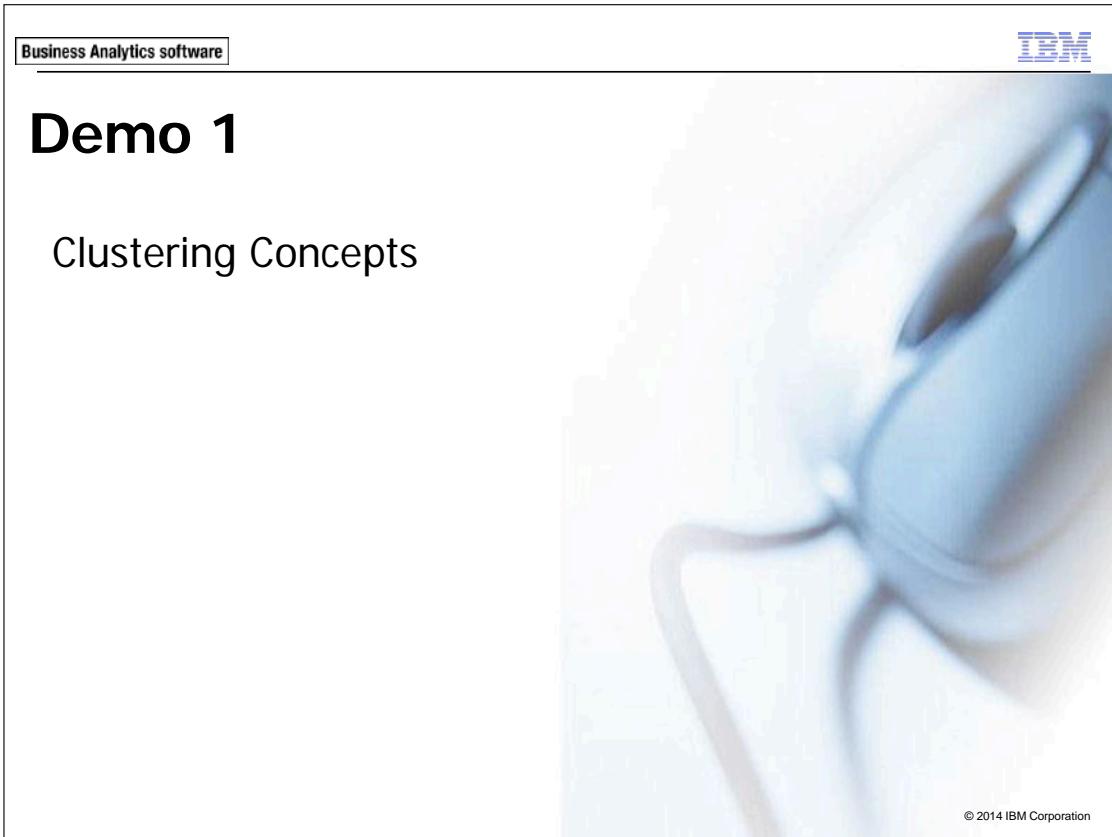
- Use the Add to Category tool to add each term to the same category. However, when you do it this way, each term will be treated separate from the other. For instance, in the Astroserve data, there were 19 occurrences of poor, 16 occurrences of coverage, but only 7 co-occurrences. Customer responses that either contained one of the terms or both terms would be placed in the same category.
- Use the Create categories for each descriptor tool to create separate two separate categories, one for "poor" and other for "coverage".
- Use the Add as AND (&) rule to category tool to create a category which pairs poor and coverage together. Only customers who mentioned both "poor" and "coverage" together would be put into this category.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

10-15



The slide has a light blue background with a faint, stylized graphic of a person's head and shoulders. In the top left corner, there is a small white box containing the text "Business Analytics software". In the top right corner, the "IBM" logo is displayed. The main title "Demo 1" is centered at the top in a large, bold, black font. Below it, the subtitle "Clustering Concepts" is also centered in a smaller, regular black font. In the bottom right corner of the slide area, there is a small copyright notice: "© 2014 IBM Corporation".

The following file(s) are used in this demo:

- Clustering_Concepts_demo1_start.str - a Modeler stream that reads a file containing call center data for March and April

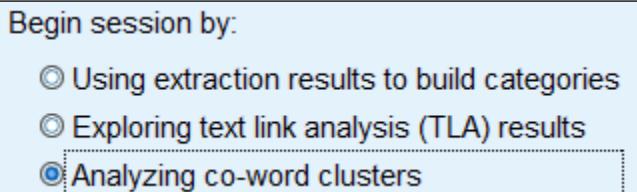
Demo 1: Clustering Concepts

Purpose:

You would like to find out if there are any pairs of concepts that tend to co-occur which may help you to predict churn. For example, if the words "poor" and "service" were often mentioned.

Task 1. Requesting a cluster analysis.

1. From the **File** menu, click **Open Stream**.
2. Navigate to **C:\Train\0A105\10-Clustering_Concepts**, and then double-click **Clustering_Concepts_demo1_start.str**.
3. Edit the **Text Mining** node labeled **query**.
4. On the **Model** tab, ensure that **Analyzing co-word clusters** is selected.



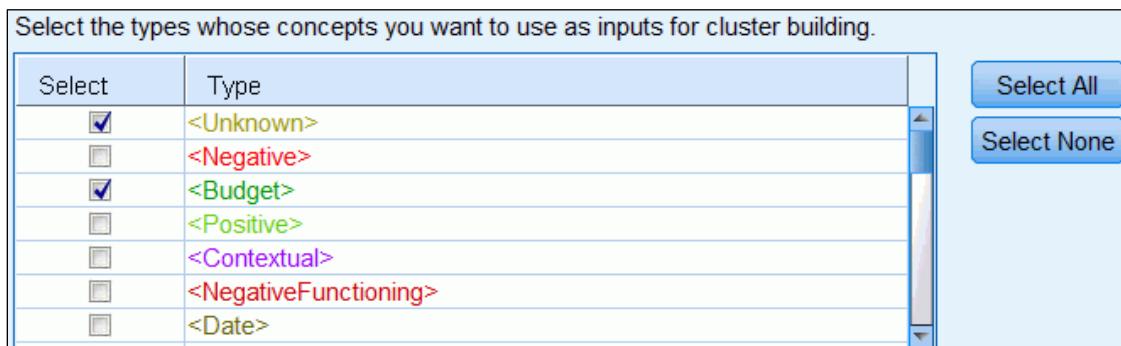
5. Click **Run**.

Cluster	Concepts	Internal	External	Saturated	Threshold
network ser...	9	20	12		66
cmd lk	7	13	0		23
mc6	5	10	0		100
address ra...	9	9	12		22
ceo file	5	6	0		26
hargrove	5	4	0		20
tcsc consul...	3	3	0		40
acc	3	3	0		20
isdn home	3	2	0		44
login	3	2	0		33
faxing	3	2	0		20
kana case ...	3	2	0		20

Because these clusters were created using the default settings, it is not surprising that none of the clusters appears to be very useful to help predict churn because you took all the default settings. You will see better results if you change some of these settings.

6. Click **Build**.

The Cluster Settings dialog appears.



At the top of the dialog is a list of the Types that were used to build the clusters. In this case, you would certainly want to use the types that were created previously, such as <Phones> and <Competitors>. In addition, it might be interesting to explore whether there are any combinations of concepts that co-occur together that may help us understand better what products or services, if any, that astroserve customers have negative feelings about, but neither <Negative> or <NegativeFunctioning> were selected by default.

7. Click **Select All** to select all of the Types, and then click **Build Clusters**.

Cluster	Concepts	Internal	External	Saturated	Threshold
graphs	7	20	20	●	100
data usage	8	9	20	●	44
cmd lk	9	19	0		23
mc6	5	10	0		100
ceo file	5	6	0		26
hargrove	5	4	0		20
pmto	3	3	0		66
tcsc consul...	3	3	0		40
acc	3	3	0		20
isdn home	3	2	0		44
login	3	2	0		33
workload	3	2	0		33
not legal	3	2	0		25
faxing	3	2	0		20
kana case ...	3	2	0		20

On the surface, it seems again that none of the clusters would tell us much about what products or services customers are negative about. At this point, you should probably consider adjusting some of the settings to get some additional clusters. Based on the threshold values, the minimum link value within a cluster is already fairly low (20), if you reduce it even further, you may get some more meaningful clusters. The drawback is that the concepts within the clusters may also be less connected than they were before.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

8. Click **Build**.
9. Change the **Minimum link value** to **10**.

Output Limits

Maximum number of clusters to create:	<input type="text" value="50"/>	Maximum number of internal links:	<input type="text" value="20"/>
Maximum concepts in a cluster:	<input type="text" value="10"/>	Maximum number of external links:	<input type="text" value="20"/>
Minimum concepts in a cluster:	<input type="text" value="3"/>	Minimum link value:	<input type="text" value="10"/>
<input type="checkbox"/> Prevent pairing of specific concepts		Manage Pairs...	

10. Click **Build Clusters**.

The results appear as follows:

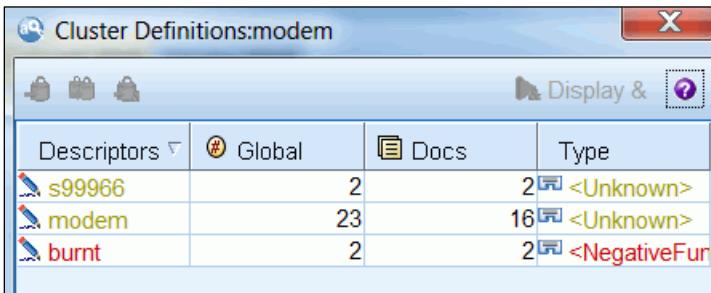
Build...

Cluster	Concepts	Internal	External	Saturated	Threshold
dealer	6	6	0		11
firewall	6	6	0		10
rescheduled	6	6	0		10
august	5	5	0		10
login	4	4	0		13
cp	4	4	0		12
varied	5	4	0		10
kana case	4	3	1		13
tcsc consu...	3	3	0		40
acc	3	3	0		20
username	3	3	0		14
cool	4	3	0		12
coverage	4	3	0		11
not conveni...	4	3	0		11
supervisor	4	3	0		11
mob	4	3	0		10
06/03/03	3	3	0		10
not legal	3	2	0		25
intermittent	3	2	0		19
barrier	3	2	0		14
driveway	3	2	0		13
modem	3	2	0		12
shop	3	2	0		12
installid	3	2	0		11
no problem	3	2	0		11
water	3	2	0		11
christian	3	2	0		10
internet sites	3	2	0		10
loss of busi...	3	2	0		10
outgoing c...	3	2	0		10
pit	3	2	0		10
user	3	2	0		10
would reco...	3	2	0		10

Based on the names, it appears you have a few more meaningful clusters. For example one cluster is named "modem" and another is called "coverage". Both of these topics would certainly be of interest to Astroserve management. However, both of them need to be examined further before you can evaluate how useful they are. You will review the cluster named "modem".

Task 2. Exploring the cluster analysis results.

- Double-click **modem** to open the Cluster Definition window.

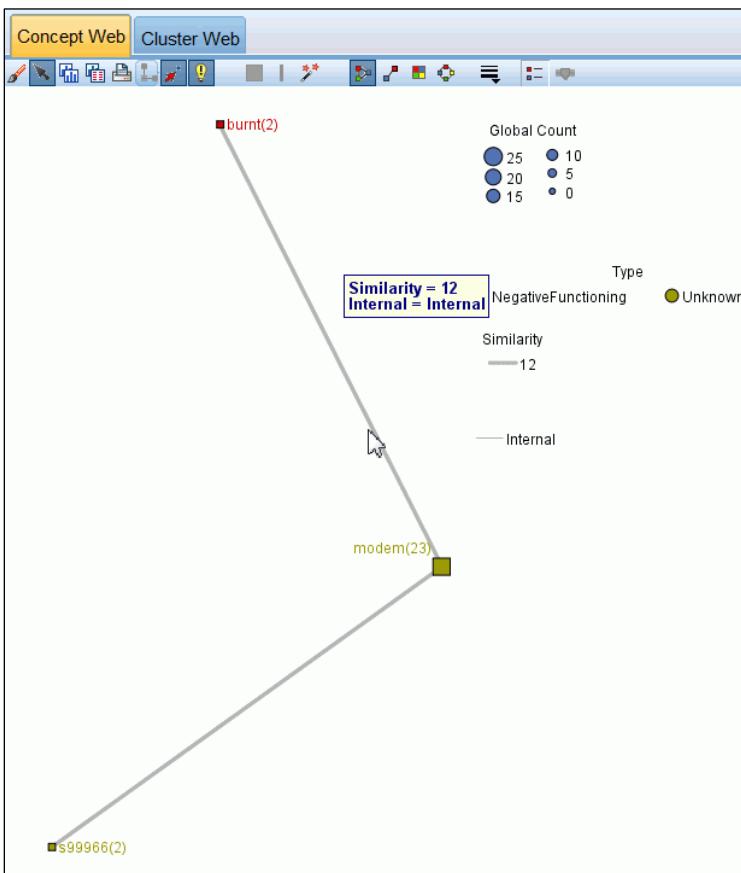


The screenshot shows the 'Cluster Definitions:modem' window. It has tabs for Descriptors, Global, Docs, and Type. The Global tab is selected, showing three rows of data:

Descriptors	Global	Docs	Type
s99966	2	2	<Unknown>
modem	23	16	<Unknown>
burnt	2	2	<NegativeFun>

This window contains a list of the terms in the cluster. While 16 customers mentioned modems in their call, only 2 customers mentioned the term "burnt", although both of them used it in conjunction with "modem". Because two is not a lot, it appears that burnt-modems are just an occasional problem rather than a pervasive one, which accounts for why the similarity value for the connection is only 12, which means that the terms do not co-occur very often.

On the right side of the Cluster View is a concept graph for the modem cluster.



You can get an exact value by hovering over the link for the connection (12).

2. Close the **Cluster Definitions:modem** window, and then double-click **coverage**.

Cluster Definitions:coverage			
Descriptors	# Global	Docs	Type
reception	4	4	<Unknown>
poor	19	19	<Negative>
coverage	16	14	<Unknown>
cdma	7	6	<Unknown>

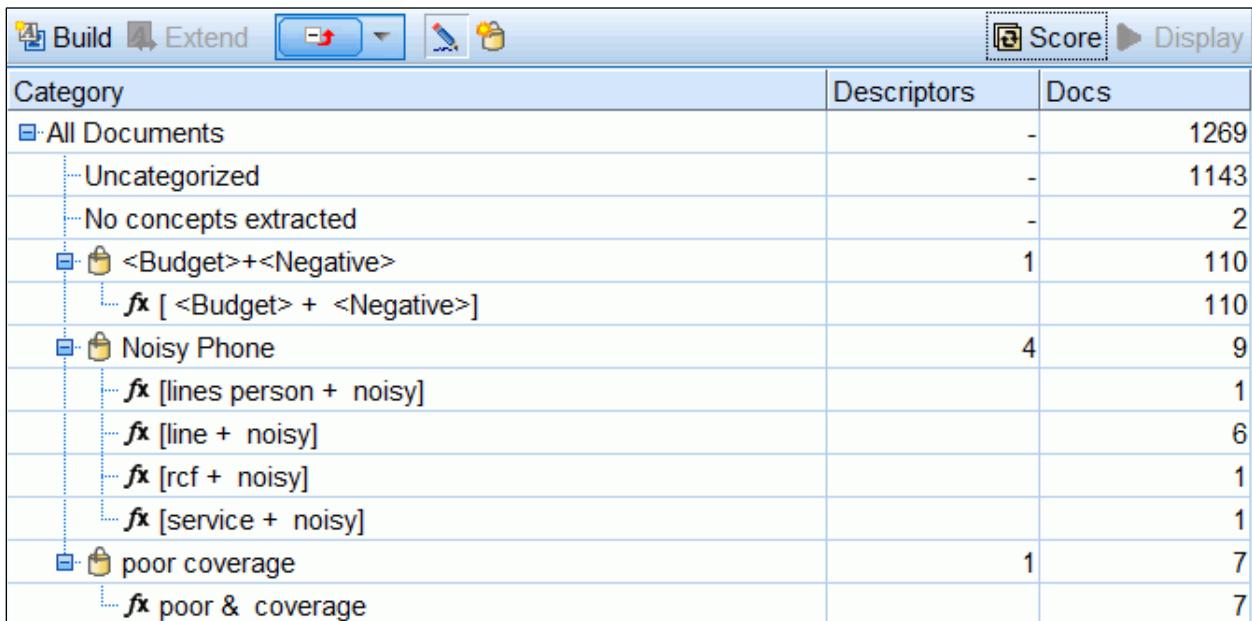
The terms "poor" and "coverage" are members of the same cluster, but it is not possible to tell from this table how many times they were used together. In other words, it is quite possible that certain customer responses contained the word "poor" or "coverage" but not both. You will review the actual text records to see how many times they occurred together.

3. Ctrl+click **coverage** and **poor**, and then click **Display &**.

	query (7) ▾	Categories
5	Customer needs monthly access fee of \$38.50 for service for the past 24 months to be credited. after connecting to cdma, she found she had poor coverage she connected a new gsm number, however the cdm...	
6	Customer raises issue with within the gisborne area about mobile coverage, customer advises coverage is poor with 5/6 km around his address. Customer needs to be contacted to dicuss.	
7	Customer advised he has poor coverage in worongary qld & had better coverage with voda & requests to port back without penalty.	

The term "coverage" appears in 7 of the records in which the word "poor" is also present, which suggests that coverage is something that Astroserve customers regularly complain about. Even though only a small number of records have this concept pair, you will create a category from the cluster because the issue of poor coverage is definitely something that Astroserve will undoubtedly want to investigate further.

4. Click **Add as AND (&) rule to category**  in the **Cluster Definitions:coverage** toolbar.
5. Select **Create New Category**, and then click **OK**.
6. Close the **Cluster Definitions:coverage** window.
7. Switch to the **Categories and Concepts** view.
8. Right-click **rule 1**, and then click **Rename Category**.
9. Name the new category **poor coverage**, and then click **OK**.
10. Click **Score**.



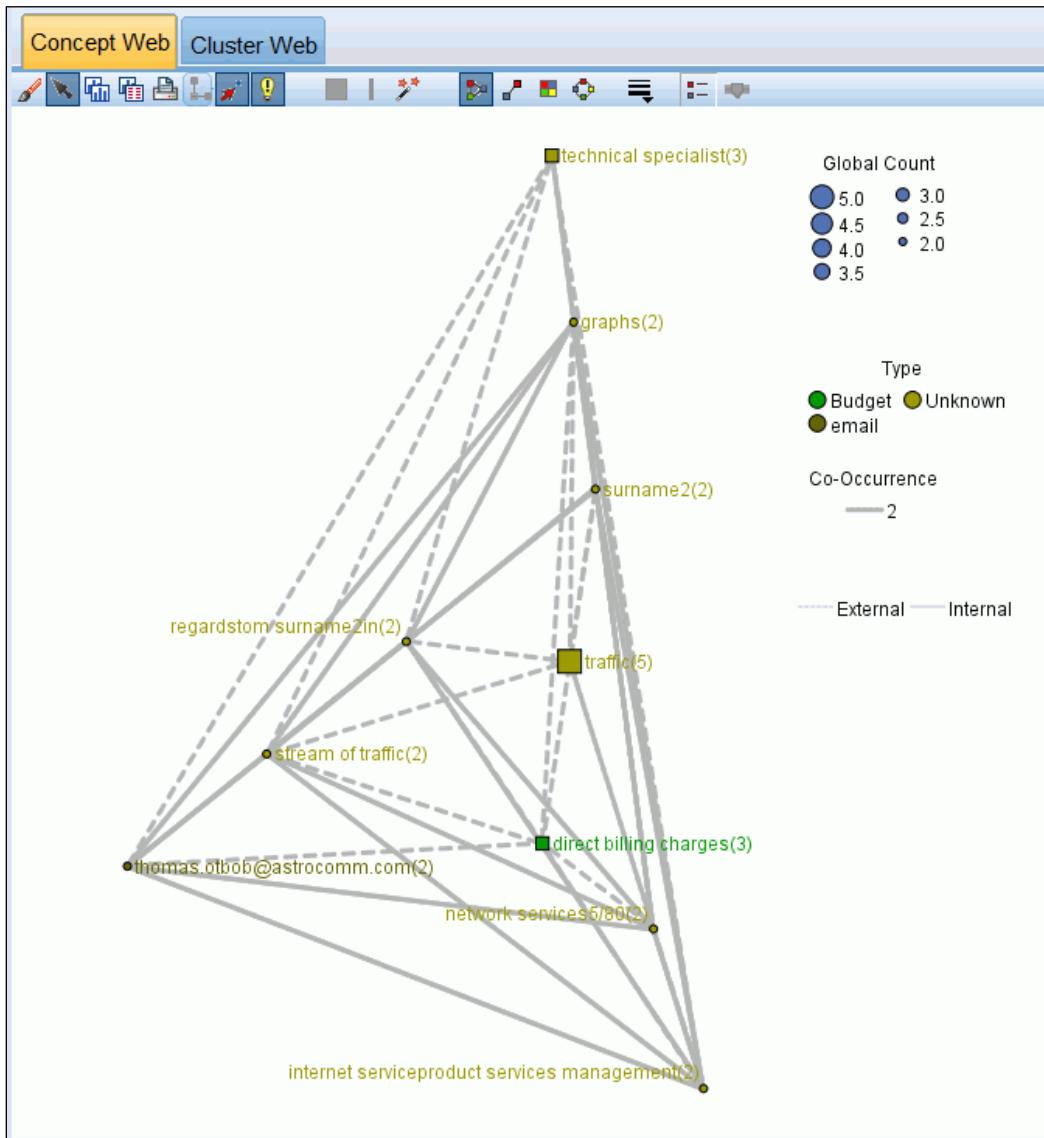
The screenshot shows the 'Categories and Concepts' view in IBM SPSS Modeler. The toolbar at the top includes 'Build', 'Extend', and other icons. The main area is a table with three columns: 'Category', 'Descriptors', and 'Docs'. The 'Category' column displays a hierarchical tree structure. The 'Descriptors' and 'Docs' columns show the count of each descriptor and the total number of documents it covers. The data is as follows:

Category	Descriptors	Docs
All Documents	-	1269
Uncategorized	-	1143
No concepts extracted	-	2
<Budget>+<Negative>	1	110
fx [<Budget> + <Negative>]		110
Noisy Phone	4	9
fx [lines person + noisy]		1
fx [line + noisy]		6
fx [rcf + noisy]		1
fx [service + noisy]		1
poor coverage	1	7
fx poor & coverage		7

The 7 customers have all been placed into the "poor coverage" category.

11. Switch to the **Clusters** view.

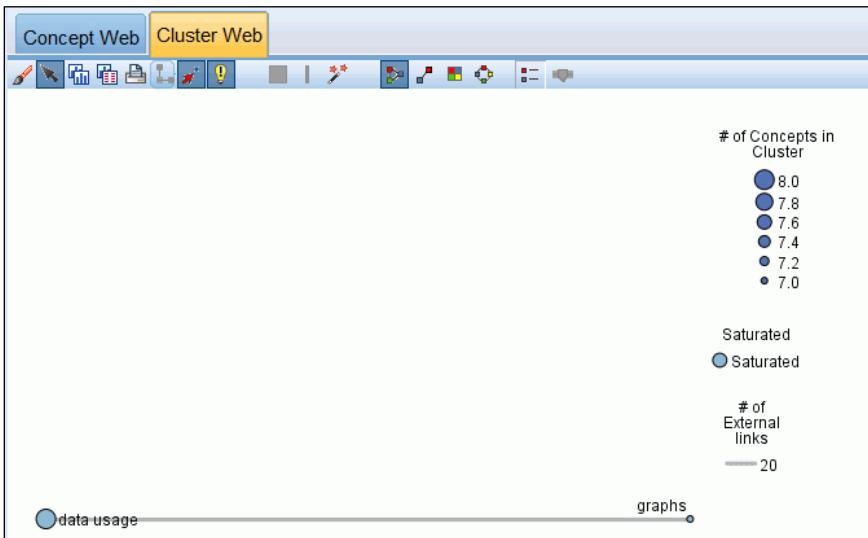
12. At the top of the list of clusters, select the **graphs** cluster.



The external links are represented by dashed lines, the internal by solid lines. This cluster has 7 concepts, and least 20 internal links, and 20 external links, so it is quite complicated. Undoubtedly these numbers would go higher if raised the maximum number of maximum number of internal and external links to above the default value of 20 when the clusters were built.

You will examine the cluster Web graph.

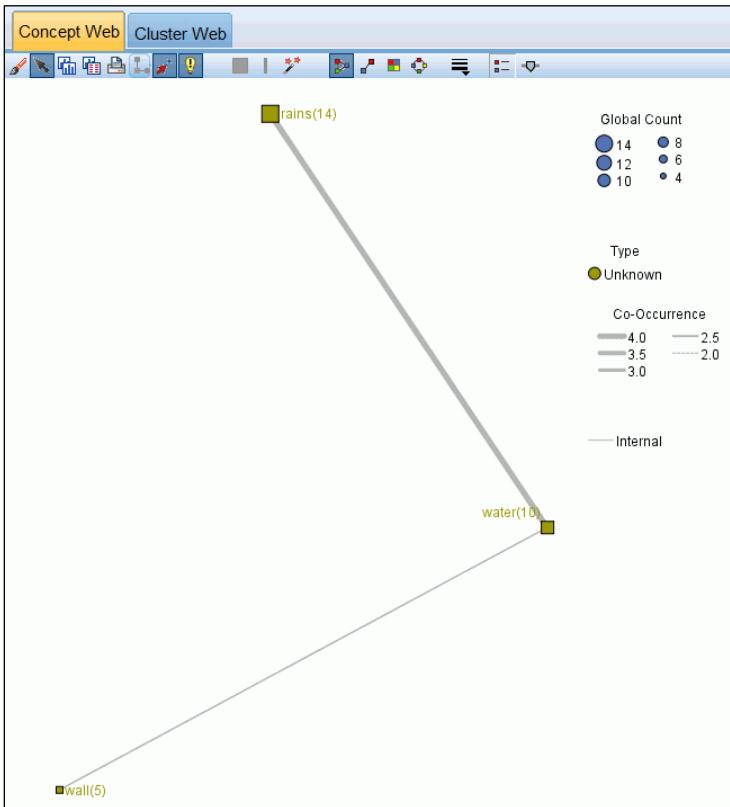
13. Click the **Cluster Web** tab.



The "graph" cluster has external links with the "data usage" cluster.

The clustering algorithm will also find clusters that occur often but are uninteresting. To verify this:

14. Select the **water** cluster, and then click the **Concept Web** tab in the Visualization pane.



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

The concept pair of water and rain means that these two terms go together in customer comments. Nothing new is learned from the graph (except perhaps for the frequency of occurrence), and so you will need to drill down to find out what the problems are with water and rain, as the cluster does not provide that. You could continue exploring the clusters to see what else can be gleaned from this analysis. You could also change some of the default settings to obtain more, or fewer, clusters, to see if anything different can be learned. This process can take some time, just as when doing cluster analysis on records. You will have to decide how deeply to review and modify the cluster results.

Task 3. Saving the cluster analysis results.

1. From the **File** menu, click **Update Modeling Node**.
2. Click **OK** and then **OK** again when you get a message that the Modeling node has been updated.
3. From the **File** menu, click **Close**, and then click **Exit**.
4. From the **File** menu, click **Save Stream As**.
5. Name the stream **Clustering_Concepts_demo1_end.str**, and then click **Save**.
6. From the **File** menu, click **Exit**.

Result:

You have successfully identified concepts that co-occur together that may help to account for churn among Astroserve customers.

Apply Your Knowledge

Purpose:

Test your knowledge of the material covered in this module.

Question 1: True or False: Cluster saturation occurs when additional merging would exceed the limit set for building the clusters.

- A. True
- B. False

Question 2: True or False: A low threshold link value for a cluster indicates that all concept pairs in the cluster are not strongly associated.

- A. True
- B. False

Question 3: True or False: Clusters can be converted into categories, but concepts within clusters cannot be turned into categories.

- A. True
- B. False

Question 4: True or False: You cannot name the clusters yourself.

- A. True
- B. False

Question 5: True or False: A similarity value of 100 indicates that the pair of concepts do not co-occur together very often.

- A. True
- B. False

Apply your Knowledge - Solutions

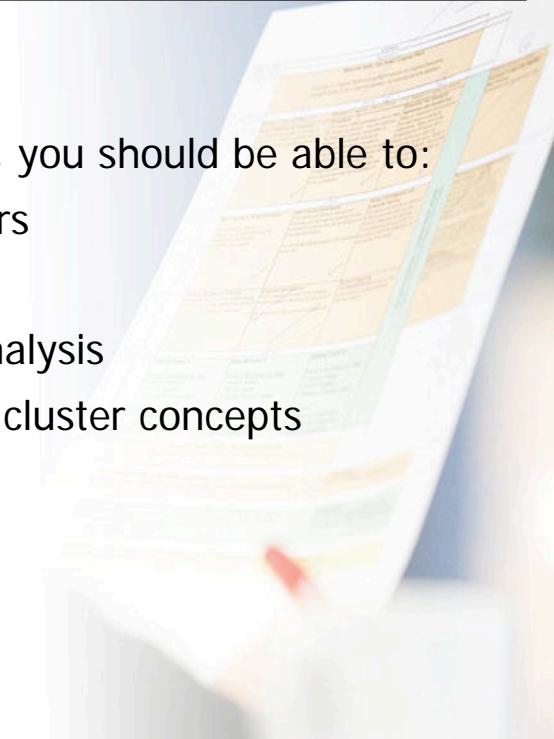
- Answer 1: A. True.
- Answer 2: B. False. The threshold level is assigned to the pair of terms that has the lowest similarity link value of all in the cluster. Other pairs of terms could have much higher values.
- Answer 3: B. False. Only concepts within clusters can be turned into categories.
- Answer 4: A. True. You cannot name the clusters yourself. The name is automatically assigned based on the concept with the highest number of internal links.
- Answer 5: B. False. A similarity value of 100 indicates that the terms usually occur together.

Business Analytics software

IBM

Summary

- At the end of this module, you should be able to:
 - analyze co-word clusters
 - build clusters
 - fine tune the cluster analysis
 - create categories from cluster concepts



© 2014 IBM Corporation

Business Analytics software



Workshop 1

Clustering Concepts



© 2014 IBM Corporation

The following file will be used:

- Music_Survey with Text Link Analysis.str - a Modeler stream that reads from a file containing customer likes and dislikes about a portable music player

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

10-29

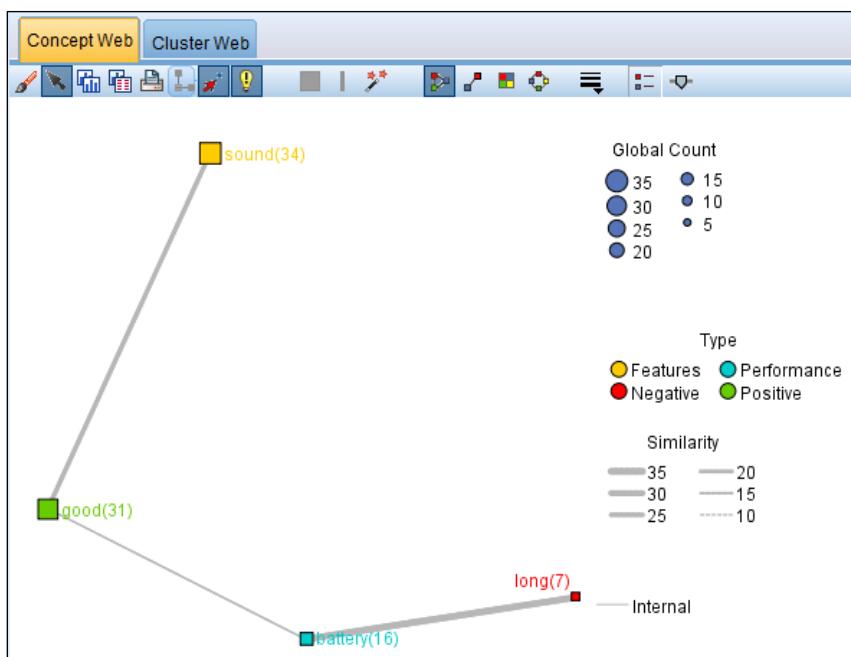
Workshop 1: Clustering Concepts

The objective of this workshop is to become familiar with the Clusters pane and to examine if there are any pairs of concepts that frequently co-occur together. For example, it would be interesting to find out which player features, if any, are linked with positive or negative sentiment.

- Open the C:\Train\0A105\10-Clustering_Concepts\Music_survey with Text Link Analysis.str stream.
- Edit the Text Mining node.
- Begin session by Analyzing co-word clusters.
- Run the Text Mining node.
- Build the clusters.
- Create co-word clusters by:
 - selecting all the types
 - changing the Maximum number of internal links value to 50
 - changing the Maximum number of external links value to 50
 - changing the Minimum link value to 10
 - click Build Clusters
- Explore the clusters using the visualization pane.
- Within the good cluster, select the sound and battery descriptors to view the data.
- There is no need to update the modeling node.

Workshop 1: Tasks and Results

- Open the C:\Train\0A105\10-Clustering_Concepts\Music_survey with Text Link Analysis.str.
- Edit the **Text Mining** node.
- Begin session by **Analyzing co-word clusters**.
- Run the **Text Mining** node.
- Click **OK** when you see the Message box that says that No clusters could be created with current settings.
- Build the clusters:
 - Select all of the types.
 - Change the **Maximum number of internal links** value to **50**.
 - Change the **Maximum number of external links** value to **50**.
 - Changing the **Minimum link** value to **10**.
 - Click **Build Clusters**.
- Explore the clusters using the visualization pane.
- Select the **good** cluster to view the graph.



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

10-31

- Double-click the **good** cluster to view the descriptors.

Cluster Definitions:good			
Descriptors	Global	Docs	Type
sound	34	33	<Features>
long	7	7	<Negative>
good	31	30	<Positive>
battery	16	16	<Performance>

- Ctrl+click **good** and **sound**.

Cluster Definitions:good			
Descriptors	Global	Docs	Type
sound	34	33	<Features>
long	7	7	<Negative>
good	31	30	<Positive>
battery	16	16	<Performance>

- Click **Display &**.

	Q1_What_do_you_like_most_about_this_portable_music_pl	Categories
1	i have a Product A. i like the small size and good sound.	
2	high quality, durable, good sound	
3	Good quality sound, easy to carry	
4	Long battery life. Good sound quality.	
5	I can always bring it with me. The sound quality is very good.	
6	Product A is fantastic. Really sharp design. Headphones are great too, not like the old in-the-ear types that always fell out. Sound quality seems good as well.	
7	good sound quality	

- There is no need to update the Text Mining node.



Categorization Techniques

IBM SPSS Modeler Text Analytics (v16)

Business Analytics software

© 2014 IBM Corporation



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Objectives

- At the end of this module, you should be able to:
 - use Text Analysis Packages to categorize data
 - use automated categorization with linguistic-based techniques
 - use frequency based categorization
 - import pre-existing categories from a Microsoft Excel file

© 2014 IBM Corporation

There are various strategies to work with IBM SPSS Text Analytics to code your text data. Several categorization techniques are available in IBM SPSS Text Analytics. It is important to understand how each of these techniques function so you can anticipate which techniques will be used and how to edit the linguistic resources to take advantage of these techniques.

The logic of text analysis is to try various techniques and combine them to improve and refine the results. Categorization proceeds so quickly with Text Analytics that there is no reason not to try alternative techniques (and settings within each method). In the module, you will review four different ways to categorize your text data in Text Analytics:

- automatic classification
- loading Text Analysis Packages
- importing pre-existing categories from a Microsoft Excel file
- manual categorization

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

11-3

As you learned earlier, in text analysis, a category refers to a group of closely related concepts, opinions, or attitudes that you find meaningful for the goals of the analysis. To be useful, a category should also be easily described by a short phrase or label that captures its essential meaning. There are three approaches to building categories, depending on the type of information you want to capture from the text.

- Simple mention of an object, thing, or person: If you are interested in how many times some concept is mentioned, without differentiating between positive and negative attitudes, then a simple coding scheme may suffice. As an illustration, you could create a category called Pace which recorded all instances in which a customer mentioned the pace of the course. This would not tell us whether the comment was favorable, neutral, or unfavorable, but you would know when the pace was mentioned.
- General sentiments: If you are interested in how many respondents make positive remarks, or negative remarks, regardless of the subject, you can create a coding scheme that focuses less on the subject than the opinion (sentiment). Thus there could be a category called General Satisfaction to group together all remarks that fell into such an attitude.
- Sentiments about specific objects, things, or persons: If you are interested in knowing that people did not like the pace of the course, or thought the teaching style kept the interest of the student, then you need to capture opinions about specific objects. This is the most complex coding scheme, and it would create categories that you might name Pace too fast or Style kept interest.

All three of these schemes can be created with the software. Naturally, it will take more work to create categories in the third scheme, but some categories that capture opinions about objects will automatically be created by the software, either from concept patterns or from rules. You need goals in mind for the text analysis, including which of these schemes are of interest in a particular set of text data. You can certainly use more than one scheme in the same project).

Thus, if you are analyzing responses from consumers about a new laundry soap, you might construct a category labeled odor that contains all the responses describing the smell of the product. Such a category would not differentiate between those who found the smell pleasing and those who found it offensive or too strong. You could then create two other categories to identify respondents who enjoyed the odor and those who disliked the odor.

Strategies for Creating Categories

- Use a Text Analysis Package.
- Build categories automatically using advanced linguistic settings.
- Import a file with predefined categories.
- Manually creating the categories and then dragging and dropping the most interesting ones to the Categories pane.

© 2014 IBM Corporation



There are many different approaches you can take when creating categories and it is not always easy to decide on which strategy is best for you. The benefit of using a Text Analysis Package is obvious: there are complex categories already defined. The downside is that many categories may have few responses, so you may have to review lots of irrelevant information.

Automated categorization techniques that come with the software offer fast and easy ways to create categories as well. For example, one of the techniques, Concept Inclusion, automatically locates concepts that contain the same term and then combines them together into a category. Thus, if many concepts contain the word "mobile" it will locate all those concepts and group them together into a category called "mobile". This is a fast and easy way to identify buzz words that customers are using, without having to manually go through all the responses to find these terms yourself.

Unfortunately, if you already have some categories in mind, almost certainly the automated categorization techniques will not find them. In that case, it may be best to import pre-existing categories from a Microsoft Excel file into the Text Analytics. After doing that, you can use automatic categorization techniques to identify concepts, patterns, and rules related to these existing categories and add them as descriptors.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

11-5

Text Analysis Package (TAP)

- A Text Analysis Package (TAP) includes a template AND categories (at least, one).
- TAPs are especially useful when:
 - your business goal is to create a category model
 - your field of application is covered by one of the shipped TAPs (opinions or CRM)
 - you do not have any predefined categories to import

© 2014 IBM Corporation



A text analysis package, also called a TAP, serves as a template for text response categorization. Using a TAP is an easy way for you to categorize your text data with minimal intervention since it contains the prebuilt category sets and the linguistic resources needed to code a vast number of records quickly and automatically. Using the linguistic resources, text data is analyzed and mined in order to extract key concepts. Based on key concepts and patterns found in the text, the records can be categorized into the category set you selected in the TAP.

You can make or update your own TAP. TAPs are usable across IBM SPSS text products. User-defined TAPs help users share or re-use categories and resources more easily.

A TAP is made up of the following elements:

- Category Set(s): A category set is essentially made up of predefined categories, category codes, descriptors for each category, and lastly, a name for the whole category set.
- Linguistic Resources: Linguistic resources are a set of libraries and advanced resources that are tuned to extract key concepts and patterns. These extraction concepts and patterns, in turn, are used as the descriptors that enable records to be placed into a category in the category set.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Business Analytics software

IBM

Demo 1

Using a Text Analysis Package to Categorize Data



© 2014 IBM Corporation

The following file(s) are used in this demo:

- Categorization_Techniques_demo1_start.str - a Modeler stream that reads a file containing call center data for May

Demo 1: Using a Text Analysis Package to Categorize Data

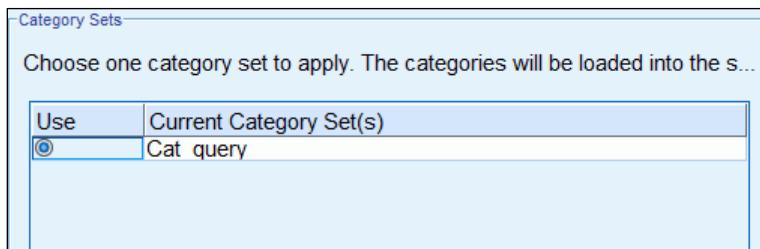
Purpose:

You already created a Text Analysis Package on Astroserve data that was collected in March and April. The company recently finished collecting data from May and would like you to use the TAP to categorize the new data so they can see whether the same patterns continued on from previous months.

Task 1. Using a Text Analysis Package to categorize data for the month of May.

1. From the **File** menu, click **Open Stream**.
2. Navigate to **C:\Train\0A105\11-Categorization_Techniques**, and then double-click **Categorization_Techniques_demo1_start.str**.
3. Edit the **Text Mining** node.
4. On the **Fields** tab, click **Field Chooser** , and then select **query**.
5. Click the **Model** tab.
6. In the **Copy Resources From** section, select **Text analysis package**, and then click **Load**.
7. Navigate to **C:\Train\0A105\11-Categorization_Techniques**, select **Astroserve0304.tap**.

A section of the results appear as follows:

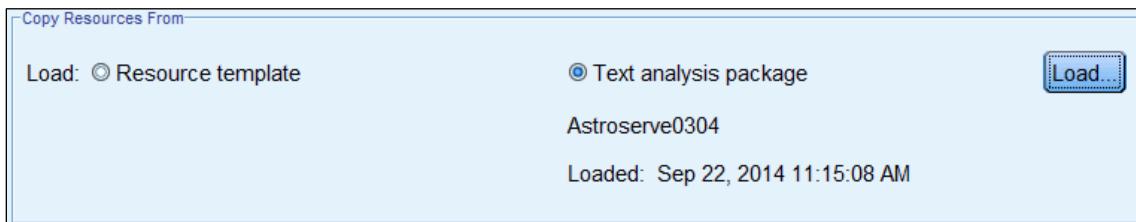


In this case there is only one set of categories, and it is already selected, but many times there is more than one to choose from. For example, most satisfaction TAPs contain three sets of categories:

- Mixed opinions (when question is « what do you think about...?»)
- Negative opinions (when question is more « what do you like least in ...?»)
- Positive opinions (when question is more « what do you like most in ...?»)

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

8. Click **Load**.



9. Click **Run**.

10. If necessary, switch to the **Categories and Concepts** view.

11. Click the **Collapse**

12. Click **Score**.

13. Click the **Docs** column header to sort the categories in descending order based on occurrence.

The results appear as follows:

Category	Descriptors	Docs ▾
All Documents	-	3138
Uncategorized	-	444
No concepts extracted	-	1
service	137	920
fee	80	476
billing	89	415
internet	39	405
number	93	352
tia	1	345
Not Working	1	342
mobile	63	318
resolved	1	302
phone	53	258
time	40	243
claims	17	230
pay	21	229
tech	1	216
cable	37	203

Based on the results, in May, service, fees and billing were at the top of the list. Also, a sizeable number of customers reported that some product they purchased from Astroserve is not working.

14. From the **File** menu, click **Close**, and then click **Exit** to end the Interactive Workbench session.
15. From the **File** menu, click **Close Stream**.
16. From the **File** menu, click **New Stream**.
Do not close Modeler. Leave it open for the next demo.

Results:

You have successfully used a Text Analysis Package, which you created from March and April data, to categorize Astroserve call center data you collected in May.

Importing Predefined Categories

- Coding frames can only be imported from Microsoft Excel.
- The coding frames must be structures in one of four ways:
 - indented with codes
 - indented with no codes
 - compact
 - flattened

© 2014 IBM Corporation



You may have an existing coding frame that the user would like to use in IBM SPSS Text Analytics for Modeler. Although you can manually re-create each category one-by-one that you had in Microsoft Excel by using the Categories \ Create New Empty Category menu, a more effective solution is to import the coding frame from Microsoft Excel. If you do so, the coding frame can be immediately used for categorization based on the category names.

Coding frames must be imported from a Microsoft Excel file, structured in one of four ways: Indented with codes; Indented without codes; Compact; or Flattened.

This is an example of the Indented with codes format:

A	B	C	D	E	F	G	H	I
1	1 Technical Features							
2	_reliable							
3	_durably constructed							
4		10 Battery	any positive comment about long battery life					
5		_long-lasting						
6		11 Storage Capacity	any positive comment about the amount that can be stored or memory capacity					
7		12 Sound Quality	any positive comment about sound, quality or music quality					
8								
9	2 Comfort							
10		20 Ease of Use	any positive comment indicating that it is convenient, easy and user-friendly					
11		21 Portability	any positive comment about mobility or indicating that it is handy and easy to transport					
12		22 Size	any positive comment indicating that it is small or compact					
13		23 Weight	any positive comment indicating that it is lightweight					
14		_light						

The content is hierarchical. It has top categories (Technical Features and Comfort) and sub-categories (Battery, Storage Capacity, Sound Quality, Ease of Use, Portability, Size, Weight). The codes for the top categories are in Column A and the codes for the sub-categories are in Column B.

This is an example of the Indented without codes format. It is the same as the previous example except that it does not have codes:

A	B	C	D	E	F	G	H	I
1	Technical Features							
2	_reliable							
3	_durably constructed							
4		Battery	any positive comment about long battery life					
5		_long-lasting						
6		Storage Capacity	any positive comment about the amount that can be stored or memory capacity					
7		Sound Quality	any positive comment about sound, quality or music quality					
8								
9	Comfort							
10		Ease of Use	any positive comment indicating that it is convenient, easy and user-friendly					
11		Portability	any positive comment about mobility or indicating that it is handy and easy to transport					
12		Size	any positive comment indicating that it is small or compact					
13		Weight	any positive comment indicating that it is lightweight					
14		_light						

The indented structure is used for code frames containing the names and codes for nets, subnets, and categories. The hierarchy among them is imposed by offsetting, or indenting, the data in the columns to the right. With the indented structure, each row in the code frame contains a net pair, subnet pair, or category pair. Basically, subnets are indented from the nets; and categories, which represent the most specific level of the code frame, are indented from both nets and subnets. The terminology of net and subnet is imported from market research.

Next is an example of Flattened format:

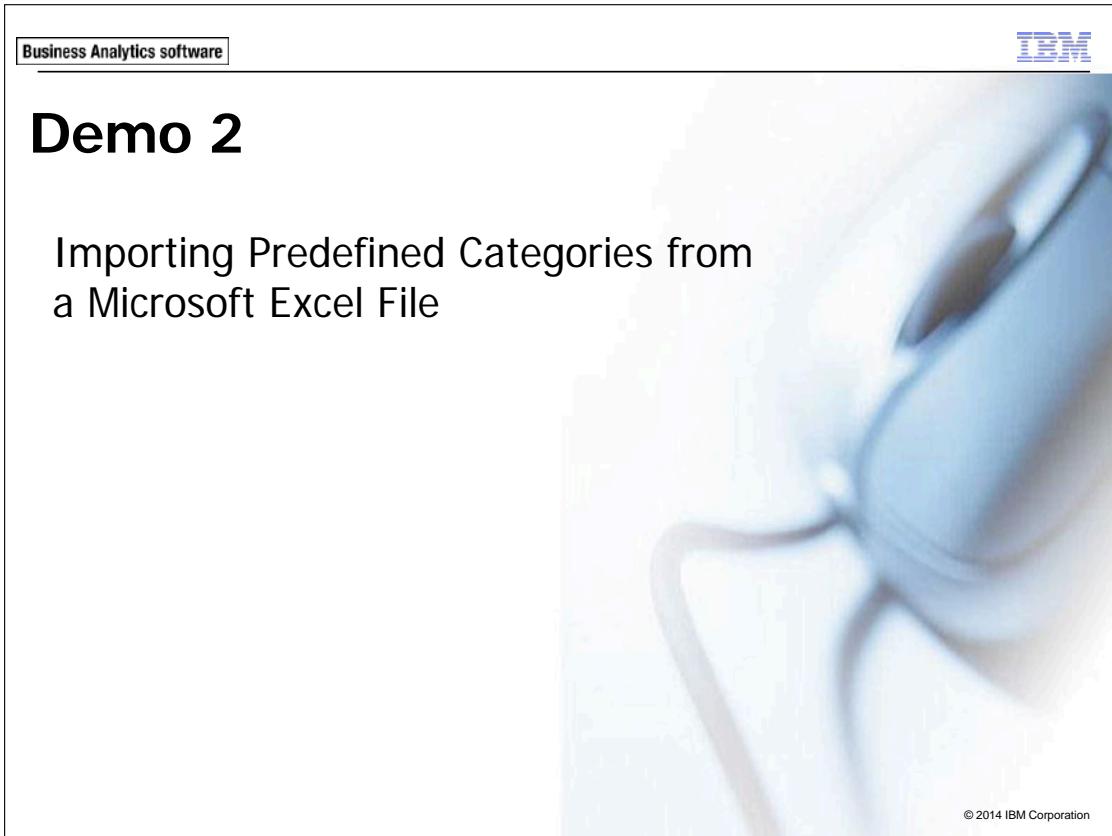
A	B	C
1	1 Technical Features	
2	_reliable	
3	_durably constructed	
4	10 Technical Features/Battery	any positive comment about long battery life
5	_long-lasting	
6	11 Technical Features/Storage Capacity	any positive comment about the amount that can be stored or memory capacity
7	12 Technical Features/Sound Quality	any positive comment about sound, quality or music quality
8	20 Comfort/Ease of Use	any positive comment indicating that it is convenient, easy and user-friendly
9	21 Comfort/Portability	any positive comment about mobility or indicating that it is handy and easy to transport
10	22 Comfort/Size	any positive comment indicating that it is small or compact
11	23 Comfort/Weight	any positive comment indicating that it is lightweight
12	_light	

There is only one top level of categories without any hierarchy, meaning no subcategories. Category names are in a single column. Optional codes column contains numerical values that uniquely identify each category. If you specify that the data file does contain codes (contains the category codes option in the Content Settings step), then a column containing unique codes for each category must exist in the cell directly to the left of category name. (In this example, column A).

This is an example of Compact format:

A	B	C	D	E
1	1	1	1 Technical Features	
2			_reliable	
3			_durably constructed	
4	2	2	10 Battery	any positive comment about long battery life
5			_long-lasting	
6	2	2	11 Storage Capacity	any positive comment about the amount that can be stored or memory capacity
7	2	2	12 Sound Quality	any positive comment about sound, quality or music quality
8	1	1	2 Comfort	
9	2	2	20 Ease of Use	any positive comment indicating that it is convenient, easy and user-friendly
10	2	2	21 Portability	any positive comment about mobility or indicating that it is handy and easy to transport
11	2	2	22 Size	any positive comment indicating that it is small or compact
12	2	2	23 Weight	any positive comment indicating that it is lightweight
13			_light	

The compact format is structured similarly to the flat list format except that it is used with hierarchical categories. Therefore, a code level column is required to define the hierarchical level of each category and subcategory.



The slide is titled "Demo 2" and discusses "Importing Predefined Categories from a Microsoft Excel File". It features the IBM logo in the top right corner and a copyright notice at the bottom right.

Demo 2

Importing Predefined Categories from
a Microsoft Excel File

© 2014 IBM Corporation

The following file(s) are used in this demo:

- Categorization_Techniques_demo2_start.str - a Modeler stream that reads a file containing evaluations of IBM SPSS Instructors gathered from customers who took IBM SPSS courses.
- Instructor Evaluations Coding Frame.xls - A Microsoft Excel file that contains pre-defined categories

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

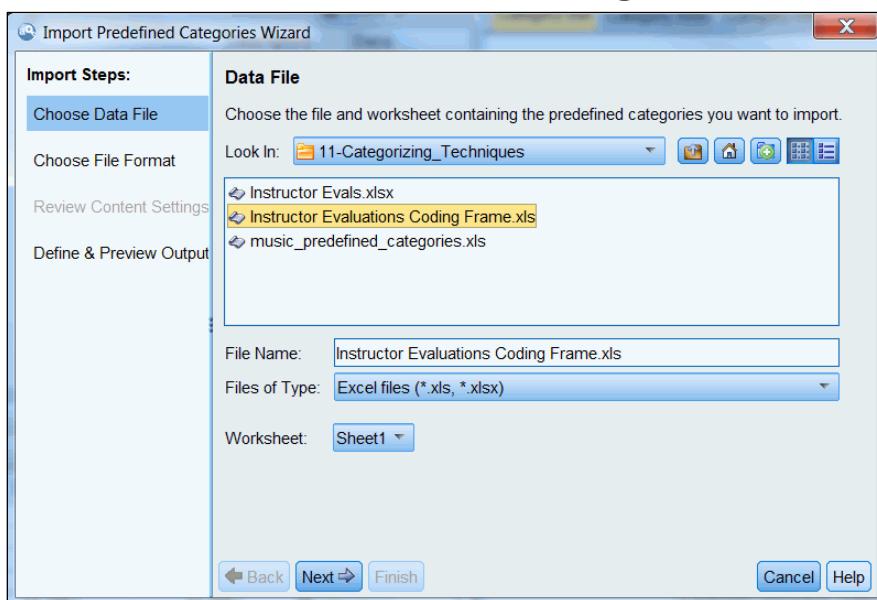
Demo 2: Importing Pre-Defined Categories from a Microsoft Excel File

Purpose:

You have a Microsoft Excel file that contains categories that you will use to analyze customer evaluations of IBM SPSS instructors. The categories were developed from prior evaluations that were gathered from customers who took IBM SPSS courses.

Task 1. Applying pre-existing categories to instructor evaluations.

1. From the **File** menu, click **Open Stream**.
2. Navigate to **C:\Train\0A105\11-Categorization_Techniques**, and then double-click **Categorization_Techniques_demo2_start.str**.
3. Edit the **Text Mining** node, and then click **Run**.
4. If necessary, select the **Categories and Concepts** view.
5. From the **Categories** menu, point to **Manage Categories**, and then click **Import Predefined Categories**.
6. Navigate to **C:\Train\0A105\11-Categorization_Techniques**.
7. Select **Instructor Evaluations Coding Frame.xls**.



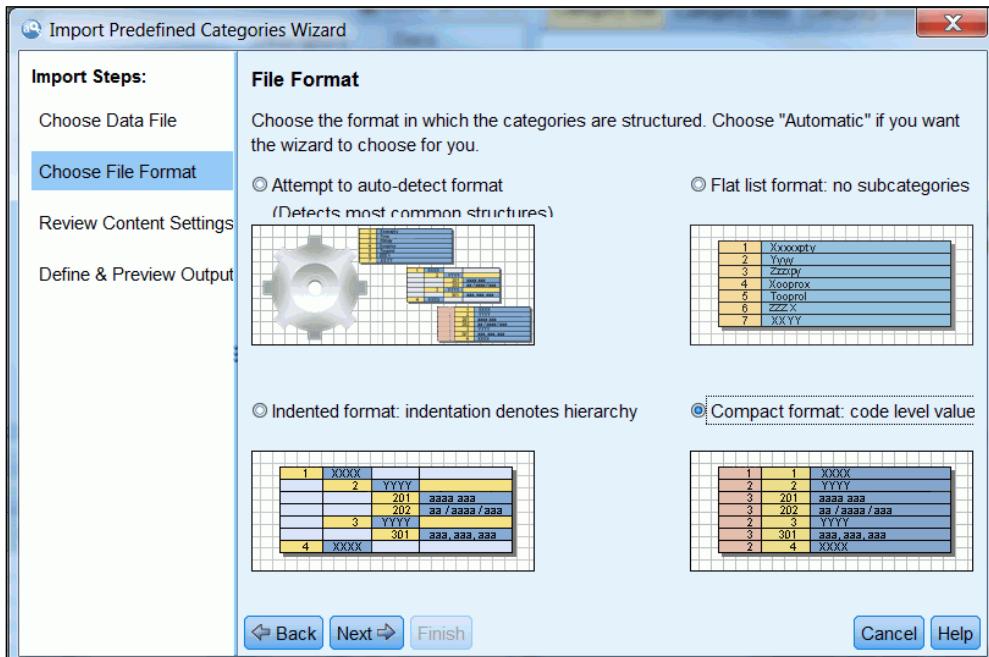
8. Click **Next**.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

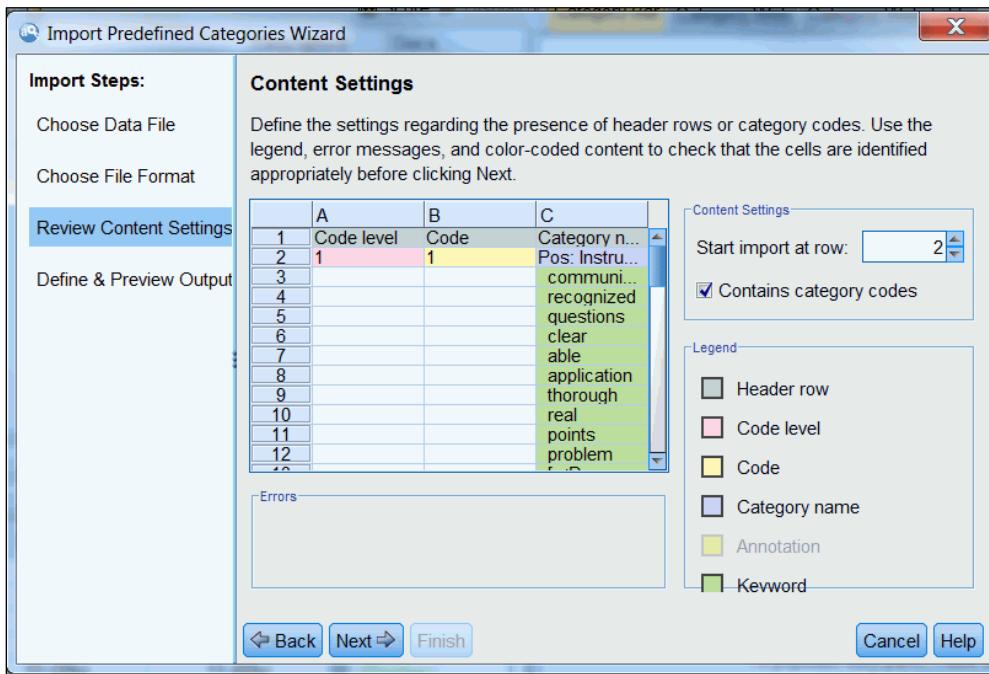
9. Select Compact format.



10. Click Next.

In this example the first row of the coding frame contains the column names.

11. Beside **Start Import at row** enter **2**.
12. Select **Contains category codes**.

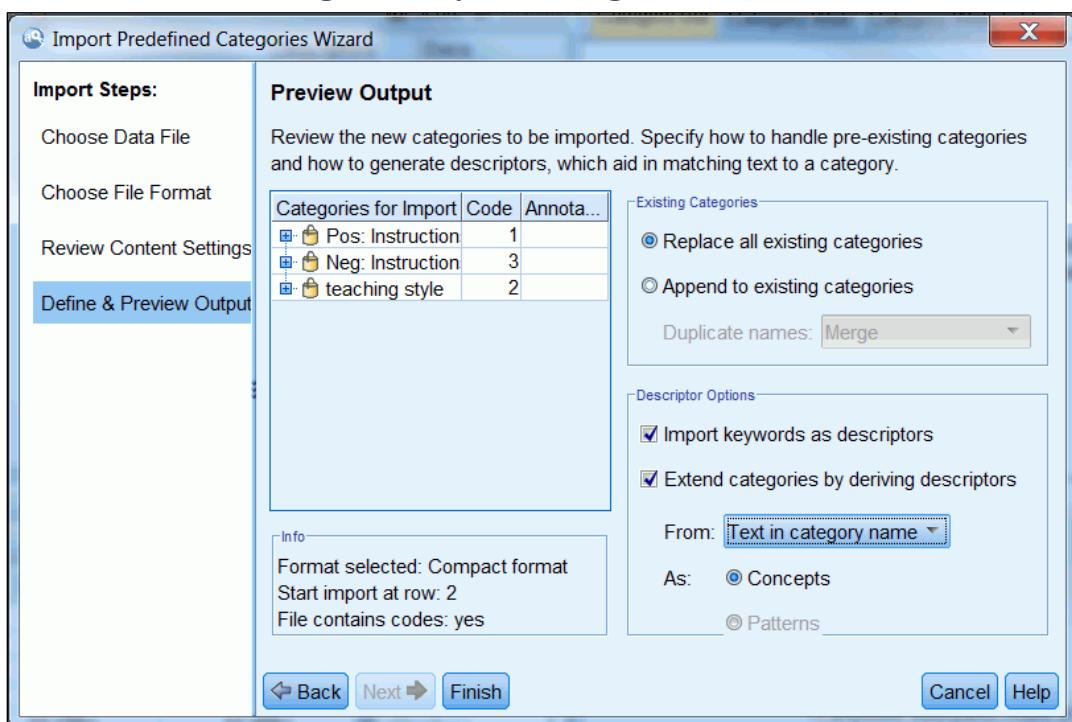


This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

13. Click Next.

In the last screen of the Wizard, the categories in the Code Frame are previewed. You also have the option to append or replace existing categories. In addition, you can import keywords as descriptors. Finally, categories can be extended by deriving descriptors by either using the category names, annotations, or both. The category name is scanned to see if words in the name match any extracted concepts or patterns. This option produces the best results when the category names are both long and descriptive. This is a quick method for generating category descriptors, which in turn enable the category to capture records that contain those descriptors.

14. Select Extend categories by deriving descriptors.



15. Click Finish.

16. Click **Score**.

Category	Descriptors	Docs
All Documents	-	188
Uncategorized	-	154
No concepts extracted	-	106
Pos: Instructions	28	23
teaching style	19	14
Neg: Instructions Extended	9	3

The code frame is read in, and categorization occurs based on the descriptors. Categories are populated using the methods to extend categories discussed above. The "Neg: Instructions" category was extended which means that additional descriptors were added during the categorization process.

The imported coding frame is not ideal for the instructor evaluations data, but this example does illustrate how to structure a coding frame and import it. After it is imported, assigning cases can be straightforward, if the category names are simple. If not, you will probably need to create rules and text matches to place cases correctly in categories.

- From the **File** menu, click **Close** and then click **Exit** to end the Interactive Workbench session.
- From the **File** menu, click **Close Stream**.
- From the **File** menu, click **New Stream**.

Do not close Modeler. Leave it open for the next demo.

Results:

You have successfully categorized data with a coding frame you imported from a Microsoft Excel spreadsheet.

Automated Classification

- When the business goal is to create a model and there are no TAPs, nor predefined categories available, it is always recommended to launch an automatic categorization process.
- You may not end up with the exact categories, but it is a good starting point.

© 2014 IBM Corporation



Besides using text analysis packages (TAPs, *.tap) with prebuilt category sets, you can also categorize your responses using any combination of the following methods:

- Automatic building techniques. Several linguistic-based and frequency-based category options are available to automatically build categories for you.
- Automatic extending techniques. Several linguistic techniques are available to extend existing categories by adding and enhancing descriptors so that they capture more records.
- Manual techniques. There are several manual methods, such as drag-and-drop.

In this section, the focus will be on automated techniques.

In general, categories can be made up of different kinds of descriptors (types, concepts, TLA patterns, category rules). When you build categories using the automated category building techniques, the resulting categories are named after a concept or concept pattern (depending on the input you select) and each contains a set of descriptors. These descriptors may be in the form of category rules or concepts and include all the related concepts discovered by the techniques.

Automated Classification Methods

- There are three different methods to create categories automatically:
 - based on frequency
 - based on types
 - based on type patterns

© 2014 IBM Corporation



The Settings dialog box lets you choose which inputs will be used to create and populate categories, whether you want to use linguistic or frequency based techniques, and when the user wants to see this dialog box. The categories are formed using descriptors derived from either type patterns or types.

If you select type patterns, categories are built from concept patterns belonging to the selected type patterns. They are not built from individual concepts. However, if a concept pattern contains only one concept (for example, excellent +), effectively the category can be built with the single concept excellent. If you select types instead of type patterns, the categories will be generated from the concepts belonging to the selected types.

There is a selection check box that is used to select the specific type patterns, or types, to be used in category building. You can also choose to select all or none of the types or type patterns. If you are creating categories from type patterns, the user has the additional ability to control which types will be used in creating the type patterns by making a selection in the Structure categories by type patterns box.

Automatic Categorization based on Frequency: Simply takes the list of extracted concepts and compares the number of times each concept occurs with the minimum number of terms to build their own category, set in advanced settings. By default, the minimum number is 15. Depending on the frequency of concepts extracted, you may increase or decrease this minimum. Any concept not caught by this minimum value will be grouped into a category called "Other". You can delete this category afterwards, or If you don't want this "Other" category to be created, just remove the string "Other" from the Advanced Settings: Frequencies window. Once this box is empty, the category will no longer be created and all concepts under the minimum value will remain as unused concepts.

This technique is most useful when your documents are:

- Market research-oriented and mainly contain lists of brands (who is your internet provider, where do you do grocery shopping)
- Your documents are rather regular lists of items (what is your favorite sport)

Keep in mind that each category will be made of a single concept; but once your set of categories is built and reviewed, you can extend those categories.

Automated Categorization Based on Types: If you select types, the categories will be built from the concepts belonging to the selected types. So if you select the <Budget> type in the table, categories such as cost or price could be produced since cost and price are concepts assigned to the <Budget> type. By default, only the types that capture the most records or documents are selected. This pre-selection allows you to quickly focus in on the most interesting types and avoid building uninteresting categories. The table displays the types in descending order starting with the one with the greatest number of records or documents (Doc. count). Types from the Opinions library are deselected by default in the types table.

Automated Categorization Based on Type Patterns: If you select type patterns, categories are built from concept patterns belonging to the selected type patterns. They are not built from individual concepts. However, if a concept pattern contains only one concept (for example, excellent + .), effectively the category can be built with the single concept "excellent. Usually you will want to delete such concepts.

To summarize, if you want to focus on the individual concepts, you should categorize based on types; if you want to focus more on patterns, you should choose type patterns. Even if you select types, you can still create categories that associate two or more concepts or types by requesting co-occurrence rules, or by creating your own rules.

Linguistic Categorization Techniques

- Concept root derivation
- Semantic network
- Concept inclusion
- Co-occurrence

© 2014 IBM Corporation



The three methods of linguistic categorization use the extracted concepts to produce categories. These techniques primarily work with nouns or noun phrases, excluding adjectives. They identify terms that are likely to have the same meaning, or are linguistically related (defined below for each method). By default, they are used together.

Linguistic Based Categorization-Concept Inclusion: The "concept inclusion" method uses algorithms to create categories by taking a term and finding other terms that include it. When determining inclusion, word order and the presence of such words as "in" or "of" are ignored. As an illustration, if you have the term "skill", term inclusion will group terms such as "computer skills" and "skill set" in a "skill" category. The root term used to create the category can have words before it, after it, or both before and after ("computer skill set"). As an example of how word order is ignored, the term "advanced Spanish course" will be included with "course in Spanish".

Concept inclusion may give better results when the responses contain a lot of domain-specific terminology or jargon. This is especially true if you have tuned the dictionaries beforehand so that the special terms are extracted and grouped appropriately (with synonyms).

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Linguistic Based Categorization--Semantic Networks: The semantic networks method creates categories using a semantic/lexical network based on WordNet, a linguistic project based at Princeton University. WordNet is a reference system of "English nouns, verbs, adjectives and adverbs...organized into synonym sets, each representing one underlying lexical concept." WordNet uses the word "synonym" in a broader sense than the usual meaning (information on WordNet can be found at <http://wordnet.princeton.edu> at the time of writing this course).

This technique begins by identifying extracted terms that are known synonyms (in the usual sense) and hyponyms. A hyponym is a word that is more specific than the category represented by a term. Thus, if "animal" is a term, "cat", "dog" and "kangaroo" are hyponyms of "animal"; that is, things that are examples of an animal.

Thus, if you extracted terms like "cat", "dog" and "kangaroo", a category containing all these terms could be created representing "animals". If you also had the terms of "oak tree" and "sunflower", a second category could be created representing "plants". These two categories could then be combined in another representing "living things".

Taken individually, many terms, especially single words, are ambiguous in meaning. For example, the term "buffet" can denote a type of meal or a piece of furniture. If a set of terms to be categorized includes "buffet", "meal" and "furniture", then the semantic network will have to choose between grouping "buffet" with one of the other two (because they are more generic). The choice made by the software may not be appropriate for the project, so results need to be reviewed carefully.

The semantic network is mainly restricted to synonym and hyponym relationships. The technique can also group terms representing place names in part/whole relationships (terms of type Location). Thus, the technique will group U.S. states into a category representing the United States.

The semantic network approach can yield better results than term inclusion with two types of data. First, when you expect to have terms that are related ("computer" and "hard drive"), and are interested in the relationships, the semantic network method is ideal. Second, when the open-ended responses are longer and contain more complex phrases, this method can often capture this information. The semantic network performs less well with highly technical or specialized terms.

Linguistic Based Categorization--Concept Root Derivation: The concept root derivation method groups terms by looking at the endings (suffixes) of each component in a term and finding other terms that have corresponding components with a related ending (suffix). It uses a set of linguistic derivation rules to accomplish this. For example, there is a rule that says that a term component ending with the suffix "ical" might be derived from a term having the same stem and ending with the suffix "ic." Using this rule, the algorithm would be able to group the terms "geological study" and "geologic studies".

Concept root derivation ignores function words (such as "in" and "of") and component order. Thus, by using another rule relating components ending in "y" to those ending in "ical," the algorithm would also be able to group the terms "studies in geology" and "geological studies". The set of component derivation rules has been chosen so that most of the terms grouped by this algorithm are synonyms. To increase completeness, there are some derivation rules that allow the algorithm to group terms that are situationally related. For example, the algorithm can group terms such as "career builder" and "career building".

You can use concept derivation on any sort of text. It produces fairly few categories, and each category tends to contain few terms. You may find it helpful to use this algorithm even if you are building categories manually; the synonyms it finds may be synonyms of those terms you are particularly interested in.

Linguistic Based Categorization--Co-occurrence: The co-occurrence rules method enables users to discover and group concepts that are strongly related within the set of records. The idea is that when concepts are often found together in records, their co-occurrence reflects an underlying relationship that is probably of value in the category definitions. For each co-occurrence discovered, one category is built. This method is not selected by default.

Additional Categorization Options

- There are a number of additional options for linguistic categorization. These options are listed in the dialog under the headings:
 - input and output
 - grouping techniques
 - other options

© 2014 IBM Corporation



Input and Output: Category input specifies from what the categories will be built. Category output specifies the general structure for the categories that will be built.

- Unused extraction results. This option enables categories to be built from extraction results that are not used in any existing categories. This minimizes the tendency for records to match multiple categories and limits the number of categories produced.
- All extraction results. This option enables categories to be built using any of the extraction results. This is most useful when no or few categories already exist. There will be some redundancy with this approach, and there will be more categories to sort through, but categorization will be richer.
- Hierarchical with subcategories. This option enables the creation of subcategories and sub-subcategories. You can set the depth of the categories by choosing the maximum number of levels (Maximum levels created field) that can be created.
- Flat categories (single level only). This option enables only one level of categories to be built, meaning that no subcategories will be generated.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

11-25

Grouping Techniques: In addition to selecting the grouping techniques, you can edit several other build options:

- Maximum search distance: The search distance value, which ranges from 1 to 4, tells the software how far, in relative terms, you want the techniques to search before producing categories. The lower the value, the fewer results you will get—however, these results will be less error-filled and are more likely to be significantly linked or associated with each other. The higher the value, the more results you might get—however, these results may be less reliable or relevant. This option is used with co-occurrences and semantic networks. The default setting of 3 is reasonably aggressive.
- Prevent pairing of specific concepts: Select this checkbox to stop the process from grouping or pairing two concepts together in the output.
- Generalize with wildcards where possible: Select this option to allow the product to generate generic rules in categories using the asterisk wildcard. This option has the advantage of reducing the number and simplifying category descriptors. Additionally, this option increases the ability to categorize more records using these categories on new text data.

Other Options: These are additional build options that you can modify:

- Maximum number of categories created: Use this option to limit the number of categories that can be generated. In some cases, you might get better results if this value is set higher and then any of the uninteresting categories are deleted.
- Minimum number of descriptors and/or subcategories per category: Use this option to define the minimum number of descriptors and subcategories a category must contain in order to be created. This option helps limit the creation of categories that do not capture a significant number of records.
- Allow descriptors to appear in more than one category: When selected, this option allows descriptors-concepts, types, patterns-to be used in more than one of the categories that will be built. This option is generally selected since items commonly or "naturally" fall into two or more categories and allowing them to do so usually leads to higher quality categories. If you do not select this option, you reduce the overlap of records in multiple categories and depending on the type of data you have, this might be desirable.

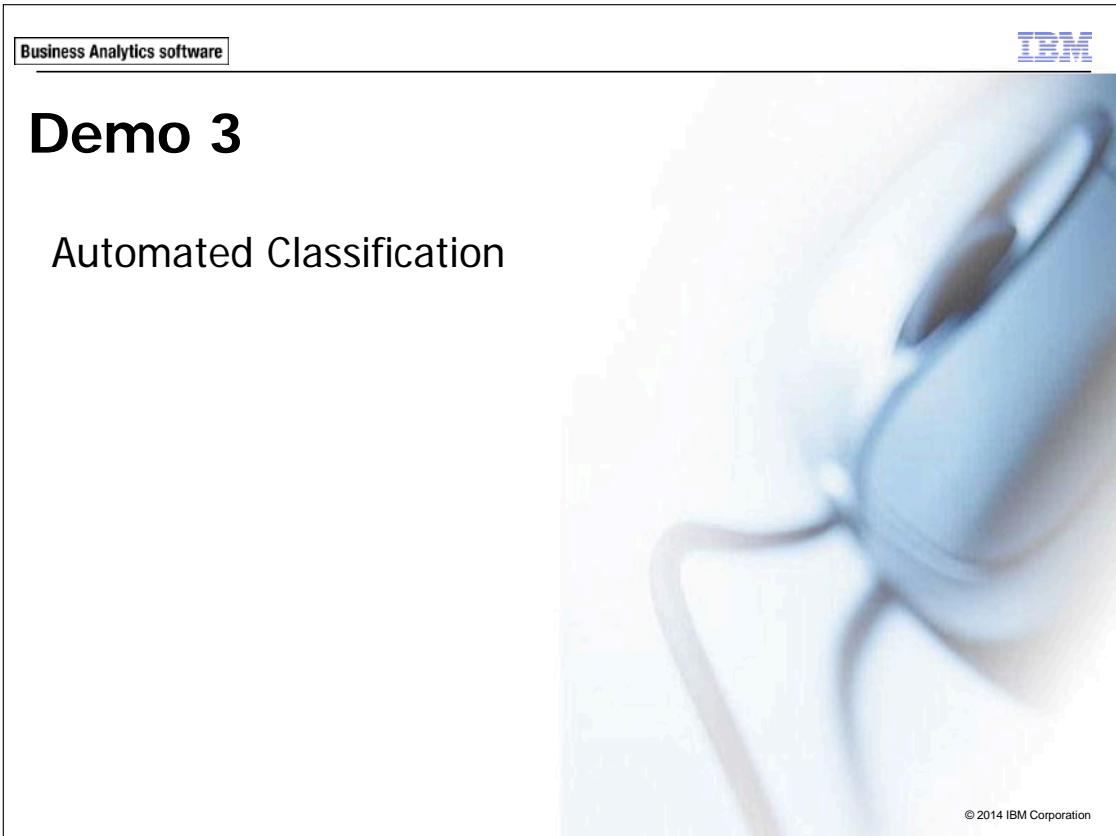
- Resolve duplicate category names by: Select how to handle any new categories or subcategories whose names would be the same as existing categories. You can either merge the new ones (and their descriptors) with the existing categories with the same name or you can choose to skip the creation of any categories if a duplicate name is found in the existing categories.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

11-27



The slide features a large, abstract blue and white background image of what appears to be a stylized flower or perhaps a brain-like structure. In the top left corner, there is a small rectangular box containing the text "Business Analytics software". In the top right corner, the "IBM" logo is displayed. The main title "Demo 3" is centered at the top in a large, bold, black font. Below it, the subtitle "Automated Classification" is also centered in a smaller, regular black font. At the bottom right of the slide, there is a small, faint copyright notice: "© 2014 IBM Corporation".

The following file(s) are used in this demo:

- Categorization_Techniques_demo3_start.str - a Modeler stream that reads a file containing call center data for March and April

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

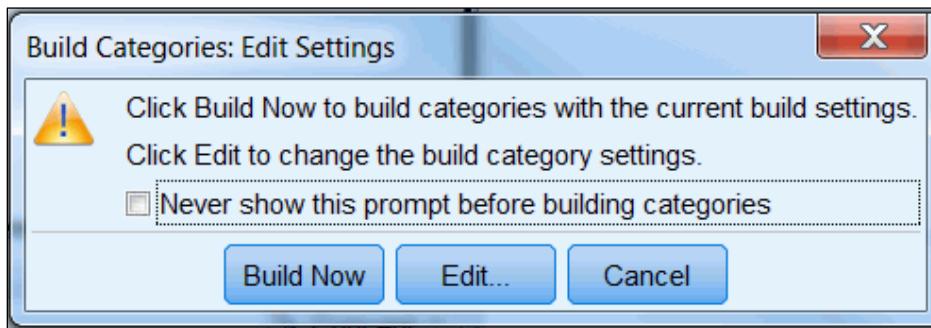
Demo 3: Automated Classification

Purpose:

You want to use automated techniques to create categories when predefined categories are not available in order to help predict which Astroserve customers are likely to churn.

Task 1. Requesting an automated classification technique.

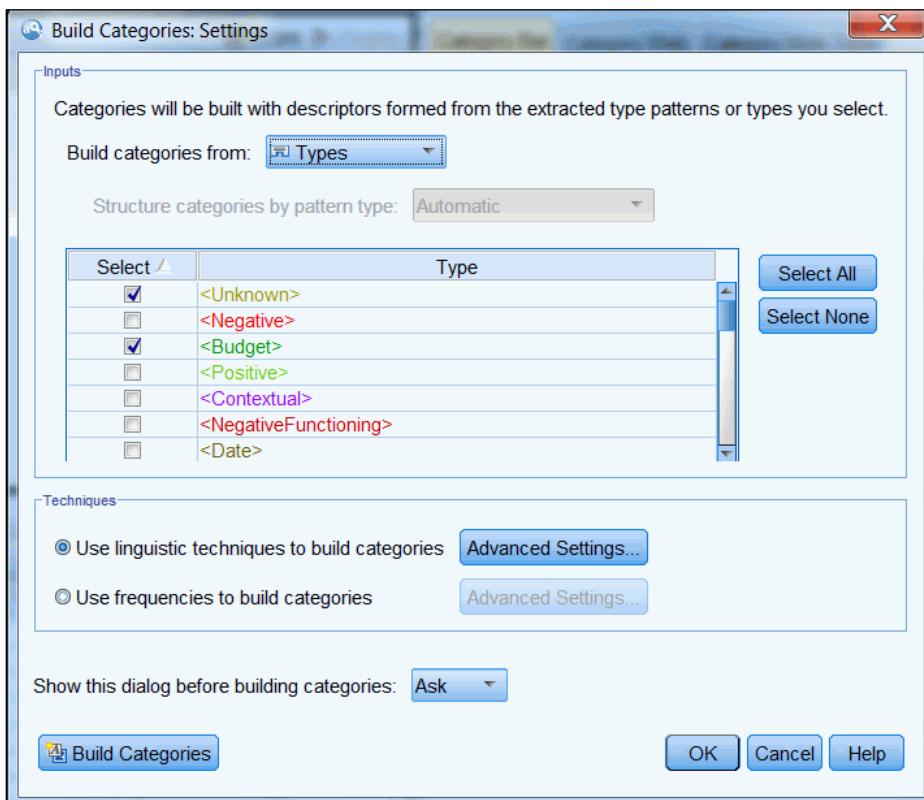
1. From the **File** menu, click **Open Stream**.
2. Navigate to **C:\Train\0A105\11-Categorization_Techniques**, and then double-click **Categorization_Techniques_demo3_start.str**.
3. Edit the **Text Mining** node.
4. Click **Run**.
5. Switch to the **Categories and Concepts** view.
6. In the left corner of the **Categories** pane, click **Build**.



The Build Now button uses the default settings. Although these settings are often sufficient to begin the categorization process, you will select Edit to review and modify the build settings.

7. Click **Edit**.

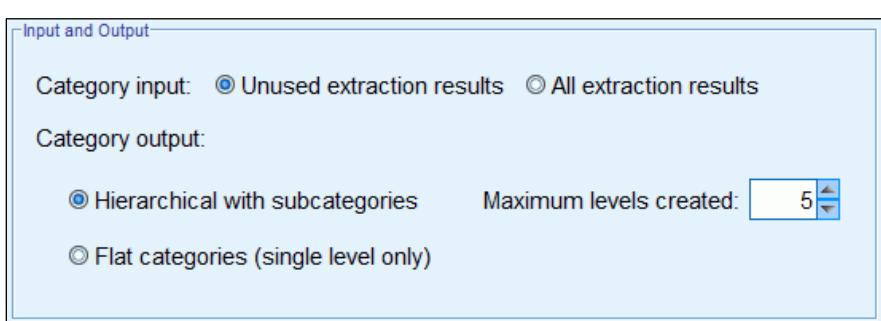
The results appear as follows:



You will start by building categories from Types, which is the default.

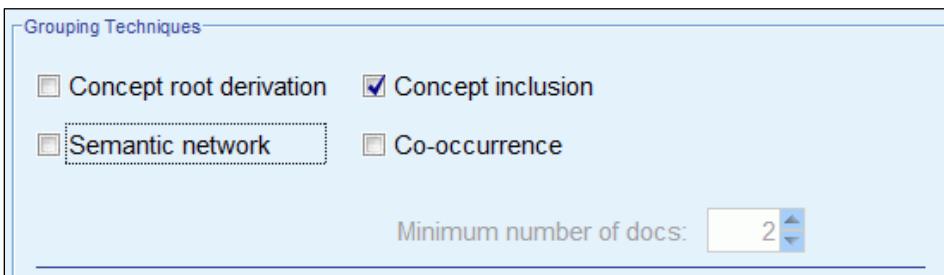
8. In the **Inputs** section, click **Select All** so that all Types will be used to build categories.
9. In the **Techniques** section, click **Advanced Settings** to select the Linguistic method you want.
10. Ensure that **Hierarchical with subcategories** is selected in the Category output section.

This is the default.



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

11. Clear the **Concept root derivation** and **Semantic network** check boxes, leaving only **Concept inclusion** selected.



12. Click **OK**.

13. Click **Build Categories**.

Category	Descriptors	Docs
All Documents	-	1269
Uncategorized	-	171
No concepts extracted	-	2
service	147	431
number	95	204
on line	90	145
bill	69	182
mobile	63	159
phone	53	158
account	49	146
charge	47	188
credit	44	80
time	40	141

In this window you see the names of the categories.

14. Click **Score** to see how many records are in each category.

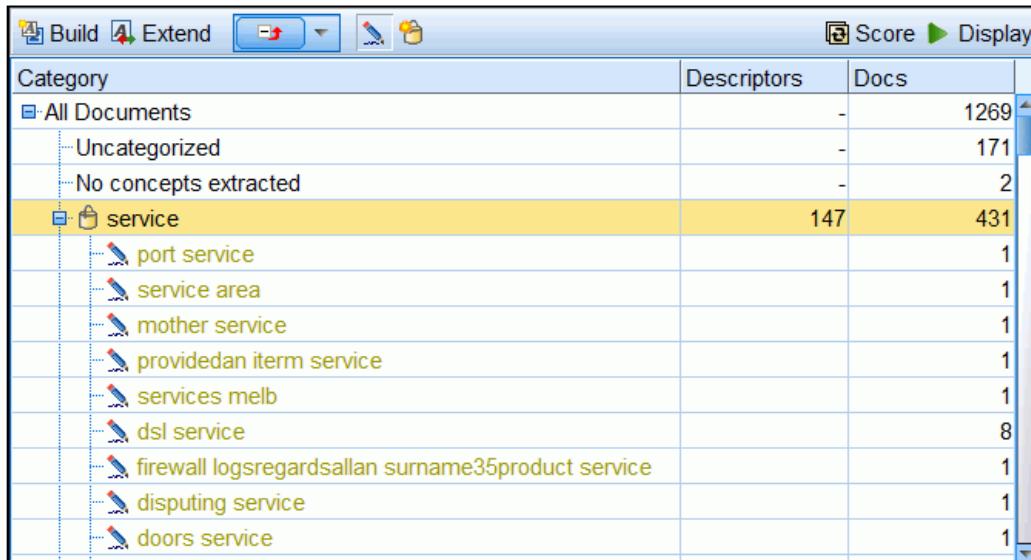
Category	Descriptors	Docs
All Documents	-	1269
Uncategorized	-	171
No concepts extracted	-	2
service	147	431
number	95	204
on line	90	145
bill	69	182
mobile	63	159
phone	53	158
account	49	146
charge	47	188
credit	44	80
time	40	141

Task 2. Examining the results.

The Concept Inclusion method appears to have done a good job categorizing the records, considering that only 171 of the records were not categorized. It also did a good job identifying a number of the issues customers were calling about, for example, "service", "mobile", "charge", and so on. It is certain that Astroserve would be interested in finding out what customers are saying about these sorts of topics.

Of course, one drawback to these automated categorization methods is that they tend to create a number of not very interesting categories as well. In this example, Concept Inclusion created categories for "time" and "number" because those concepts both occurred so often in the data. While no doubt that is true, neither category looks particularly useful. Keep in mind that your ultimate goal is to predict which Astroserve customers are likely to churn. It seems unlikely that categories like "time" or "number" will make very good predictors.

1. Beside **service**, click the **expand**  button to view a list of the 147 descriptors in the category.



The screenshot shows the IBM SPSS Text Analytics interface. At the top, there are tabs for Build, Extend, Score, and Display. Below the tabs is a toolbar with icons for search, edit, and file operations. The main area is a tree view of categories. The 'service' category is expanded, showing its sub-descriptors. The 'Category' column lists the category names, and the 'Descriptors' and 'Docs' columns show the count of descriptors and documents respectively. The 'service' category has 147 descriptors and 431 documents. Some of the descriptors listed under 'service' include 'port service', 'service area', 'mother service', 'providedan item service', 'services melb', 'dsl service', 'firewall logsregardsallan surname35product service', 'disputing service', and 'doors service'.

Category	Descriptors	Docs
All Documents	-	1269
Uncategorized	-	171
No concepts extracted	-	2
service	147	431
port service	1	
service area	1	
mother service	1	
providedan item service	1	
services melb	1	
dsl service	8	
firewall logsregardsallan surname35product service	1	
disputing service	1	
doors service	1	

Concept inclusion has grouped together all the concepts that include the word "service" in this category.

2. Scroll down to view the **service** categories that have sub-categories, and then expand **mobile service**.

The screenshot shows a software interface for managing categories. At the top, there are menu options: Build, Extend, Score, and Display. The main area is a table with three columns: Category, Descriptors, and Docs. The Category column displays a hierarchical tree. The first level includes 'Service guarantee' (with 5 descriptors and 1 doc), 'mobile service' (with 4 descriptors and 16 docs), 'service provider' (with 3 descriptors and 3 docs), and 'connect service' (with 3 descriptors and 5 docs). The 'mobile service' category is expanded, showing its sub-categories: 'iterim mobile service', 'deactivation of mobile service', 'mobile service' (which has 13 sub-docs), and 'provision of a mobile service'. The 'mobile service' entry under sub-categories also has 4 descriptors and 16 docs.

Category	Descriptors	Docs
Service guarantee		1
USD100 service guarantee		1
USD50 service guarantee		1
USD40 service guarantee		1
reconnect service guarantee		1
requests service guarantee		1
mobile service	4	16
iterim mobile service		1
deactivation of mobile service		1
mobile service		13
provision of a mobile service		1
service provider	3	3
connect service	3	5

Some people find sub-categories quite useful. For example, if Astroserve management found out that customers were complaining about the "services" they were receiving, they might ask whether customers were complaining about services in general, or just some of them. The sub-categories help to narrow it down. For instance, the "mobile service" category differentiates between the types of "mobile service" the customer was calling about. On the other hand, you may find it perfect just to capture all mentions of the word "service" when they called. In that case, you may be just happy with a single (flattened) category called "service".

Now, look at what the categories look like when they are created from type patterns instead of types.

3. Scroll to the top of the **Categories** pane.
4. Click **All Documents**.

The screenshot shows a simplified list of categories in the 'Categories' pane. The categories listed are 'All Documents' (selected and highlighted in yellow), 'Uncategorized', 'No concepts extracted', 'service' (with a folder icon), and 'port service' (with a gear icon).

Category
All Documents
Uncategorized
No concepts extracted
service
port service

5. From the **Edit** menu, click **Select All**.

6. Press **Delete** to delete all the categories.

Otherwise when you create the new ones based on type patterns they will be appended to bottom of the current list. It will be easier to see how type patterns work if you start out with a blank slate.

7. Click **Build**, and then click **Edit**.
8. From the **Build Categories from** list, select **Type Patterns**.
9. Click **Build Categories**.
10. Click **Score**.
11. Expand the **service** category.
12. Expand the **service + <>** category.

The screenshot shows the IBM SPSS Text Analytics interface. At the top, there are tabs for 'Build' (selected), 'Extend', and other tools. Below the tabs is a toolbar with icons for search, refresh, and file operations. To the right of the toolbar are buttons for 'Score' and 'Display'. The main area is divided into two sections: a tree view on the left and a table on the right.

Category Tree:

- All Documents
 - Uncategorized
 - No concepts extracted
 - service
 - fx [service + free]
 - fx [service + not free]
 - fx [service + comfortable]
 - service+<>
 - fx [barr service + .]
 - fx [domestic service + .]
 - fx [service rep + .]
 - fx [premium service + .]

Table:

Category	Descriptors	Docs
All Documents	-	1269
Uncategorized	-	212
No concepts extracted	-	2
service	191	427
fx [service + free]	1	
fx [service + not free]	1	
fx [service + comfortable]	1	
service+<>	116	295
fx [barr service + .]	1	
fx [domestic service + .]	1	
fx [service rep + .]	13	
fx [premium service + .]	1	

In this example:

- Some descriptors are concepts (under service): *fx*[barr service +], *fx*[service rep +] (it is recommended that you delete such concepts).
 - Some sub-categories group a concept and a type: service+<>, service+<Negative>.
13. Scroll down until you get to the **service + <Negative>** category.

14. Expand the **service + <Negative>** category.

Category	Descriptors	Docs
pstn service+<>	2	3
service + <Negative>	32	117
fx [changed number information service + not offered]		1
fx [result in fees from this service + unable]		1
fx [security of wap service + problem]		1

In this example:

- Some descriptors are TLA-based business rules (under **service + <Negative>**):
fx [changed number information service + not offered]

Task 3. Categorizing with frequency based classification.

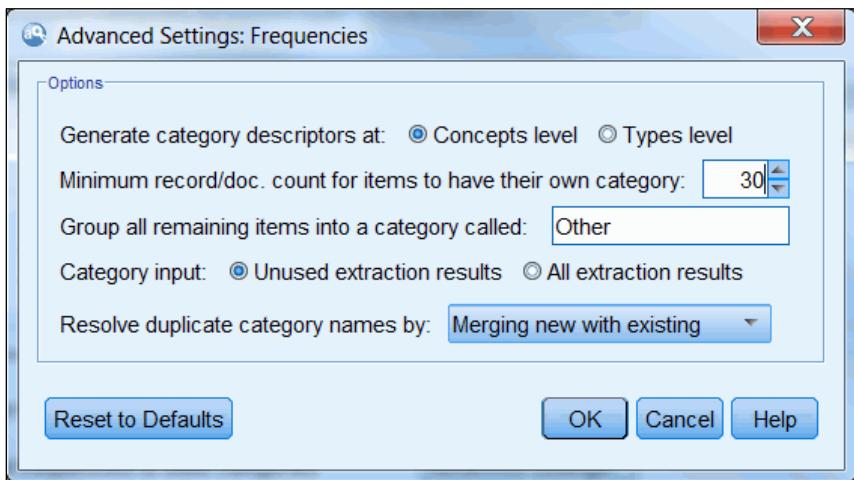
You could try additional linguistic techniques but you can try those in the workshop. Instead, at this point you will try frequency based autoclassification.

1. Scroll to the top of the **Categories** window.
2. Click **All Documents**.

Category
All Documents
Uncategorized
No concepts extracted
service

3. From the **Edit** menu, click **Select All**.
4. Press **Delete** to delete all the categories.
5. Click **Build**, and then click **Edit**.
6. Beside **Build Categories** from, select **Types**.
7. Select **Use frequencies to build categories**, and then click **Advanced Settings**.
8. Ensure that **Concepts level** is selected.

9. In the **Minimum number of record/doc. count for items to have their own category** box, enter **30**.



10. Click **OK**, and then click **Build Categories**.

11. Click **Score**.

Category	Descriptors	Docs
All Documents	-	1269
Uncategorized	-	2
No concepts extracted	-	2
Other	5764	1252
service	1	296
complaint	1	296
fault	1	229
problem	1	227
cruel	1	171
line	1	157
charge	1	155
not satisfied	1	139
not working	1	131

It appears that the categorization was incredibly successful because almost no records were uncategorized. However, this is misleading because there is an "Other" category which groups all the unused concepts, and that category has 1252 records.

12. Delete the **Other** category, and then click **Score**.

Category	Descriptors	Docs
All Documents	-	1269
Uncategorized	-	59
No concepts extracted	-	2
service	1	296
complaint	1	296
fault	1	229
problem	1	227
cruel	1	171
line	1	157
charge	1	155
not satisfied	1	139
not working	1	131
new	1	129

Now notice that there are 59 records uncategorized. Many of the same simple categories were created with Concept Inclusion, such as "line", "service", and "phone". All are certain categories that will be needed when the data are categorized because all of them are certainly things Astroserve needs to pay attention to.

13. From the **File** menu, click **Close**, and then click **Exit** to end the Interactive Workbench session.
14. From the **File** menu, click **Exit**, and then click **Exit** to end the Modeler session.

Results:

You have successfully categorized data using the Concept Inclusions and Frequency automated categorization techniques.

Apply Your Knowledge

Purpose:

Test your knowledge of the material covered in this module.

Question 1: True or False: Sentiments about specific objects, things, or persons can be captured by categories built from types.

- A. True
- B. False

Question 2: True or False: The frequency-based categorization methods are based on either the concepts or types that were extracted.

- A. True
- B. False

Question 3: True or False: Pre-existing categories can be copied and pasted from a Microsoft Word document into the Text Analytics Categories pane.

- A. True
- B. False

Question 4: True or False: The Concept Root Derivation method should be used to group terms like biology, biological, biology studies into a category.

- A. True
- B. False

Question 5: True or False: Semantic Networks should not be used to group terms like Chicago, New York, Los Angeles, and Denver into a category called Cities.

- A. True
- B. False

Apply your Knowledge - Solutions

- Answer 1: B. False. Type patterns would have to be used to identify sentiment.
- Answer 2: A. True. Frequency categories can be based on either types or concepts.
- Answer 3: B. False. Pre-existing categories can only be imported from Microsoft Excel files.
- Answer 4: A. True.
- Answer 5: B. False. Semantic networks are designed to identify hyponyms. The city names are all qualify as hyponyms of the term City.

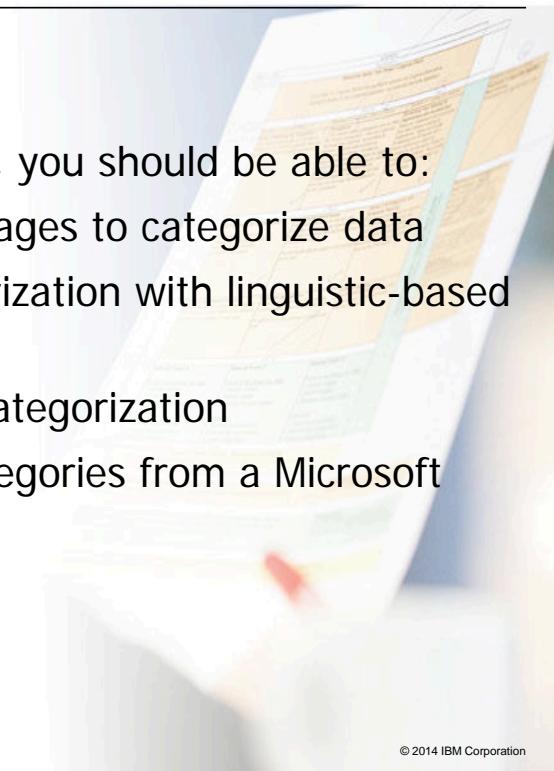
Business Analytics software

IBM

Summary

- At the end of this module, you should be able to:
 - use Text Analysis Packages to categorize data
 - use automated categorization with linguistic-based techniques
 - use frequency based categorization
 - import pre-existing categories from a Microsoft Excel file

© 2014 IBM Corporation



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

11-40

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Workshop 1

Importing Predefined Categories



© 2014 IBM Corporation

The following files will be used:

- Music_Survey with Text Link Analysis.str - a Modeler stream that reads from a file containing customer likes and dislikes about a portable music player
- music_predefined_categories.xls - a Microsoft Excel file that contains the categories you will be importing into the analysis

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Workshop 1: Importing Pre-defined Categories

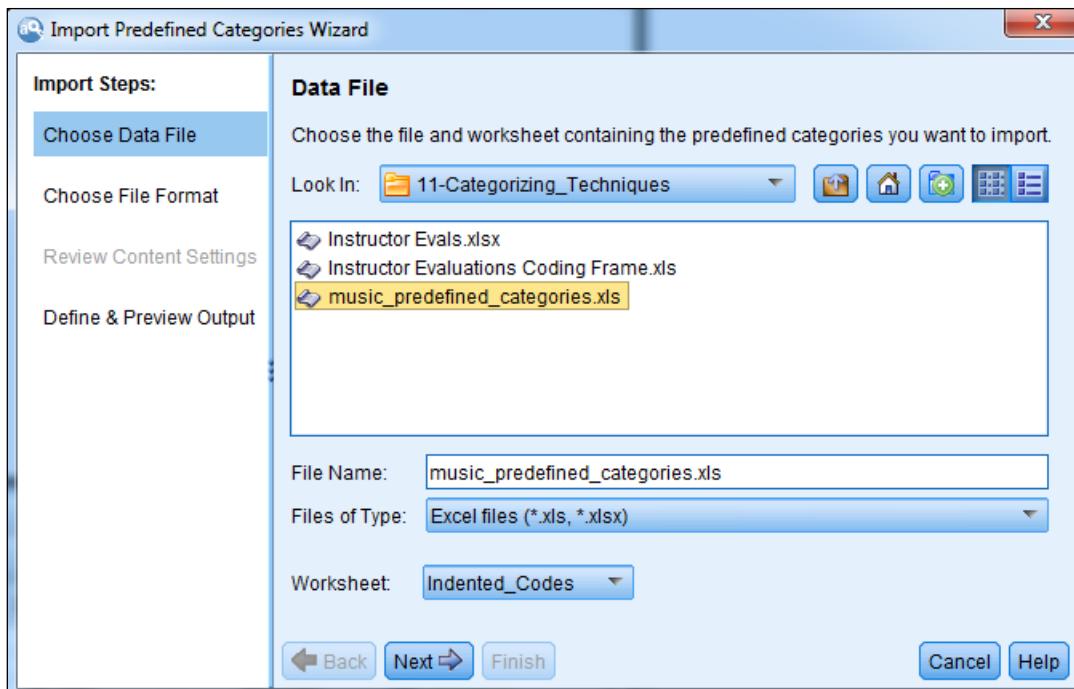
You have conducted several surveys in the past on portable music players so you already have a good idea of what most of the categories are and have them stored in a Microsoft Excel file. Your goal is to use this file to categorize some new data in Text Analytics for Modeler.

- Open C:\Train\0A105\11-Categorizing_Techniques\ Music_survey with Text Link Analysis.str.
- Run the Text Mining node.
- Switch to the Categories and Concepts view.
- Import categories from the file music_predefined_categories.xls.
- Select Indented Codes from the Worksheet menu.
- Select Attempt to auto-detect format on the File Format dialog.
- Check Extend categories by deriving descriptors in the Descriptor Options section.
- Beside From, select Text in both in the Descriptor Options section.
- Score the categories.
- End the interactive session.
- Close the stream, and then create a new stream (do not Exit from Modeler).

Workshop 1: Tasks and Results

- Open C:\Train\0A105\11-Categorizing_Techniques\Music_survey with Text Link Analysis.str.
- Run the **Text Mining** node.
- Switch to the **Categories and Concepts** view.
- Import Predefined Categories:
 - C:\Train\0A105\11-Categorization_Techniques\music_predefined_categories.xls.

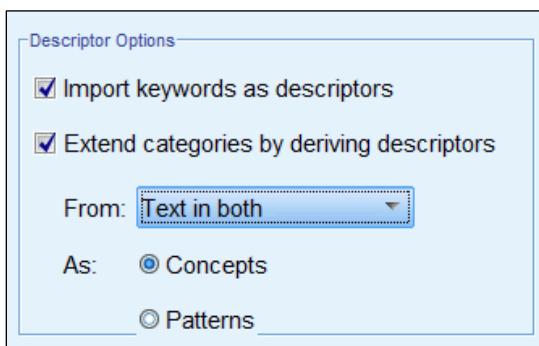
A section of the result appears as follows:



- Beside **Worksheet**, ensure that **Indented_Codes** is selected, and then click **Next**.
- From the **File Format** dialog, ensure that **Attempt to auto-detect format** is selected, and then click **Next**.
- In the **Descriptor Options** section, select **Extend categories by deriving descriptors**.

- Beside **From**, select **Text in both**.

A section of the result appears as follows:



- Click **Finish** to close the Import Predefined Categories Wizard, and score the categories.

A section of the result appears as follows:

Category	Descriptors	Docs
All Documents	-	405
Uncategorized	-	78
No concepts extracted	-	4
Comfort Extended	16	196
Technical Features Extended	14	127
Usage	9	95
Appearance Extended	9	44
Price Extended	4	12

- Exit the interactive session.
- Close the stream and leave Modeler open for the next workshop.

Workshop 2

Using Autoclassification Techniques to Categorize Data



© 2014 IBM Corporation

The following file will be used:

- Music_Survey with Text Link Analysis.str - a Modeler stream that reads from a file containing customer likes and dislikes about a portable music player

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

11-45

Workshop 2: Use Autoclassification Techniques to Categorize Data

The goal of the exercises is to see which categorization techniques work best with the music survey data. In this exercise there are no right or wrong answers, the main point here is to get familiar with the categorization techniques. Remember that the purpose of this study is to analyze what factors respondents use to rate portable music players.

- Open C:\Train\0A105\11-Categorizing_Techniques\Music_survey with Text Link Analysis.str.
- Run the Text Mining node.
- Switch to the Categories and Concepts view.
- Categorize the data with the concept inclusion technique only. Be sure to build the categories from Types and to select all types.
How well did this categorization technique perform?
Are there many categories you might use?
- Next try the semantic network method.
How would you compare its performance to the previous method?
Are there categories you might want to retain?
- End the interactive session, and then exit Modeler.

Workshop 2: Tasks and Results

- Open C:\Train\0A105\11-Categorizing_Techniques\Music_survey with Text Link Analysis.str.
- Run the **Text Mining** node.
- Switch to the **Categories and Concepts** view.
- Categorize the data with the **concept inclusion** technique only, using all types:
 - Click **Build**, and then click **Edit**.
 - Beside **Build categories from**, ensure that **Types** is selected.
 - Click **Select All** to select all types.
 - Beside **Use linguistic techniques to build categories**, click **Advanced Settings**.
 - In the **Grouping techniques** section, deselect **Concept root derivation** and **Semantic network**, leaving only the **Concept inclusion** option enabled.
 - Click **OK**, and then click **Build Categories**.
 - Click **Score**.

A section of the result appears as follows:

Category	Description	Docs
All Documents	-	405
Uncategorized	-	134
No concepts extracted	-	4
music	25	74
easy	10	62
color	5	10
radio	5	9
feature	5	6
storage	4	19
sound	4	36
tunes	4	7
memory	4	6
photo	4	5
well	4	8
size	3	38

The method performed reasonably well, although it may be concerning that only $405 - 134 = 271$ (66.9%) of the documents were categorized. Still there are certainly a number of potentially useful categories, such as "storage", "sound", "size", and "color", that could be used to gauge customer likes and dislikes of the portable music player.

- To try the **semantic network** method:
 - Select all of the documents, and then click **Delete**.
 - Click **Build**, and then click **Edit**.
 - Beside **Use linguistic techniques to build categories**, click **Advanced Settings**.
 - In the **Grouping techniques** section, deselect **Concept inclusion**, and enable the **Semantic network** option.
 - Click **OK** to return to the main dialog, and then click **Build Categories**.
 - Click **Score**.

A section of the result appears as follows:

Category	Descript...	Docs
All Documents	-	405
Uncategorized	-	280
No concepts extracted	-	4
memory device	5	20
electronics	4	51
hardware	3	4
computers	2	2
computer network	2	4
occupation	2	3
place of business	2	7
sports by type	2	2
music	2	52
outdoor activities	2	2
lighter	2	3
screen	2	2

This technique only categorized $405 - 280 = 125$ documents (30.8%), so it does not appear that it is very appropriate for this data set. In addition, on the surface there only appears to be a few categories that may be worth retaining ("screen", "light", and "memory device").

- Exit the Interactive Workbench session, and Modeler.



Creating Categories

IBM SPSS Modeler Text Analytics (v16)

Business Analytics software



© 2014 IBM Corporation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Objectives

- At the end of this module, you should be able to:
 - create categories using various categorization techniques
 - discuss strategies for categorization
 - use linguistic based categorization techniques
 - visualize relationships among categories
 - use conditional rules to create categories
 - extend categories
 - create text analysis packages

© 2014 IBM Corporation

There are a number of techniques you can use to categorize text data: you can use automated categorization techniques such as Concept Inclusion or Semantic Networks; you can import predefined categories from a Microsoft Excel spreadsheet; you can use a Text Analysis Package; or you can manually categorize your data. The intent of this module is not to review the different types of categorization in depth. Instead, the focus will be on how to categorize a specific set of data, in this case call center data collected from Astroserve customers.

Because the categories are unknown, it was decided that the automated classifications techniques are the best approach for categorizing this particular set of data. If the categories were known, you may have considered a different approach, such as to import the categories from a Microsoft Excel spreadsheet. The categorization you use will depend on the type of data you have and whether or not you already know your categories. Thus, while not all of the techniques demonstrated in this module will be appropriate for your set of data, it is hoped presenting you with a real-life example of categorizing a set of data, will in the end provide you with ideas and strategies that you will find useful in categorizing your own data.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

12-3

Which Technique Should You Use?

- Categories can be built in different ways:
 - already defined in a Text Analysis Package (TAP)
 - already defined in a code frame to be imported
 - automatic classification
 - manually

© 2014 IBM Corporation



There are many ways you can build your categories in Text Analytics for Modeler and sometimes it difficult to decide on which one to use. Of course, the answer is fairly easy if you already have categories, in which case you will probably lean toward importing categories from Microsoft Excel, build them manually, or using a Text Analysis Package if you have one. If this is not the case, you will need to use the automatic classification techniques that come with the software.

If you decide to use automatic classification, it is recommended that you try all of the methods to see which one is best suited for your data. As you have seen already in an earlier module, all of these techniques produce a lot of categories, and it is up to you decide if the categories they create will be useful in your analysis. For example, if your goal is to predict which customers are likely to churn, you would evaluate each category based on whether or not you considered it relevant. In the end, you would probably keep only the categories that you thought made sense and throw out the rest.

There are several contenders for best method, but without specific goals in mind for the text analysis, it is hard to choose the best method, except on the simple criteria of fewest records left uncategorized. In this module, you will primarily focus on automatic classification techniques to categorize the data.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Using Automated Classification

- Which method is best for your data?
- Is it better to use each method simultaneously, or one at a time?

© 2014 IBM Corporation



Categories can be created with linguistic methods, with co-occurrence rules, from the most frequent types, and by manually creating a category (this is in addition to categories created from TLA). Is there any best practice for how categories should be created? Should one of these methods be used first? Should all the linguistic methods be used simultaneously or instead sequentially?

Although there is no methodology that is ideal for every type of data and text mining project, here is some guidance about strategy:

- When a concept is used in a category, it is no longer available for further categorization. As a consequence, using the linguistic methods one at a time, as compared to using all three together, will yield different results. Using the methods one at a time lets the method used earlier utilize more concepts than those used later. Therefore, if you use the methods one at a time, it is recommended that you apply the method most applicable to the data first. Alternatively, you can change the order of the methods and compare the results.

- Co-occurrence and conditional rules should generally be used after you have done TLA and created categories from those patterns. These rules are a simpler method to create rules for concepts that occur together, but they are not a substitute for TLA patterns.
- As you categorize the data, go back and modify the linguistic resources as you see other changes that can be made. The categories that you create, including any done manually, will be retained and can be rescored after the text has been re-extracted.
- There is no absolutely correct set of categories that can be constructed for a particular data set, except for small and uncomplicated text data. That said, a particular set of categories can be more, or less, effective for a specific data-mining project, but two analysts will often create different categories from the same set of concepts.
- It is usually better to initially create more, rather than fewer, categories. It is easy to delete and combine categories. However, too many categories can make the job of reviewing the categories time-consuming, so a balance must be struck. That is why the default for the software is to create only 30 categories.
- Naming categories appropriately can be important for the understanding and use of a model's results by others in an organization (the same is true for naming factors and clusters in those statistical techniques). If a category name does not apply, then rename it.
- There is a law of diminishing returns to creating categories when you begin with thousands of concepts. You do not need to include every concept in a category.

Follow this advice by using the most applicable linguistic method first, and using the default number of categories.

Business Analytics software

IBM

Demo 1

Categorizing Astroserve Call Center Data



© 2014 IBM Corporation

The following file(s) are used in this demo:

- Creating Categories_demo1_start.str - a Modeler stream that reads a file containing call center data for March and April

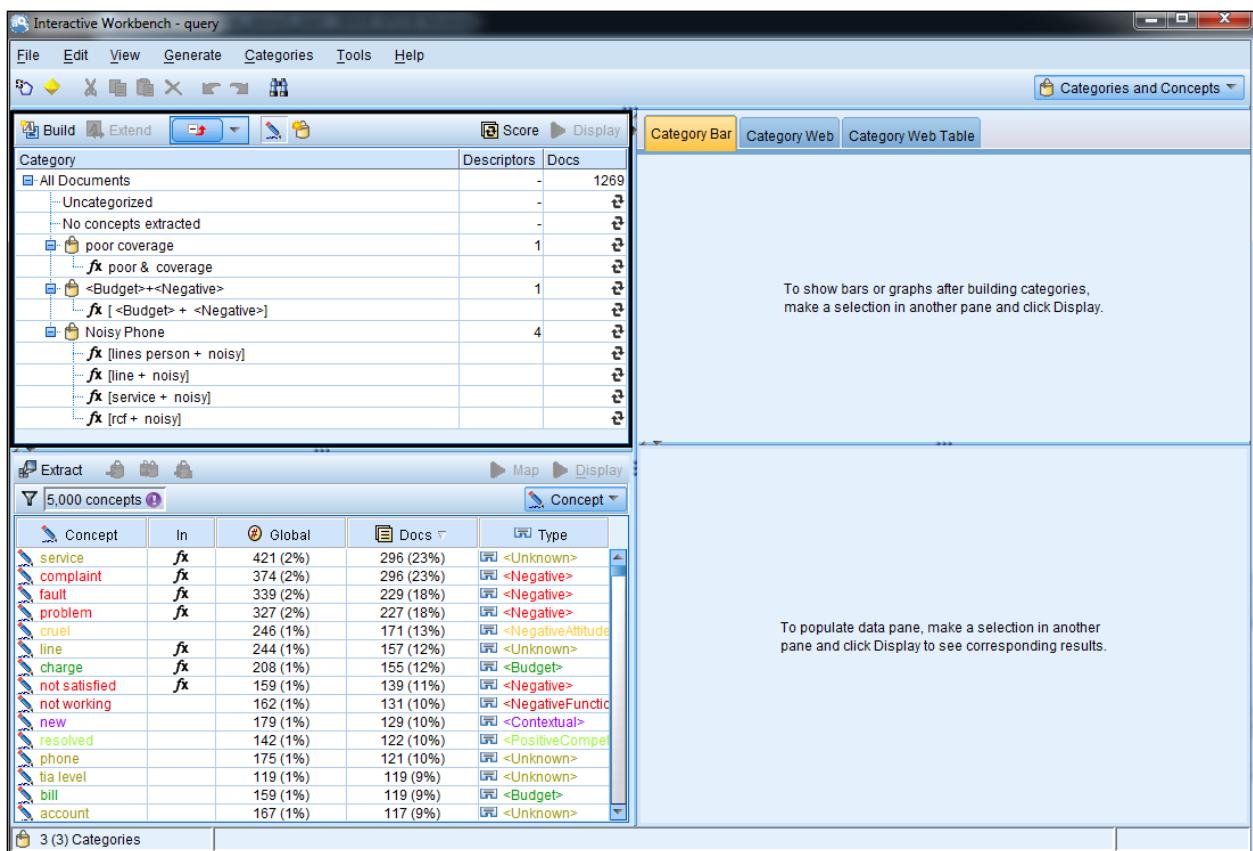
Demo 1: Categorizing Astroserve Call Center Data

Purpose:

After examining the call center data and fine tuning the dictionary resources so that they correctly extract the information that you need to analyze the data, the next step is to categorize the data. Because you do not already have pre-defined categories, you have decided to use auto classification techniques to create the categories.

Task 1. Review the data.

1. From the **File** menu, click **Open Stream**.
2. Navigate to **C:\Train\0A105\12-Creating_Categories**, and then double-click **Creating Categories_demo1_start.str**.
3. Run the **Text Mining** modeling node.
4. Switch to the **Categories and Concepts** view.



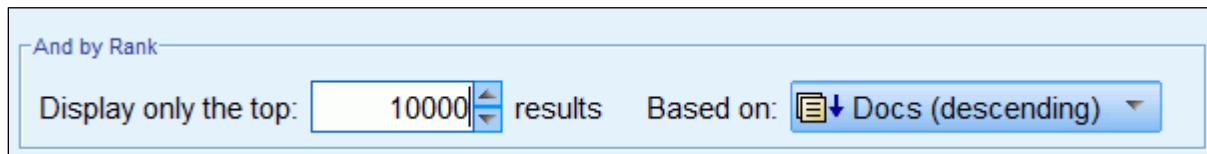
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Notice that all the categories you created using Text Link and Cluster Analysis are still there. You will use auto classification techniques to add to these categories.

There are two steps to prepare for categorization. The first is to determine the total number of concepts extracted. The information message in the Extraction pane lists 5,000 concepts, but that is not necessarily the total number of extracted concepts. This is a limit set in the Filter dialog box.

5. From the **Tools** menu, click **Filter**.
6. Change the value of **Display only the top** to **10000**.

This change allows up to 10,000 concepts to be displayed and used in the workbench.



7. Click **Filter**.

Concept	In	Global	Docs	Type
service	fx	421 (2%)	296 (23%)	<Unknown>
complaint	fx	374 (2%)	296 (23%)	<Negative>
fault	fx	339 (2%)	229 (18%)	<Negative>
problem	fx	327 (2%)	227 (18%)	<Negative>
cruel		246 (1%)	171 (13%)	<NegativeAttitude>
line	fx	244 (1%)	157 (12%)	<Unknown>
charge	fx	208 (1%)	155 (12%)	<Budget>
not satisfied	fx	159 (1%)	139 (11%)	<Negative>
not working		162 (1%)	131 (10%)	<NegativeFunctioni
new		179 (1%)	129 (10%)	<Contextual>
resolved		142 (1%)	122 (10%)	<PositiveCompete
phone		175 (1%)	121 (10%)	<Unknown>
tia level		119 (1%)	119 (9%)	<Unknown>
hill		159 (1%)	119 (9%)	<Burdnet>

Notice that there are 5840 concepts, so the majority of them were visible. By default, categorization will use the top 5,000 concepts based on document count, but it is a good idea to know the total number of concepts extracted in case you want to modify this limit.

8. In the **Categories** pane, click **Score**.

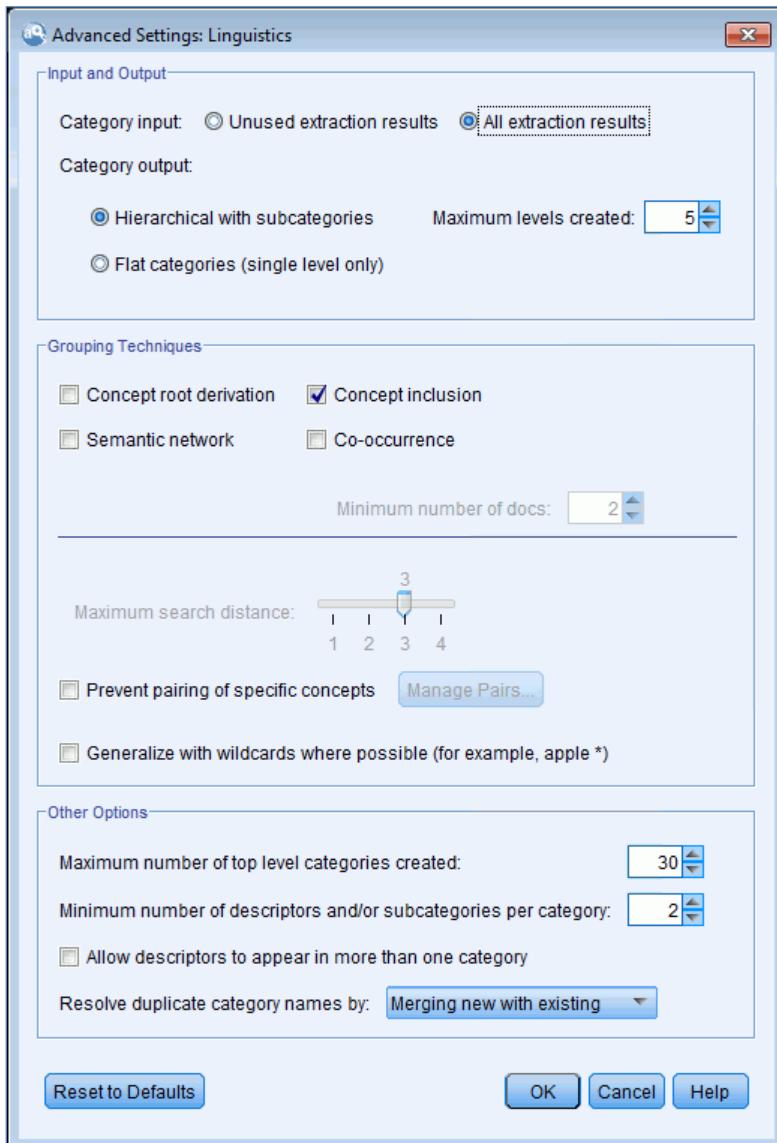
Category	Descriptors	Docs
All Documents	-	1269
Uncategorized	-	1143
No concepts extracted	-	2
poor coverage	1	7
fx poor & coverage		7
<Budget>+<Negative>	1	110
fx [<Budget> + <Negative>]		110
Noisy Phone	4	9
fx [lines person + noisy]		1
fx [line + noisy]		6
fx [service + noisy]		1
fx [rcf + noisy]		1

This is not essential, but just as with knowing the total number of concepts, it can help to know how many records are assigned to these categories as you begin. There are 110 records assigned to the "Budget+Negative" category, nine records assigned to the "Noisy phone" category, and seven records assigned to the "poor coverage" category.

Task 2. Use concept inclusion to create an initial set of categories.

1. In the **Categories** pane, click **Build**, and then click **Edit**.
2. Click **Select All**.
3. Ensure that **Use linguistic techniques to build categories** is selected, and then click **Advanced Settings**.
4. Clear the **Concept root derivation** and **Semantic network** check boxes.

5. Select All extraction results.



6. Click OK.

As mentioned above, 30 categories will be created by default. If you are eventually going to create a predictive model, you might prefer more categories, but this depends on the richness of the text data, the total number of records, and the minimum number of records you prefer to be included in a category.

7. Click Build Categories.

8. Click Score.

Category	Descriptors	Docs
All Documents	-	1269
Uncategorized	-	162
No concepts extracted	-	2
service	147	431
number	95	204
on line	90	145
bill	69	182
mobile	63	159
phone	53	158
account	49	146
charge	47	188
credit	44	80
time	40	141
cable	77	100

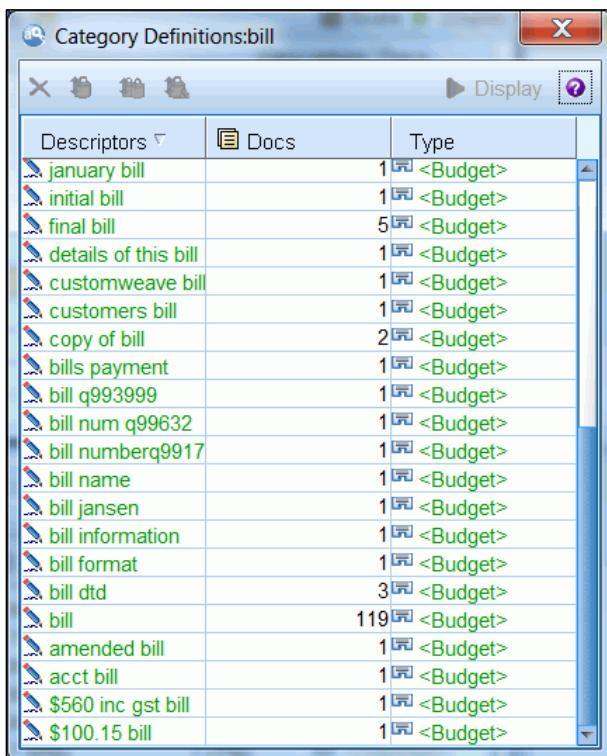
Task 3. Explore the results.

The category names come from the root term used for inclusion. Thirty categories were created, which you can see from the message in the status bar in the Extraction pane, which lists 33 categories total. Since there were two pre-existing categories, that implies that you reached the limit of the number of categories to create.

In the Extract results pane, notice that some of the concepts have a  symbol in the In column to indicate that the concept has been included in a category. The other concepts have not been used.

1. Double-click the **bill** category and scroll to the end.

The results appear as follows:



The screenshot shows a software window titled "Category Definitions:bill". The window has a toolbar with icons for file operations and a "Display" button. The main area is a table with three columns: "Descriptors", "Docs", and "Type". The "Type" column contains mostly "<Budget>" entries. The "Descriptors" column lists various terms related to bills, such as "january bill", "initial bill", "final bill", "details of this bill", "customweave bill", "customers bill", "copy of bill", "bills payment", "bill q993999", "bill num q99632", "bill numberq9917", "bill name", "bill jansen", "bill information", "bill format", "bill dtd", "bill", "amended bill", "acct bill", "\$560 inc gst bill", and "\$100.15 bill". The "Docs" column shows the count of documents for each descriptor, with "bill" having 119 documents.

Descriptors	Docs	Type
january bill	1	<Budget>
initial bill	1	<Budget>
final bill	5	<Budget>
details of this bill	1	<Budget>
customweave bill	1	<Budget>
customers bill	1	<Budget>
copy of bill	2	<Budget>
bills payment	1	<Budget>
bill q993999	1	<Budget>
bill num q99632	1	<Budget>
bill numberq9917	1	<Budget>
bill name	1	<Budget>
bill jansen	1	<Budget>
bill information	1	<Budget>
bill format	1	<Budget>
bill dtd	3	<Budget>
bill	119	<Budget>
amended bill	1	<Budget>
acct bill	1	<Budget>
\$560 inc gst bill	1	<Budget>
\$100.15 bill	1	<Budget>

Many of these descriptors are of the type "Budget", and they seem reasonable to include in a "bill" category. But there are at least a few concepts that seem to be referencing a "bill number", which could be a telephone number. This is not the same thing as a request about monthly billing or bill payment, so you will check this possibility.

2. Ctrl+click **bill num q99632** and **bill numberq99172**, and then click **Display**.

3. In the **Data** pane, click the second record.

query (2) ▾		Categories	Text Preview
1	As per letter. Customer advises that she is disputing directory chatges on bill number99172	bill	TIA level 1 complaint - no ref num customer called ASTROSERVE via the TIA dipsuting idd call made from the internet on acct num , bill num q99632 totalling \$80.399sep 07:52p tokelau is 9979 every 29:....
2	TIA level 1 complaint - no ref num customer called ASTROSERVE via the TIA dipsuting idd call made from the internet on acct num , bill num q99632 totalling \$80.399sep 07:52p tokelau is 9979 every 29:4999.60 _d 8999sep 08:25p tokelau is 9914 every 0.999.10 _d 8999sep 08:28p tokelau is 9978 every 29:4999.60 total for 3999475 \$80.30customer said that he son was on the internet before these calls and downloaded screen savers but never at any time he has sn any warnings about charges, customer also wants to know the site name for the phone num provided on the bill and what service they do provide, customer needs these calls invesTIaged.	bill acct charge internet name phone	

These text data use the word "bill" in a different sense than the other entries. However, these records are still appropriate to include in a "bill" category because reading the records in full reveals that each is a call about a bill or charging dispute. Remember that it is always important to investigate concepts that fall into a category.

Categories can also be combined when that makes sense. There is a charge category and a fee category. Both of them include concepts about an extra cost to the customer for some service or change in their internet or phone plans.

Now you will combine them by adding the first to the second category.

4. Close the **Categories Definitions:bill** window.
5. Right-click the **charge** category, and then click **Move to Category**.
6. In the **All Categories** window, scroll down, select **fee**, and then click **OK**.

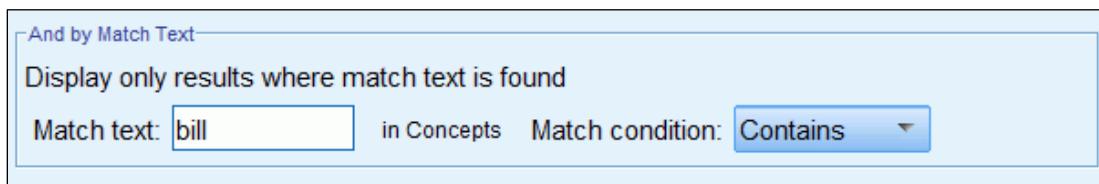
7. Click **Score**, scroll down to **fee**, and then collapse **fee**.

Category	Description	Docs
account	49	146
credit	44	80
time	40	141
cable	37	109
business	36	102
email	36	80
fee	82	230
work	29	65
internet	27	87
dsl	27	41
connection	24	63
payment	24	39

The two categories are combined, but now the data need to be rescored to display the number of records in the combined category. The combined category has the name **fee** because the charge category was added to **fee**. The number of records in the **fee** category increased from 64 to 230.

Another way to modify categories is by adding more concepts to existing categories. The "bill" category should pick up all mentions of bills or billing, but it can be a good idea to double-check by filtering the extracted concepts for terms with the text string "bill."

8. Click an extracted concept to activate the **Extraction** pane.
 9. From the **Tools** menu, click **Filter**.
 10. In the **Match text** box, type **bill**.



11. Click Filter.

The screenshot shows the Extraction pane with the following data:

Concept	In	Global	Docs	Type
bill		159 (1%)	119 (9%)	<Budget>
billing		61 (0%)	52 (4%)	<Budget>
single bill	fx	18 (0%)	14 (1%)	<Budget>
final bill		6 (0%)	5 (0%)	<Budget>
phone bill	fx	4 (0%)	4 (0%)	<Budget>
billing address		3 (0%)	3 (0%)	<Budget>
direct billing		3 (0%)	3 (0%)	<Budget>
current bill		3 (0%)	3 (0%)	<Budget>
monthly billing		7 (0%)	3 (0%)	<Budget>
bill dtd		3 (0%)	3 (0%)	<Budget>
account		2 (0%)	2 (0%)	<Budget>
billing name		2 (0%)	2 (0%)	<Budget>

At the bottom left, it says "33 (157) Categories".

There are 96 concepts with the text string "bill." However, not all of them were included in the billing category. A quick glance at the unused concepts will help you determine whether or not to add them to the category. To more easily examine them, first sort the concepts according to whether or not they were used in a category.

12. In the Extraction pane, click the In column header.

The screenshot shows the Extraction pane with the following data, sorted by the In column:

Concept	In	Global	Docs	Type
pre-bill		1 (0%)	1 (0%)	<UNKNOWN>
confirmed		1 (0%)	1 (0%)	<Budget>
times to amend		1 (0%)	1 (0%)	<Budget>
billing		61 (0%)	52 (4%)	<Budget>
billing disputehi		1 (0%)	1 (0%)	<Budget>
disputing billing		1 (0%)	1 (0%)	<Budget>
inquiries to the		1 (0%)	1 (0%)	<Budget>
monthly billing		7 (0%)	3 (0%)	<Budget>
single billing		1 (0%)	1 (0%)	<Budget>
billing		1 (0%)	1 (0%)	<Budget>
advice to pay		1 (0%)	1 (0%)	<Budget>
mobile bill		1 (0%)	1 (0%)	<Budget>

At the bottom left, it says "33 (157) Categories".

This put all the unused concepts at the top. It turns out that there were many concepts related to billing that were not used. After a quick examination of them, it appears that they all involve billing questions and should be added to the "bill" category.

13. Shift+click all of the unused concepts, right-click the highlighted concepts, and then click **Add to Category**.
14. Select **bill**, and then click **OK**.

15. Click Score.

Category	Descriptors	Docs
All Documents	-	1269
Uncategorized	-	161
No concepts extracted	-	2
service	147	431
number	95	204
on line	90	145
bill	94	229
customweave bill	1	
amended bill	1	
reading bill	1	
details of this bill	1	
\$560 inc gst bill	1	
\$100 15 bill	1	

The number of descriptors for the billing category increases to 94 and the number of categorized documents rose to 229.

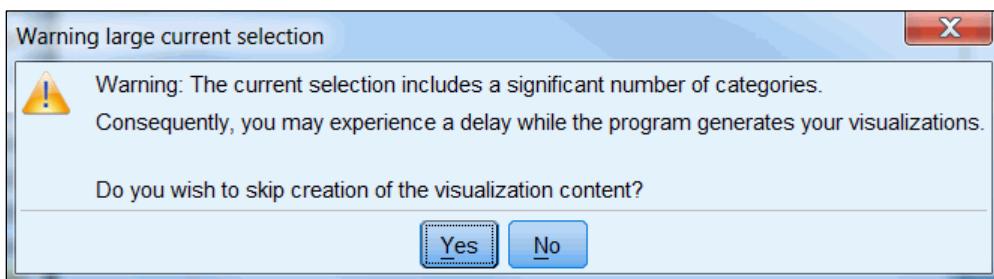
You could investigate where there are any unused concepts that should be added to other categories (for example, "mobile"), but in the interest of time, you will not go through another example.

Task 4. Examine the categories in the Visualization pane.

Next you will use of the Visualization pane to display relationships between categories. The Visualization pane provides graphs and tables that allow you to understand the relationship between a selected category and all other categories. This can help you to decide whether two categories should be combined (they seem to refer to the same thing but do not occur together often), the meaning of a category, or to give you other hints about categorization.

1. From the **Tools** menu, click **Filter**.
2. Remove **bill** from the **Match text** box.
3. Click **Filter**, select the **phone** category, and then click **Display**.

If the following message displays, click No:



- Click the **Selection %** column header to sort the results into descending order based on percentages.

The selected category will display a selection % of 100, and all of the other bars represent the percentage of documents that are in the selected category and the other unselected categories.

Category	Bar	Selection %	Docs
phone	100.0	158	
service	32.3	51	
mobile	24.7	39	
number	24.1	38	
bill	21.5	34	
time	18.4	29	
charge	15.2	24	
on line	14.6	23	
consultant	10.8	17	
account	10.1	16	
business	10.1	16	
cable	8.9	14	
pay	8.9	14	
internet	7.6	12	
..	7.0	10	

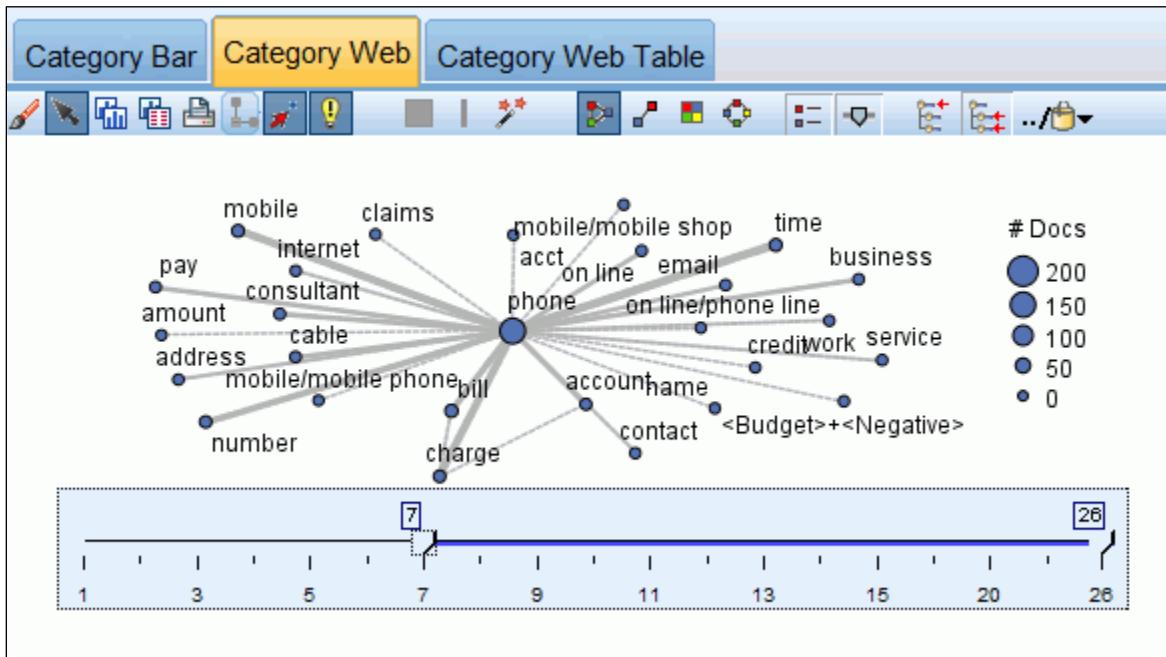
The Visualization pane displays a category bar chart for phone. This shows the number of records in this category, and, the number of these records that also include other categories. Notice that the mobile category occurs in 39 of these records, or 24.7% of this group. Many of the other categories occur at least once or twice.

If "mobile" and "phone" were always mentioned together in the same document, you could get rid of one of them because there would be complete overlap. However, in this case it is apparent that Astroserve customers were not always referring to mobile phones when they called about their phone service so it would probably be best to leave "mobile" and "phone" as separate categories unless you wanted to merge them into a more generic category called "telephones" or something like that.

You can also see the same information in either a Web graph or in a table.

- Click the **Category Web**.

6. Click the **Show slider** tool, and then move the slider to 7.



You may prefer to display the information in a Category Web table.

7. Click the **Category Web Table** tab.
8. Click the **Count** column heading until the counts are listed in descending order.

The figure shows the Category Web Table interface. At the top, there are three tabs: "Category Bar" (blue), "Category Web" (light blue), and "Category Web Table" (yellow, selected). Below the tabs is a toolbar with a "refresh" icon. The main area is a table with three columns: "Count", "Category 1", and "Category 2". The "Count" column is sorted in descending order, showing values like 38, 26, 24, 24, 20, etc. The "Category 1" and "Category 2" columns list pairs of concepts separated by parentheses, such as "service(39)", "time(26)", etc.

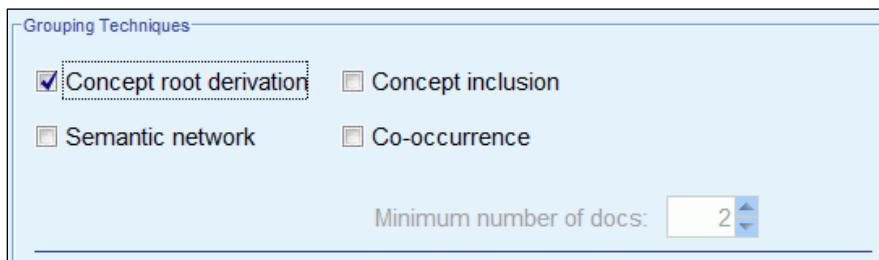
Count	Category 1	Category 2
38	service(39)	phone(153)
26	time(26)	phone(153)
24	charge(24)	phone(153)
24	mobile(26)	phone(153)
20	number(21)	phone(153)
18	phone(153)	bill(30)
16	account(16)	phone(153)
16	consultant(17)	phone(153)
14	business(14)	phone(153)
14	pay(14)	phone(153)
14	cable(14)	phone(153)
13	on line(13)	phone(153)

In this table, the Count column contains the number of shared records. The numbers in parentheses are the number of records in that category, among the group of records that are selected. In this example, the concept "mobile" was contained in 26 of the documents that also contained the "phone" concept.

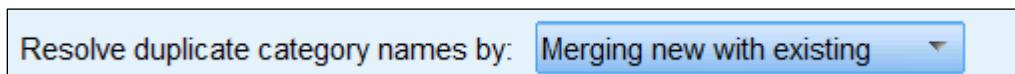
Task 5. Use concept root derivation to supplement the categories.

You will create some additional categories are created using the concept derivation method.

1. Click **Build**, and then click **Edit**.
2. Click **Advanced Settings**.
3. Select **Concept root derivation**, and deselect **Concept inclusion**.



4. If necessary, beside **Resolve duplicate category names by**, select **Merging new with existing**.

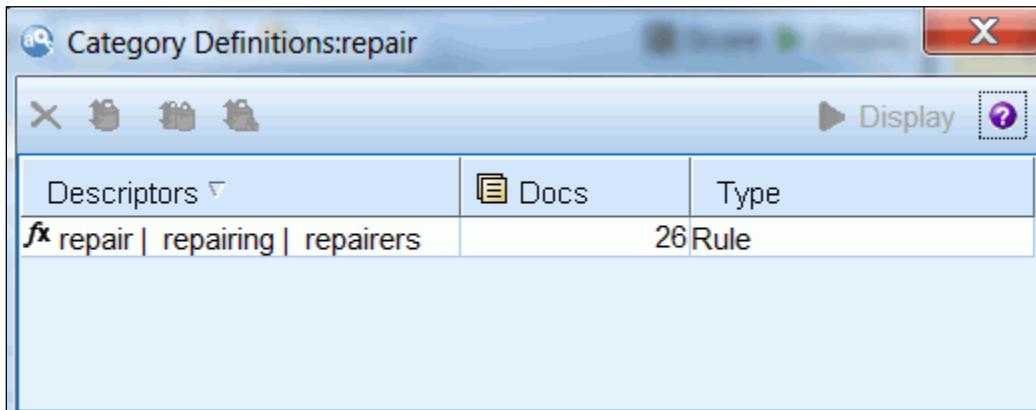


5. Click **OK**.
6. Click **Build Categories**, and then click **Score**.

The note in the Status bar in the Extraction pane lists 62 categories, so concept root derivation added another 29 categories. If the categories of the same name were not merged together, they would remain separate, though the name of one of them would end with the digit 1 so you could tell them apart.

Something else that might seem odd is that there is only one descriptor for the new categories. To understand why, view one of these categories.

- Double-click the **repair** category.



The number of descriptors is listed as one because the category is created with an OR rule that says that any record that includes the terms "repair", "repairing", or "repairers" should be assigned to this category.

- Close the **Category Definitions:repair** window.

Task 6. Add types as categories.

When the linguistic resources were edited in a previous module, several types were created to eventually use as categories. The automatic categorization of types in the Build dialog will not be used because that will simply take the most frequent types. Instead, the created types need to be added manually.

- Click the list in the **Extraction pane** and select **Type**.
- Ctrl+click to select the <Phones> and <competitors> types.
- In the **Extraction** pane, click **Create categories for each descriptor** to add each to a separate category.

4. Click **Score** and then scroll to the end of the list to see <competitors> and <Phones>.

Category	Descriptors	Docs
mgr		
list	1	21
long	1	25
connect	1	24
details	1	22
single bill	1	15
records	1	17
mrs boben	1	14
process	1	14
on hold	1	11
log	1	13
mrs bobh	1	7
<competitors>	1	42
<Phones>	1	11

There is only one descriptor listed for each type because this is not a category that was created from several concepts. There are now only 133 uncategorized records, though there are still a large number of unused concepts.

Task 7. Update the Text Mining node to save the categories.

1. From the **File** menu, click **Update Modeling Node**.
 2. Click **OK**, and then click **OK** to the message that the modeling node has been updated.
 3. From the **File** menu, click **Close** and then click **Exit** to end the Interactive Workbench session.
You will save the stream.
 4. From the **File** menu, click **Save Stream As**.
 5. Name the stream **Creating Categories_demo1_end.str**, and then click **Save**.
 6. From the **File** menu, click **Close Stream**.
 7. From the **File** menu, click **New Stream**.
- Do not close Modeler; leave it open for the next demo.

Results:

You have successfully used automated classification techniques to create an initial set of categories. In addition, you have been able to supplement these categories with unused extractions with these categories and by adding types as categories.

Extending Categories

- Extending categories is a method of enhancing categories by using automatic categorization to identify concepts, patterns and rules related to existing categories.
- You can extend one, several, or all of the categories.
- Categories should not be extended multiple times because they can become too general.

© 2014 IBM Corporation



Another method of enhancing categories is the extension of categories. Extending is a process through which descriptors are added or enhanced automatically to “grow” existing categories. The objective is to produce a better category that captures related records that were not originally assigned to that category. You can extend one, several, or all of your categories.

The automatic categorization techniques you select to extend will attempt to identify concepts, patterns, and rules related to existing category descriptors. These new concepts, patterns, and rules are then added as new descriptors or added to existing descriptors. The available techniques include the linguistic techniques and co-occurrence rules. There is an option to Extend categories with descriptors based on category names, which generates descriptors using the words in the category name itself; therefore, the more descriptive the category names, the better the results.

Extending categories can be particularly important after importing a code frame in which the categories have very descriptive names, after creating categories manually and adding simple rules and descriptors, or after refining categories from a TAP. You can always try extending any set of categories just to see the results.

A category can be extended multiple times, but sometimes this can result in too general a category. Also, since building categories and extending categories use similar algorithms, extending categories directly after building categories is unlikely to produce more interesting results.

It is not a good practice to extend all the categories at once, unless you have just a few categories. If many categories are extended, it can take some time and be somewhat confusing to review all the changes, so it is usually better to do a few categories at a time.

The Extend Categories dialog is very similar to the Build Categories dialog, so this course only mentions some of the differences. By default, the Extend categories with descriptors based on categories names option is selected. The category name is scanned to see if words in the name match any extracted concepts. If a concept is recognized, it is used to find matching concept patterns and these both are used to form descriptors for the category.

The setting for the Maximum number of items to extend a descriptor by option will limit how many concepts and types can be added to a category descriptor. You may prefer to cut down the number of extenders by using the Generalize with wildcards where possible option.

The parent/child grouping techniques list provides three options, including child concepts, parent concepts, or both. As an example, the Parents concept option concentrates on finding concepts that are more generically related. It will look upwards to extend by including only parent relationships, which are often more ambiguous. This option creates more results but may group concepts into categories that are not closely linked in the context of your data. For example, if the category has the concepts melon and apple, then Parent concepts could yield fruit as a concept to extend the category. This may or may not match your intent for the category.

Creating Rules

- You can manually build rules to classify records into a category based on a logical expression.
- Each rule is attached to a single category so that each record matching the rule is automatically categorized into that category.

© 2014 IBM Corporation



In general, category rules are statements that you can create to automatically classify records into a category based on a logical expression. The statements are based on a logical expression using extracted concepts, types, and patterns as well as Boolean operators. While some category rules are generated automatically by category building techniques such as co-occurrence rules and concept derivation, you can also create category rules manually. Each rule is attached to a single category so that each record matching the rule is automatically categorized into that category. As an example, a category rule could be an expression to include all records that contain the concepts "dinner" and "cost" or "price".

The category rule editor is used to create a rule (or edit an existing one). You open the rule editor by editing an existing rule or by right-clicking the category name and choosing Create Rule. You can add concepts, types, or patterns as well as use wildcards to extend the matches in the rule. It is best to use recognized concepts, types and patterns in a rule since the rule will then find all related concepts. For example, when you use a concept, all of its associated terms, plural forms, and synonyms are also matched to the rule. Likewise, when you use a type, all of its concepts are also captured by the rule.

The rule has a simple format: [repair | repairing | repairers]. In English, this means that a case should be assigned to this category if the response mentioned either repair or repairing, repairers.

The supported syntax symbols for rules are listed in table.

Syntax Character	Description
&	The "and" Boolean
	The "or" Boolean, which means that if any of the elements are found, a match is made
!()	The "not" Boolean
+	The pattern connector used to form an order-specific pattern. When present, the square brackets must be used
()	An expression delimiter. Any expression within the parenthesis is evaluated first
[]	The pattern delimiter that is required if a pattern is being defined
*	A wildcard representing anything from a single character to a whole word depending on how it is used

If you have a concept that contains any character that is also a syntax character you must place a backslash in front of that character so that the rule is properly interpreted. When you drag a concept into the editor, backslashes are automatically added for you.

To avoid errors when creating or modifying rules, it is best to drag concepts directly from the Extraction Results pane or the Data pane into the rule editor or add them from the context menus.

For an example of creating a rule, you can construct a category that captures those responses that mention lunch, dinner, or a meal and also mention the menu. This will be a new category, so your first step will be to create a new blank category in which to place the rule.

Business Analytics software

IBM

Demo 2

Fine Tuning Categories



© 2014 IBM Corporation

The following file(s) are used in this demo:

- Creating Categories_demo2_start.str - a Modeler stream that reads a file containing call center data for March and April

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

12-27

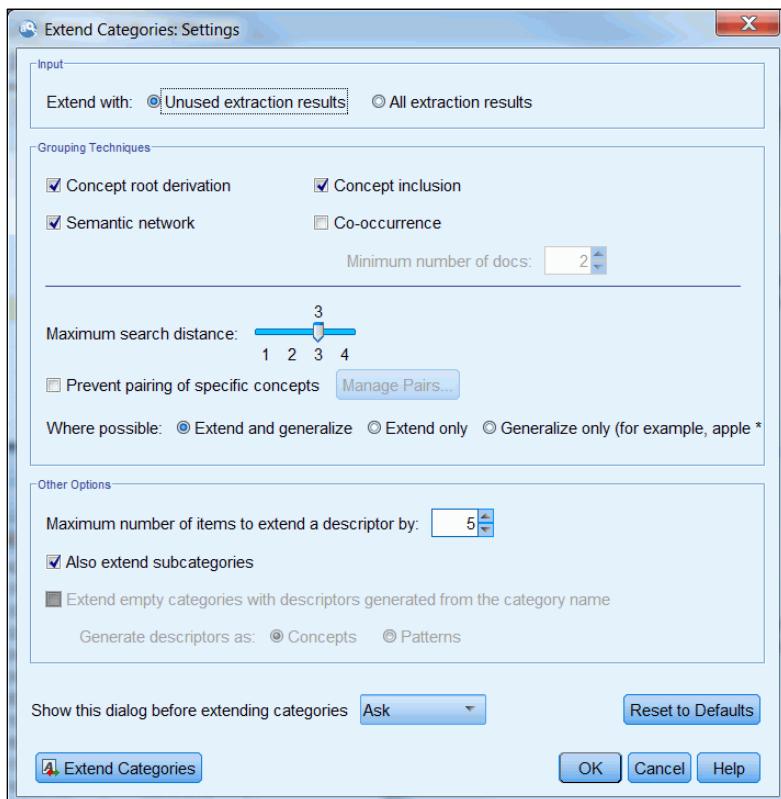
Demo 2: Fine Tuning Categories

Purpose:

Now that you have created the initial categories for the Astroserve data, the next step is to extend some of these categories with additional relevant descriptors. Also, you would like to create business rules that classify records into categories based on a logical expression

Task 1. Extend categories.

1. From the **File** menu, click **Open Stream**.
2. Navigate to **C:\Train\0A105\12-Creating_Categories**, and then double-click **Creating Categories_Demo2_start.str**.
3. Run the **Text Mining** modeling node.
4. Switch to the **Categories and Concepts** view.
5. Collapse all of the categories, and then right-click the **repair** category.
6. Click **Extend Categories**, and then click **Edit**.



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

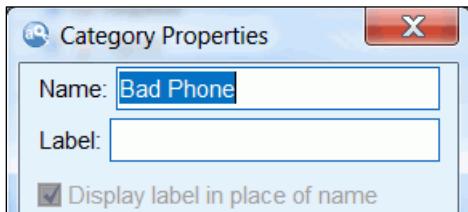
7. Select All extraction results, click Extend Categories, and then click Score.

Category	Descriptors	Docs
acct	20	48
consultant	19	77
date	18	48
order	17	61
claims	17	106
repair Extended	7	57
Noisy Phone	4	9
states	1	49
contract	1	58
house	1	29
poor coverage	1	7
report	1	44
rent	1	11
<Phones>	1	11

The label "Extended" is added to the categories that were extended. There were originally 26 responses in the repair category, so 31 have been added. The number of descriptors has increased from 1 to 7. These were not found earlier because this time several categorization techniques were used at once, including some that were not used before such as semantic network.

Task 2. Create rules.

1. From the **Categories** menu, click **Create Empty Category**.
2. Change the name from **New Category** to **Bad Phone**.



3. Click **OK**, and then scroll to the end of the list.

Category	Descriptors	Docs
high		
advising	1	2
wrong	1	2
<competitors>	1	2
mrs bobh	1	2
lodge	1	2
connect	1	2
mrs boben	1	2
exist	1	2
resolved	1	2
records	1	2
poor coverage	1	2
<Budget>+<Negative>	1	2
Bad Phone	0	2

4. Right-click the **Bad Phone** category, and then click **Create Category Rule**.
The rule editor appears.
 5. In the rule box, type the string **<phones> & <negative>**.

& | !() () [] + * Rule Name: rule 1 Category: Bad Phone
<phones> & <negative>

6. Click **Test Rule**.
 7. Click the **Selection %** heading in the bar graph to sort the percentages in descending order.

Category	Bar	Selection %	Docs
Bad Phone		100.0	8
<Phones>		100.0	8
phone		50.0	4
consultant		25.0	2
fee		25.0	2
mobile		25.0	2
bill		12.5	1
acct		12.5	1
account		12.5	1
<Budget>++		12.5	1
cancel		12.5	1

query (8) -

	Categories
1	Follow sent from nbc as follows david not happy with \$310.00cr, believes should be more. Have run thru backdater & come up with same figure. Advised david of this & have applied \$40 service guarantee. David still unhappy about the way this has been handled since the start and wants referred to mrg. Previous notes include cust should be on staff 5 plan with mro \$589 for 24 months for nokia 8310 on the 17/12 can you please adjust and back date calls as from the 17/12, Bad Phone <Phones> service

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

The results indicate that eight respondents were categorized into the **Bad Phone** category. The person in the first record said they were unhappy with their Nokia 8310. Although eight customers are not very many, you will keep the category because it is reasonable to assume that customers who are unhappy with their phones are likely to churn.

8. Click **Save & Close** to save **Bad Phone** as a category.
9. Click **Score**.

Category	Descriptors	Docs
missed		
day	1	48
list	1	21
high	1	25
advising	1	12
wrong	1	81
<competitors>	1	42
mrs bobh	1	7
lodge	1	29
connect	1	24
mrs boben	1	14
exist	1	6
resolved	1	124
Bad Phone	1	8

Task 3. Update the Text Mining node to save the changes.

1. From the **File** menu, click **Update Modeling Node**.
2. Click **OK**, and then click **OK** to the message that the modeling node has been updated.
3. From the **File** menu, click **Close** and then click **Exit** to end the interactive session.
4. From the **File** menu, click **Save Stream As**.
5. Name the stream **Creating Categories_demo2_end.str**, and then click **Save**.
6. From the **File** menu, click **Close Stream**.
7. From the **File** menu, click **New Stream**.

Do not close Modeler; leave it open for the next demo.

Results:

You were able to use Extension to add descriptors to the repair category. In addition, you were able to create a new category called Bad Phones that you designed to capture negative sentiment toward particular types of phones.

Creating a Final Set of Categories

- At the end of the categorization process, you need to decide which categories to keep and which ones to discard.
- Most analysts believe that categories with too few records can be deleted.
- Domain expertise is necessary to determine which categories to keep; without it you cannot accurately judge which categories are relevant and which are not.

© 2014 IBM Corporation



All the methods of creating categories have been discussed and demonstrated. Because this is not an exercise in creating all possible categories, at this point these categories will suffice. What then remains is to decide which categories to keep among the very many that have been created.

Whether you are interested in understanding and extracting the information in text, or in using this information in modeling, most analysts find that categories with too few records are not useful for either task (the primary exception is "needle in a haystack" problems, such as in intelligence or fraud work where analysts look for unique relationships among a small number of entities).

Creating a Text Analysis Package

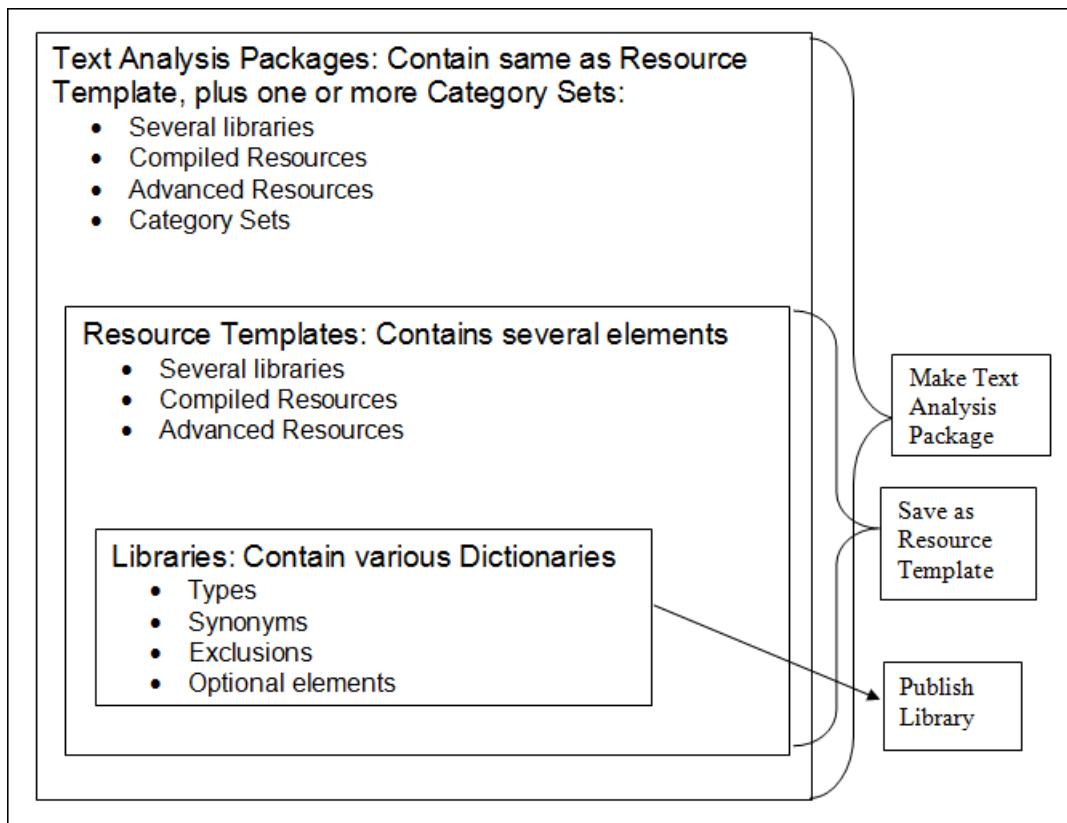
- A text analysis package (TAP) is a predefined set of libraries and advanced linguistic and nonlinguistic resources bundled with one or more sets of predefined categories.

© 2014 IBM Corporation

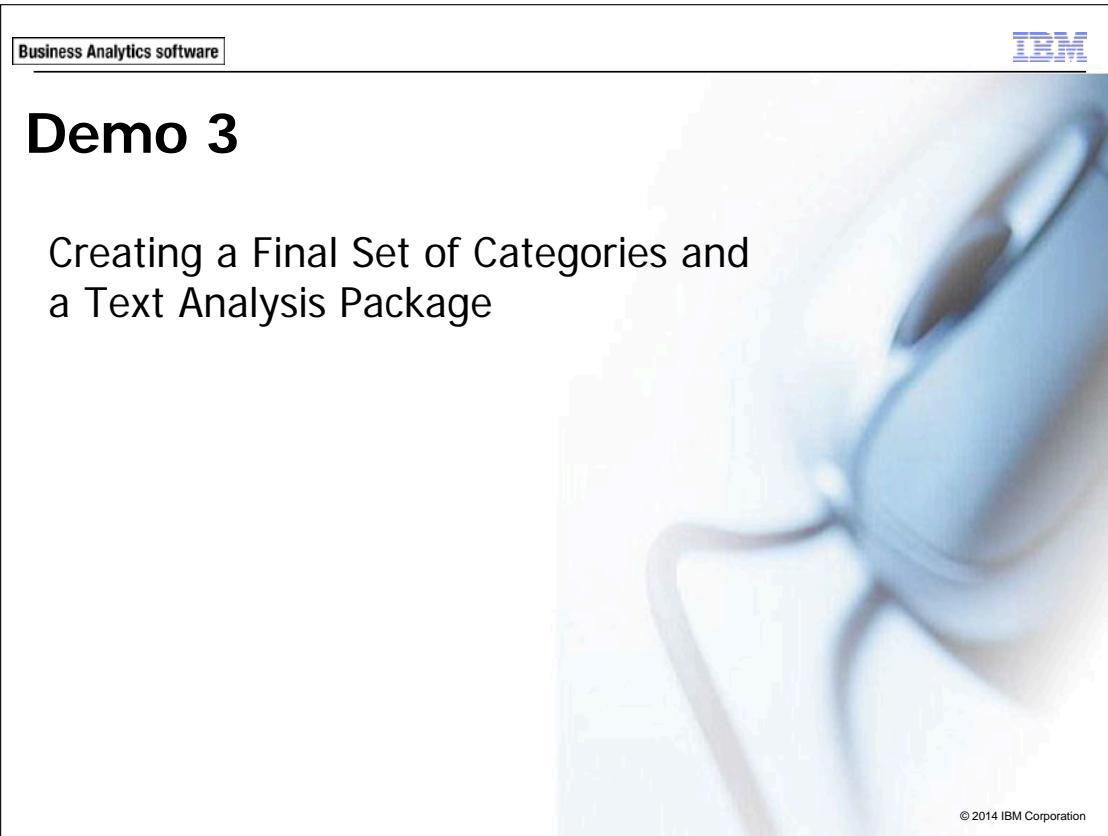


A text analysis package was used previously for the initial analysis of the Astroserve data. A TAP contains the linguistic resources in a resource template, plus an additional element. TAPs also include one or more category sets, which are predefined categories (category names and code values), plus the set of descriptors for each category-concepts, patterns, rules, and so on-that will code responses into these categories from extracted results.

A visual depiction of the relationship between libraries, resource templates, and text analysis packages is displayed on the following page. Libraries are contained within Resource Templates, and TAPs contain the equivalent of a Resource Template, plus one or more category sets. Each of these elements can be saved separately.



If you have a finished project with linguistic resources and categories that you would like to save so they can be used on future text data, you can make a TAP from the project contents.



The slide is titled "Demo 3" and discusses creating a final set of categories and a text analysis package. It features the IBM logo and a copyright notice at the bottom right.

Business Analytics software

IBM

Demo 3

Creating a Final Set of Categories and a Text Analysis Package

© 2014 IBM Corporation

The following file(s) are used in this demo:

- Creating Categories_demo3_start.str - a Modeler stream that reads a file containing call center data for March and April.
- Astroserve05.str - a Modeler stream that reads a file containing call center data for May. This file was not used to build the model that you just completed.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

12-35

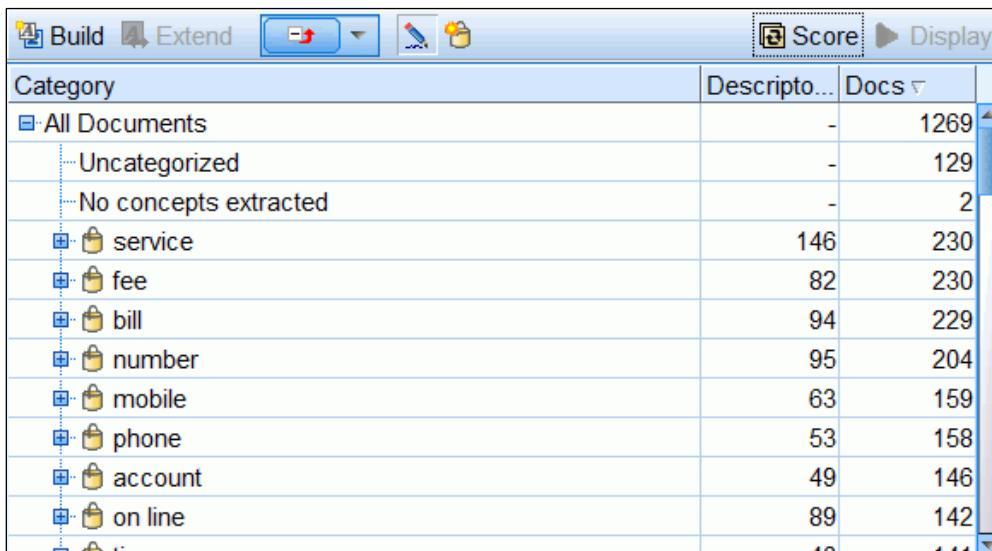
Demo 3: Creating a Final Set of Categories and a Text Analysis Package

Purpose:

After creating categories, you will decide which categories to keep among the very many that have been created. You also need to create a Text Analysis Package because you intend to use these categories to classify new customer data when you get it.

Task 1. Create a final set of categories.

1. From the **File** menu, click **Open Stream**.
2. Navigate to **C:\Train\0A105\12-Creating_Categories**, and then double-click **Creating Categories_demo3_start.str**.
3. Run the **Text Mining** modeling node.
4. Switch to the **Categories and Concepts** view, and then click the **Score** button.
5. Click the collapse  button to collapse the categories.
6. Click the **Docs** column heading until the counts are listed in descending order.



The screenshot shows the 'Categories and Concepts' view in IBM SPSS Modeler. The interface includes a toolbar with 'Build', 'Extend', and other icons, and a menu bar with 'Score' and 'Display' selected. The main area displays a table with three columns: 'Category', 'Description', and 'Docs'. The 'Category' column lists various document types, some of which are collapsed under 'All Documents'. The 'Description' column contains brief definitions, and the 'Docs' column shows the count of documents for each category. The data is sorted by 'Docs' in descending order.

Category	Description	Docs
All Documents	-	1269
Uncategorized	-	129
No concepts extracted	-	2
service	146	230
fee	82	230
bill	94	229
number	95	204
mobile	63	159
phone	53	158
account	49	146
on line	89	142
time	40	141

7. Delete all categories with less than 25 records, except the categories that you created from types (Bad Phone, poor coverage, Noisy Phone, <Phones>).

8. Right-click <Phones>, click **Rename Category**, and then remove the brackets from the name.
9. Repeat step 8 for any other categories with brackets around the name.
10. Right-click the category **acct**, and then click **Move to Category**.
11. In the **All Categories** window, select **Ascending: A-Z**, select **account**, and then click **OK**.

When this is done, you should have 48 categories. You could certainly make several more modifications but in the interest of time, you will stop here. Even though many categories have been deleted, most records have still been assigned to at least one category: there are only 133 records that have not been categorized.

Task 2. Update the Text Mining node, and generate a model.

1. From the **File** menu, click **Update Modeling Node**.
2. Click **OK**, and then click **OK** to the message that the modeling node has been updated.

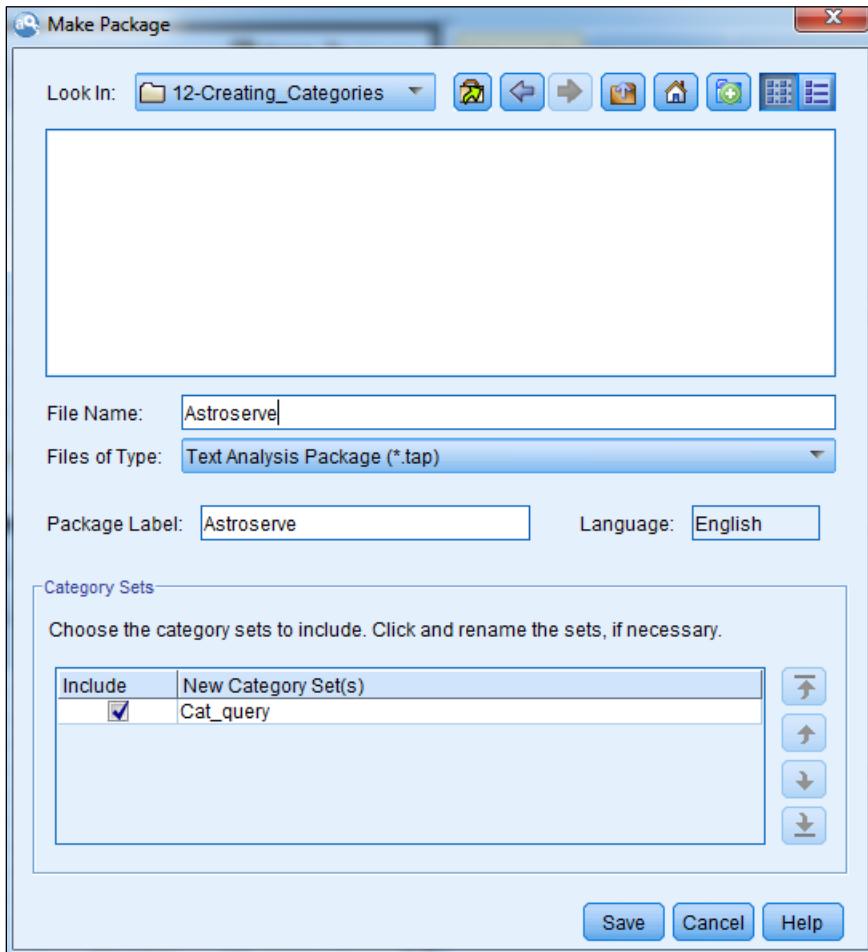
Now that the project has been completed, generate a model with the categories for later scoring.

3. From the **Generate** menu, click **Generate Model**.

Task 3. Create a Text Analysis Package.

1. From the **File** menu, point to **Text Analysis Package**, and then click **Make Package**.
2. Navigate to the **C:\Train\0A105\12-Creating_Categories** folder.

3. In the **File Name** box, type **Astroserve**.



4. Click **Save**.

The TAPs are not stored in a special database but instead are stored in individual files, as you can see for the TAPs shipped with the product. This means that you can send a TAP you created to another user, who can then place it in the TAP folder on their computer (this is where the software looks for the TAPs), and use it for new projects. (You are saving the file in a different folder just in case you do not have write access to the TAPs folder.)

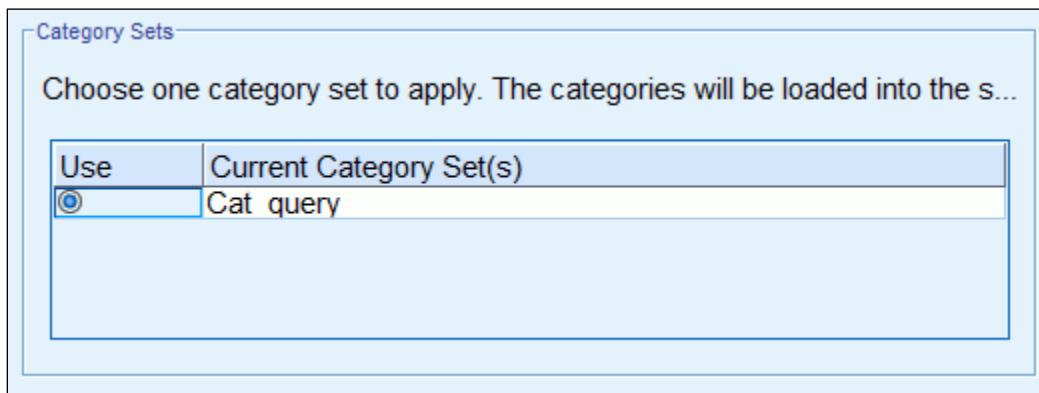
5. From the **File** menu, click **Close**, and then click **Exit**.
You will save the stream.
6. From the **File** menu, click **Save Stream As**.
7. Name the stream **Creating Categories_demo3_end.str**, and then click **Save**.

Task 4. Use the Text Analysis Package to categorize new data.

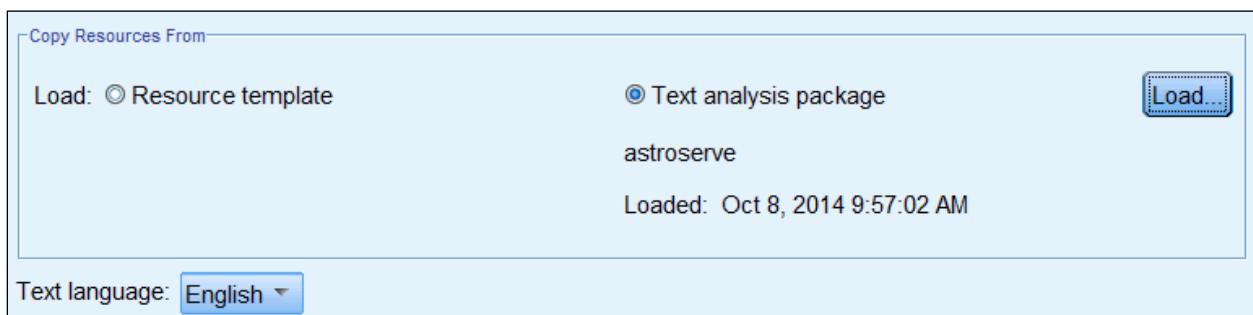
You will use a TAP to categorize newer Astroserve data.

1. From the **File** menu, click **Open Stream**.
2. Navigate to **C:\Train\0A105\12-Creating_Categories**, and then double-click **Astroserve05.str**.
3. Edit the **Text Mining** modeling node.
4. Click the **Model** tab, and ensure that **Exploring text link analysis (TLA) results** is selected.
5. Click the **Text Analysis Package** option.
6. Click **Load**.
7. Navigate to the **C:\Train\0A105\12-Creating Categories** folder, and then click **Astroserve.tap** to select it.

In the Category Sets area, you will see a list of sets of categories contained within the TAP. You will need to select the set you want to use. For example, a TAP could contain separate sets of categories for Positive, Negative or Mixed Opinions on a subject. In this case there is only one set to choose from called **Cat_Query**.



8. Click the **Load** button.



9. Click **Run**.
10. After the extraction is complete, switch to the **Categories and Concepts** view.
11. Click the **Score** button, and then if necessary, collapse the categories.

Category	Description	Docs
All Documents	-	3138
Uncategorized	-	496
No concepts extracted	-	1
account	69	431
address	23	89
amount	23	70
Bad Phone	1	18
bill	94	413
Budget+Negative	1	253
business	36	167
cable	37	203
cancel	1	92

The categories and dictionary resources that were created in this module have been successfully reused.

If you want to make improvements to a category set, linguistic resources, or make a whole new category set, you can update a text analysis package. To do so, you must be in the open project containing the information you want to put in the TAP. When you update, you can choose to append category sets, replace resources, change the package label, or rename/reorder category sets. You can access this dialog from the menus with File \ Text Analysis Package \ Update Package.

At this point you will end the interactive session and exit from Modeler.

12. From the **File** menu, click **Close** and then click **Exit** to end the Interactive Workbench session.
13. From the **File** menu, click **Exit** and then click **Exit** to end the Modeler session.

Results:

You have successfully finished categorizing the Astroserve data, generated a model, created a Text Analysis Package, and used the Text Analysis Package to categorize some new data.

Apply Your Knowledge

Purpose:**Test your knowledge of the material covered in this module.**

Question 1: True or False: Linguistic and frequency based methods can both be used in the same project.

- A. True
- B. False

Question 2: True or False: The same category can be extended multiple times.

- A. True
- B. False

Question 3: True or False: You cannot use more than one of the linguistic categorization techniques at a time in a project.

- A. True
- B. False

Question 4: True or False: The following rule "* not working" will capture those responses in which people mention that a product is "not working".

- A. True
- B. False

Question 5: True or False: If you create the category rule <Organization> & !(ibm) as a descriptor, it would match the following text "SPSS Inc. was a company founded in 1967" and not match the following text "the software company was acquired by IBM."

- A. True
- B. False

Apply Your Knowledge - Solutions

Answer 1: A. True

Answer 2: A. True

Answer 3: B. False. You can use more than one technique at a time but it is usually preferred to use one at a time so you can evaluate how well each technique does categorizing the data

Answer 4: A. True

Answer 5: A. True. The ! symbol indicates Not. Thus, the rule will only match if IBM is not contained in the sentence, which is the case in the first sentence but not the second.

Summary

- At the end of this module, you should be able to:
 - create categories using various categorization techniques
 - discuss strategies for categorization
 - use linguistic based categorization techniques
 - visualize relationships among categories
 - use conditional rules to create categories
 - extend categories
 - create text analysis packages

© 2014 IBM Corporation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

12-43

Business Analytics software

IBM

Workshop 1

Creating Categories



© 2014 IBM Corporation

The following file will be used:

- Music_Survey with Text Link Analysis.str - a Modeler stream that reads from a file containing customer likes and dislikes about a portable music player

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

12-44

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

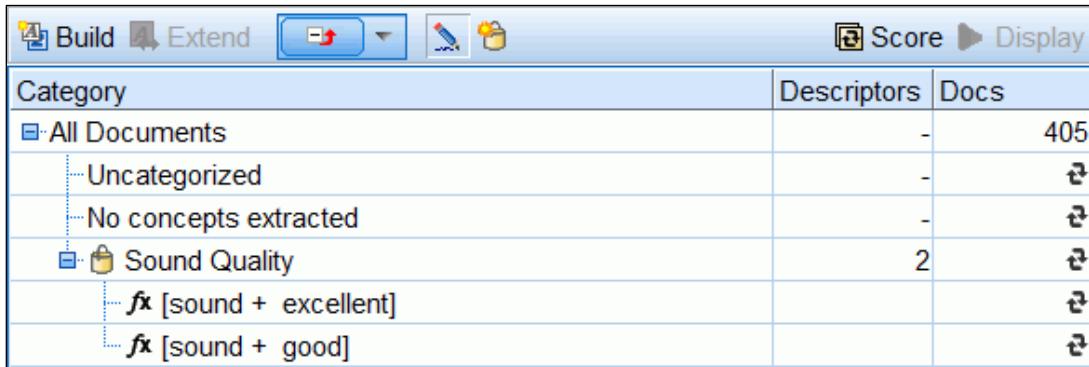
Workshop 1: Creating Categories

After preparing the dictionary resources, the next step is to categorize the responses so that you can analyze the data.

- Open and run the stream (C:\Train\0A105\12-Creating Categories folder\Music_survey with Text Link Analysis.str)
- Run the Text Mining node.
- Rename the category to Sound Quality created using text link analysis.
- Build additional categories using the linguistic categorization methods. Since responses are mainly lists, build the categories from types.
- Use the combination of concept derivation, concept inclusion, and semantic network on all of the extraction results to create the categories.
- Create two categories using the little and storage types defined earlier, and score the data.
- Use the visualization tools to determine category overlap, and combine the songs category with the storage category.
- Combine the size category with the little category and rename the little category to small size.
- Combine the following categories into a new category called location: car, exercise, sports by type, outdoor activities, commuting, work.
- Update the modeling node and save the stream as Music Survey with Categories.str.

Workshop 1: Tasks and Results

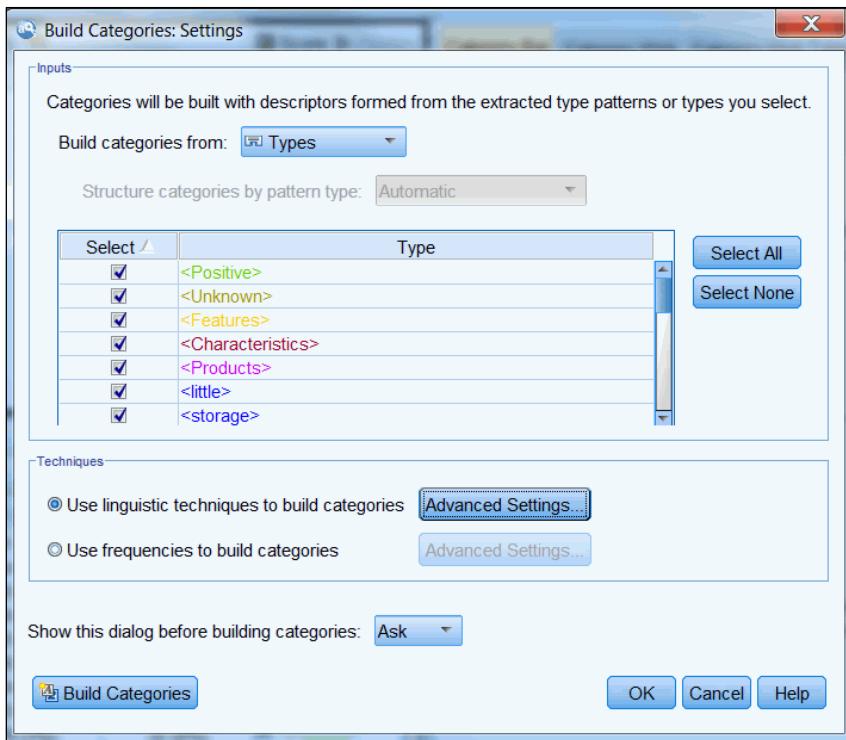
- Open the C:\Train\0A105\12-Creating Categories\Music_survey with Text Link Analysis.str file.
- Run the **Text Mining** node.
- Switch to the **Categories and Concepts** pane.
- Rename the category created using text link analysis, as **Sound Quality**.



The screenshot shows the 'Categories and Concepts' pane in IBM SPSS Modeler. The pane has a toolbar with 'Build', 'Extend', and other icons. Below the toolbar is a table with three columns: 'Category', 'Descriptors', and 'Docs'. A hierarchical tree is displayed under 'Category'. The root node is 'All Documents', which branches into 'Uncategorized' and 'Sound Quality'. 'Uncategorized' further branches into 'No concepts extracted' and two specific concepts: '<fx [sound + excellent]>' and '<fx [sound + good]>'. The 'Descriptors' column shows '-' for most nodes and '2' for 'Sound Quality'. The 'Docs' column shows '405' for 'All Documents' and '-' for the other nodes.

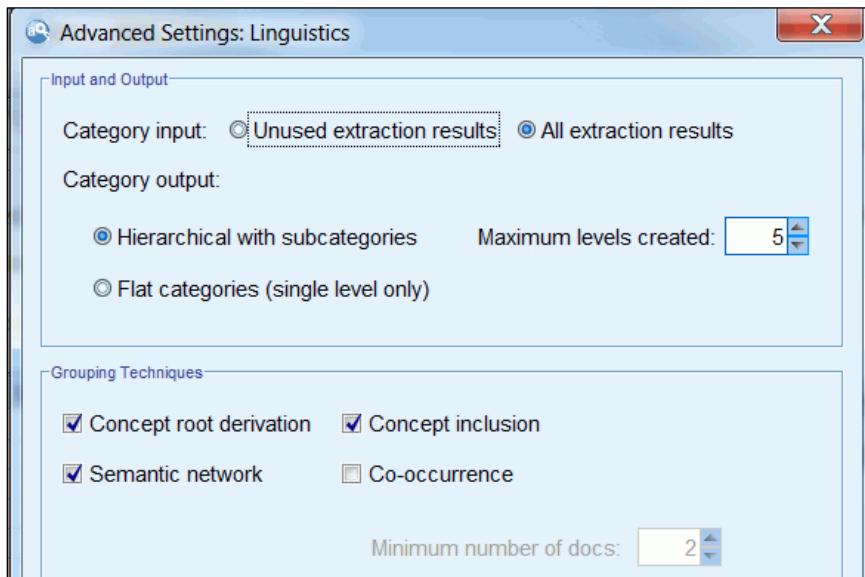
Category	Descriptors	Docs
All Documents	-	405
Uncategorized	-	
No concepts extracted	-	
Sound Quality	2	
<fx [sound + excellent]>		
<fx [sound + good]>		

- Build additional categories using the linguistic categorization methods. Since responses are mainly lists, build the categories from **all types**.



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

- Use the combination of **concept derivation**, **concept inclusion**, and **semantic network** on **all of the extraction results** to create the categories.



- In the Extracted Results pane, select **Type** from the list, and then Ctrl+click **little** and **storage**.
- Click **Create categories for each descriptor** .
- Score the data, and then scroll to the end of the **Category** list.

Category	Descripto...	Docs
+ skipping+<Contextual>	2	2
+ exercise	2	3
+ commuting	2	2
+ playlists	2	7
+ traveling	2	4
+ son	2	3
+ mix	2	2
+ quality	2	2
+ gadgets	2	1
+ unit+<Positive>	1	1
+ <little>	1	61
+ <storage>	1	57

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

- Use the visualization tools to determine which category overlaps with the storage category.
 - In the **Category** pane, click <storage> and then click **Display**.
 - Sort the **Selection %** column in descending order.

Category Bar	Category Web	Category Web Table		
Category	Bar	Selection %	Docs	
<storage>		100.0	57	
music		59.6	34	
songs		26.3	15	
easy		15.8	9	
electronics		15.8	9	
memory de		14.0	8	
cds		14.0	8	
<little>		12.3	7	
tunes		7.0	4	
listening		7.0	4	
place of bu		5.3	3	
car		5.3	3	
Sound Qua		3.5	2	
computer n		3.5	2	
size		1.8	1	
space		1.8	1	

- In the **Category** pane, move **songs** to the **storage** category.
- Score the data.
- Select **storage** and then click **Display**.
- Click the **Selection %** column header to sort in descending order.

Category Bar	Category Web	Category Web Table		
Category	Bar	Selection %	Docs	
<storage>		100.0	71	
music		52.1	37	
memory de		16.9	12	
electronics		14.1	10	
easy		12.7	9	
<little>		11.3	8	
cds		11.3	8	
listening		8.5	6	
place of bu		5.6	4	
tunes		5.6	4	
car		4.2	3	
Sound Qua		2.8	2	
well		2.8	2	

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

- Combine the **size** category with the **little** category and rename the **little** category **small size**.

Category	Bar	Selection %	Docs
small size		100.0	89
easy		10.1	9
music		10.1	9
<storage>		9.0	8
electronics		7.9	7
Sound Qua		5.6	5
listening		4.5	4
memory de		3.4	3
radio		3.4	3
exercise		2.2	2
color		2.2	2
space		1.1	1
tracks		1.1	1

- Finally, combine the categories **car**, **exercise**, **sports by type**, **outdoor activities**, **commuting**, and **work** into a new category called **location**.

Category	Description	Docs
All Documents	-	405
Uncategorized	-	87
No concepts extracted	-	4
small size	4	89
music	26	75
<storage>	4	71
easy	10	62
electronics	10	56
memory device	12	33
listening	3	31
Sound Quality	2	23
location	9	19
design	3	17

- Update the modeling node and save the stream as **Music Survey with Categories.str**.
- Close Modeler.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.



Managing Linguistic Resources

IBM SPSS Modeler Text Analytics (v16)

Business Analytics software



© 2014 IBM Corporation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Objectives

- At the end of this module, you should be able to:
 - use the Template Editor
 - save resource templates
 - understand the relationship between templates and loaded templates
 - understand the difference between local and public libraries
 - publish libraries
 - share libraries and templates

© 2014 IBM Corporation

The resources in IBM SPSS Text Analytics are stored within a project in a set of libraries, and each of the libraries is comprised of several dictionaries. Together, a set of such libraries form a Resource Template, which is simply a collection of linguistic resources for a particular content area or domain. In addition to the visible libraries and their dictionaries, resource templates include compiled resources—such as terms assigned to the Location type—and some advanced resources, such as fuzzy grouping exceptions, anti-links, and regular expression statements.

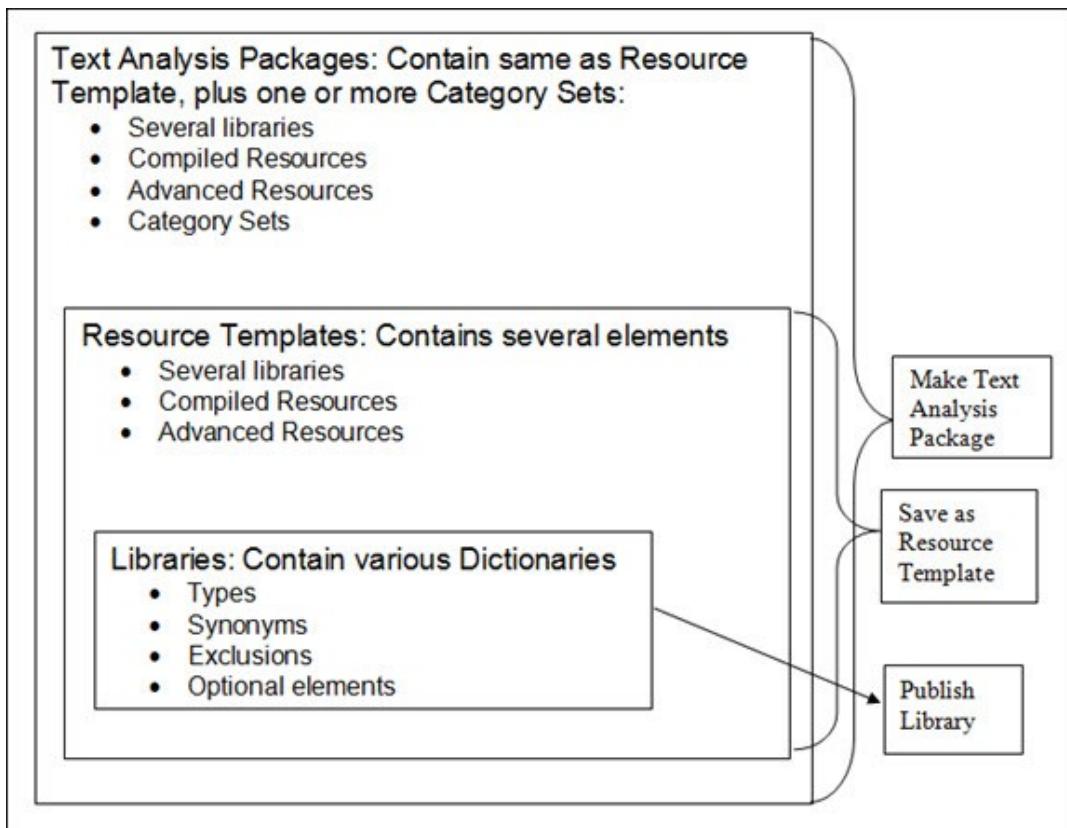
After you fine-tune a set of linguistic resources, you can save both the template and its component libraries. This will allow you to use these in new text-mining projects with similar text data. This module discussed how to accomplish these tasks. The relationship (or lack thereof) between a template and the resources from that template loaded into a text-mining node is also discussed. The ability to share resources and libraries with other users will be discussed, and there will be some notes on backing up and restoring all the resources.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

13-3



A visual depiction of the relationship between libraries, resource templates, and text analysis packages is displayed. Libraries are contained within Resource Templates and TAPs. TAPs contain the equivalent of a Resource Template, plus one or more category sets. Each of these elements can be saved separately.

Template Editor

- Used to edit templates and their resources outside of the Interactive Workbench session.
- Enables you to create and edit resources independent of a specific node or stream.
- Used to create or edit resource templates before loading them into the Text Link Analysis node and the Text Mining modeling node.

© 2014 IBM Corporation



There are two main methods for working with and editing the templates, libraries, and their resources. One is using the Template Editor, which allows you to create and edit templates and the resources they contain independent of a specific node or stream. The other method is using the Resource Editor, accessible within an Interactive Workbench session, which allows you to work with the resources in the context of a specific node and dataset.

The Template Editor can be used to create and edit templates as well as libraries directly, without an Interactive Workbench session. You can use this editor to create or edit templates before loading them into the Text Link Analysis node or the Text Mining modeling node.

Creating a Resource Template

- It is recommended that you do not modify the shipped templates.
- Instead, save your changes to a new template.

© 2014 IBM Corporation



During the data preparation portion of your project, the resources that were loaded into the text mining modeling node were changed to better extract text and create categories specific to your data. If the data you are analyzing are unique, you might not need to save the resources as a template. In contrast, if you think you might possibly want to use these resources on similar data in the future, you should save them as a resource template. Normally you would not want to save back to the Opinions (English) template, as doing so would overwrite that template. The end result would be that these changes would affect all subsequent projects that loaded the Opinions (English) template. Instead, you would choose a new name to create a new template.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

13-6

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Local and Public Libraries

- Libraries that are associated with a particular session or added to a text-mining node are local libraries.
- Libraries that are available to any other modeling node or interactive session are public libraries.
- Public libraries are contained in their own database separate from the Resource Templates database.

© 2014 IBM Corporation

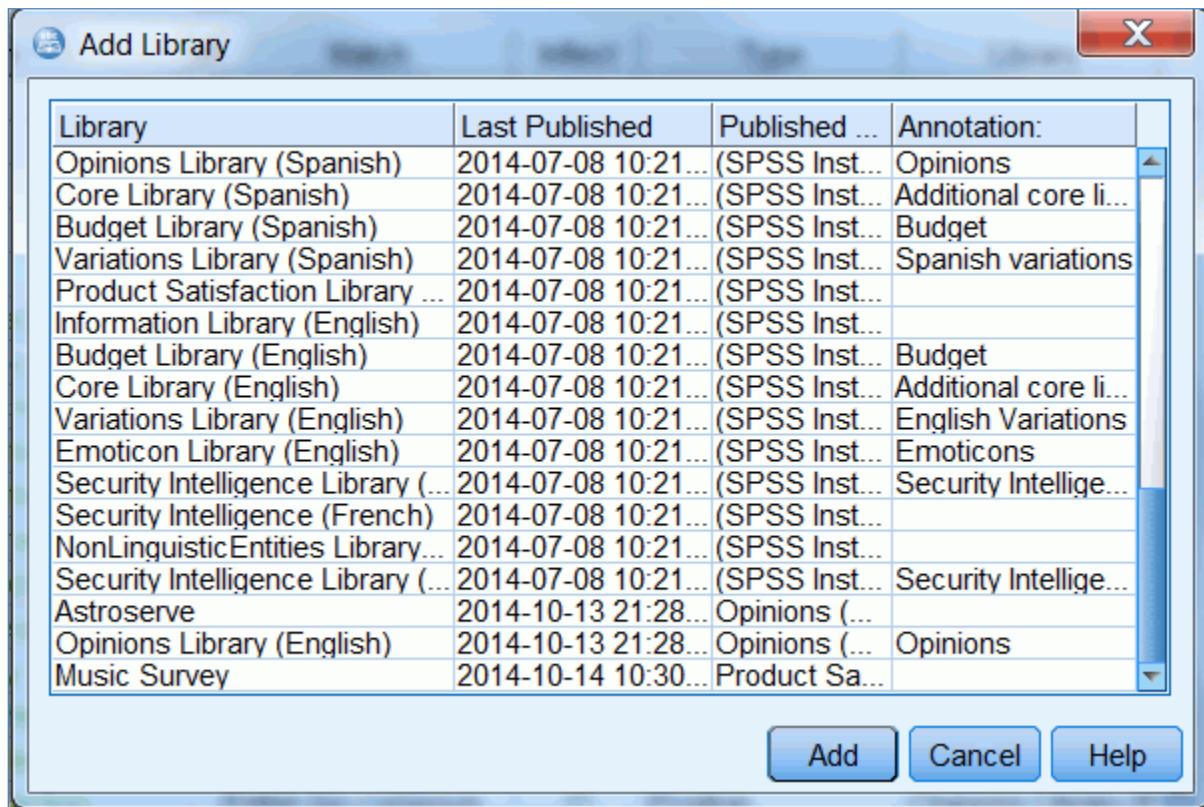


Libraries can exist in two states or versions. Libraries that are associated with a particular session or added to a text-mining node are local libraries. As you added type definitions, synonyms, and excluded terms to the Local library, you were working with a local library. The resources defined were only available to this text mining node. This is also true of the versions of the Opinions, Budget, Core, and Variations libraries that were loaded from the Opinions (English) template: once loaded, they are all now local libraries.

You can make these resources available for use with other text mining nodes by creating a public library version of a library. A public library is available to any other modeling node or interactive session. Public libraries are contained in their own database, separate from the Resource Templates database.

The various libraries supplied with the software are public libraries and can be added to text-mining projects. It is possible to edit the resources in these libraries and then create a new public version. Importantly, if a library of this name was part of one or more resource templates, this new public version would not be used in the templates, unless a set of resources including the modified library were saved as a template. The public libraries are distinct from the templates.

A partial listing of the available public libraries is as follows:



The screenshot shows a Windows-style dialog box titled "Add Library". The main area is a grid table with four columns: "Library", "Last Published", "Published ...", and "Annotation:". The table lists various public libraries, many of which have "(SPSS Inst..." in the "Published ..." column. The "Annotation:" column contains brief descriptions like "Opinions", "Additional core li...", "Budget", etc. At the bottom of the dialog are three buttons: "Add", "Cancel", and "Help".

Library	Last Published	Published ...	Annotation:
Opinions Library (Spanish)	2014-07-08 10:21...	(SPSS Inst...	Opinions
Core Library (Spanish)	2014-07-08 10:21...	(SPSS Inst...	Additional core li...
Budget Library (Spanish)	2014-07-08 10:21...	(SPSS Inst...	Budget
Variations Library (Spanish)	2014-07-08 10:21...	(SPSS Inst...	Spanish variations
Product Satisfaction Library ...	2014-07-08 10:21...	(SPSS Inst...	
Information Library (English)	2014-07-08 10:21...	(SPSS Inst...	
Budget Library (English)	2014-07-08 10:21...	(SPSS Inst...	Budget
Core Library (English)	2014-07-08 10:21...	(SPSS Inst...	Additional core li...
Variations Library (English)	2014-07-08 10:21...	(SPSS Inst...	English Variations
Emoticon Library (English)	2014-07-08 10:21...	(SPSS Inst...	Emoticons
Security Intelligence Library (...)	2014-07-08 10:21...	(SPSS Inst...	Security Intellige...
Security Intelligence (French)	2014-07-08 10:21...	(SPSS Inst...	
NonLinguisticEntities Library...	2014-07-08 10:21...	(SPSS Inst...	
Security Intelligence Library (...)	2014-07-08 10:21...	(SPSS Inst...	Security Intellige...
Astroserve	2014-10-13 21:28...	Opinions (...)	
Opinions Library (English)	2014-10-13 21:28...	Opinions (...)	Opinions
Music Survey	2014-10-14 10:30...	Product Sa...	

The libraries here show all the public libraries in your version of IBM SPSS Modeler Text Analytics. A public library is one that has been published, or stored, in a database that is available to any of your projects. Notice that the Budget Library, Core Library, Opinions Library, and Variations Library are all listed as public libraries, separately from the Opinions (English) template in which versions of them are stored. Although these public libraries resemble the libraries contained within the template, the templates have been specially tuned to particular application areas and contain additional advanced resources. Therefore, it is recommended that if you are working with, for example, opinion/survey text data, genomics, or security intelligence data, you should use the templates for these rather than adding individual libraries to a more generic template.

Publishing Libraries

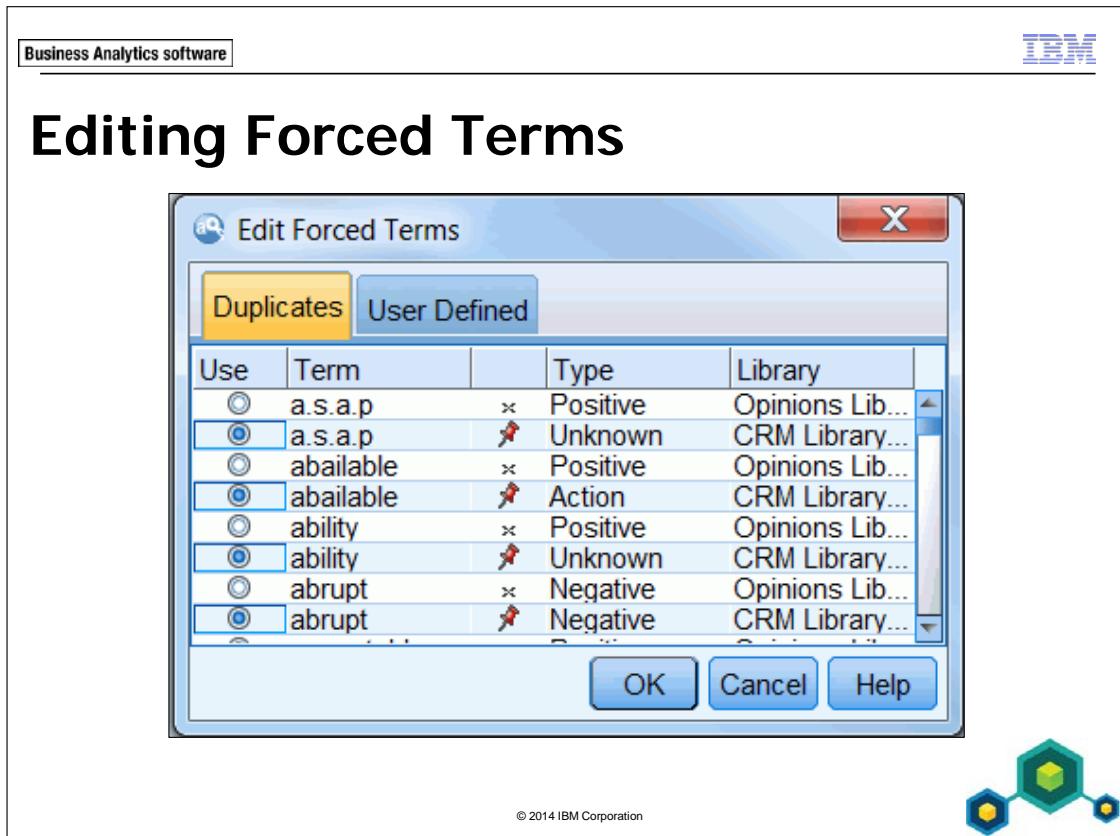
- Publishing a library creates a public copy of the local library. Until it is published, the library remains local to the current session and cannot be used by any other project.
- A public version of the local library cannot be created with its default name.
- You must rename the local library before you can publish it.

© 2014 IBM Corporation



You can publish the local libraries to make them available to all projects. Remember from the important note above that published libraries are not part of a template just because they have been published. The collection of resources in a Resource Editor session can be saved as a template, and then those libraries, whether published or not, will be part of a template that can be used for other projects.

When you add a library, if there are conflicts with definitions in an existing library a Resolve Conflicts dialog box will appear.



The Edit Forced Terms dialog box contains each pair of conflicting terms or types. Alternating background colors are used to visually distinguish each pair. The Duplicates tab contains the duplicated terms found in the libraries, their type, and library location. If a pushpin icon appears after a term, it means that this occurrence of the term has been forced to that definition. If a black X icon appears, it means that this occurrence of the term will be ignored during extraction because it has been forced elsewhere.

In this example, because the CRM library is being added, notice that by default Modeler has assigned all the terms in conflict to the CRM library. To make changes, select the radio button in the Use column for the term that you want to force.

The pushpin symbol in a term assignment for a type, such as for upgraded, means that this term has been forced to this assignment. If you decide to reset an assignment, you can do so from here by double-clicking the pushpin for a term. A Resolve Conflicts dialog will appear that is essentially identical to the Edit Forced Terms dialog, and you can follow the same procedure.



Publishing a library creates a public copy of the local library. If there is an existing public library with the same name as the local version, the contents of the local library replace the existing public version. All libraries are stored in a specific location defined at the time of program installation. Therefore, you only need to specify a library name when publishing, but not a folder location.

A public version of the local library cannot be created with its default name. It must first be renamed; then it can be published.

In addition to the library name, there is information about when the library was last published, from which project it was published, and how many types, terms, excluded words, patterns, and synonyms it contains. There is also a note that tells you that the local library has never been published, with an icon indicative of that status (see more on these icons below).

By default libraries that have never been published, and libraries that have more recent changes in the local version compared to the published version (determined by their date), will be checked to be published. In this example, the Astroserve library has been checked to be published. By clicking the Publish button, the library will be available for other text mining sessions.

When you work with the local version of a public library, the two versions will become desynchronized as you modify the local version. When this occurs, you can synchronize them again. The program shows the synchronization status of a local library with an icon next to the library name in the tree view of the Libraries pane. The synchronization states and associated icons follow:

Icon	Local Library Status
	Unpublished: the local library has never been published.
	Newer: the local library version is more recent than the public version. You can republish the local version to the public version.
	Out of date: the local library version is older than the more recent public version. You can update the local version with the differences.
	Synchronized: the local and public library versions are identical.
	Out of sync: Both the local and public libraries contain changes that the other does not. You must decide whether to update or publish the local library. If you update, you will lose the changes that you made since the last time you updated or published. If you choose to publish, you will overwrite the changes in the public version.

When a public version of a library is out of sync with a local version, you should receive a message when you launch an Interactive Workbench session, asking if you want to synchronize any libraries that need updating or republishing.

Sharing Libraries

- The Manage Libraries dialog allows you to do the following:
 - export libraries so they can be shared with other users
 - import libraries that other users have exported

© 2014 IBM Corporation



As was learned, publishing libraries makes them available to other projects on the computer on which the copy of the software is installed. If you want to share libraries with users on other computers, you need to export them from here.

Exported libraries are stored in a proprietary format (with an extension of .lib). When you export, you will be asked for a folder location and name. The file name will be taken from the library name.

Once this file has been created, you can email it or send it by other convenient means to colleagues. They can then import it into their software (so that it will become a public library), and then add it to any projects that would benefit from those resources.

Managing Resource Templates

- The Manage Template dialog allows you to do the following:
 - rename templates
 - delete templates
 - import or export templates

© 2014 IBM Corporation



As with the libraries, there are some basic tasks you can perform to manage templates, such as renaming or deleting them. The Manage Template dialog provides this functionality.

Unlike public libraries, templates can be renamed directly. There is an Export button to write out (export) a resource template as a file with an extension of .lrt. This file can be sent to other users so that a complete set of resources can be shared between users. Exported resource templates can be imported into this installation of IBM SPSS Text Analytics.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

13-14

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Backing Up Resources

- Interactive Workbench offers menu selections to backup the linguistic resources.
- When you restore a database, the entire contents of the resource database will be erased and only the backed up resources will be available.

© 2014 IBM Corporation



You should back up the files on a computer on a regular basis. The Interactive Workbench offers menu selections to make backing up the linguistic resources rather painless.

In addition to backing up the resources for the usual security reasons, you may want to back up the resources when uninstalling and reinstalling the product. A reinstall will write over the existing resources.

When you back up the entire resource database, the templates and libraries are written out into a mySQL database with an extension .tmb. This file can be used to restore the database contents later. Resources can be backed up and restored from the Resources / Backup Tools menu choices.

It is important to note that if you have updated a text mining modeling node with resources from an interactive session, those resources will still be available even if the template from which they were created was deleted, in a restore or otherwise. The resources stored in a node are no longer linked to the template or libraries from which they came. This also means that nodes can be saved and sent to other users for scoring or even additional editing of the resources.

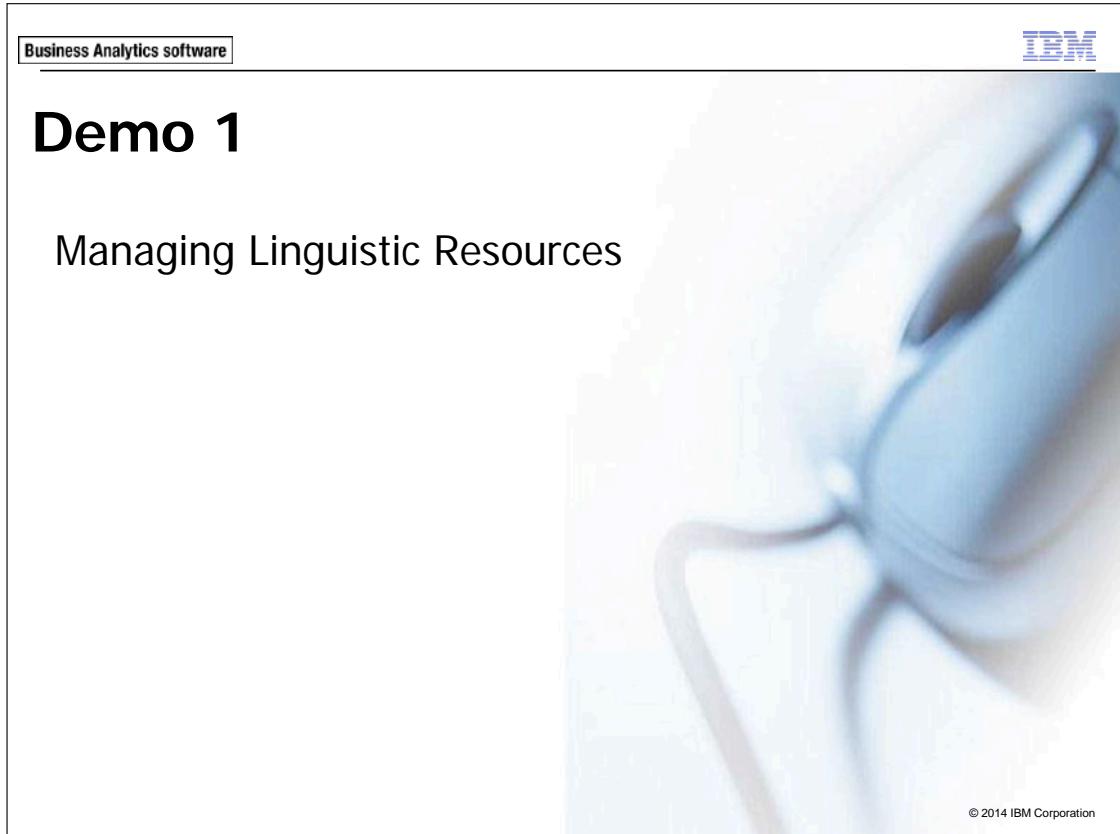
If for some reason you need to restore one or more of the shipped libraries, you can delete one of these libraries from the Manage Libraries dialog. Upon doing so, you will get a dialog stating that the original shipped library has been reinitialized and added.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

13-15



The slide features a large, semi-transparent background image of a person wearing a headset, likely a call center operator. In the top left corner, there is a small rectangular box containing the text "Business Analytics software". In the top right corner, the IBM logo is displayed. The main title "Demo 1" is centered at the top in a large, bold, black font. Below the title, the subtitle "Managing Linguistic Resources" is also centered in a smaller, black font. At the bottom right of the slide, there is a small copyright notice: "© 2014 IBM Corporation".

The following file(s) are used in this demo:

- Managing Linguistic Resources_demo1_start.str - a Modeler stream that reads a file containing call center data for March and April

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

13-16

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

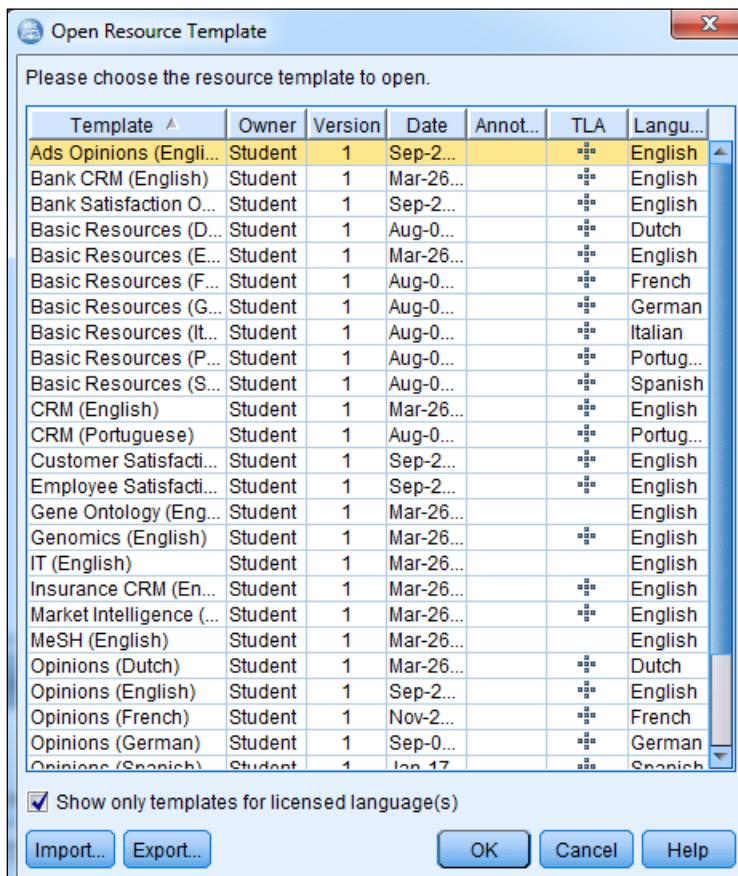
Demo 1: Managing Linguistic Resources

Purpose:

The text mining process is long and laborious. In all likelihood you will want to reuse the resources you just created in other projects, make additional modifications to them, and share them with other people. You will explore how to do this.

Task 1. Use the Template Editor.

1. From the **File** menu, click **Open Stream**.
2. Navigate to **C:\Train\0A105\13-Managing_Linguistic_Resources**, and then double-click **Managing Linguistic Resources_demo1_start.str**.
3. From the **Tools** menu, click **Text Analytics Template Editor**.



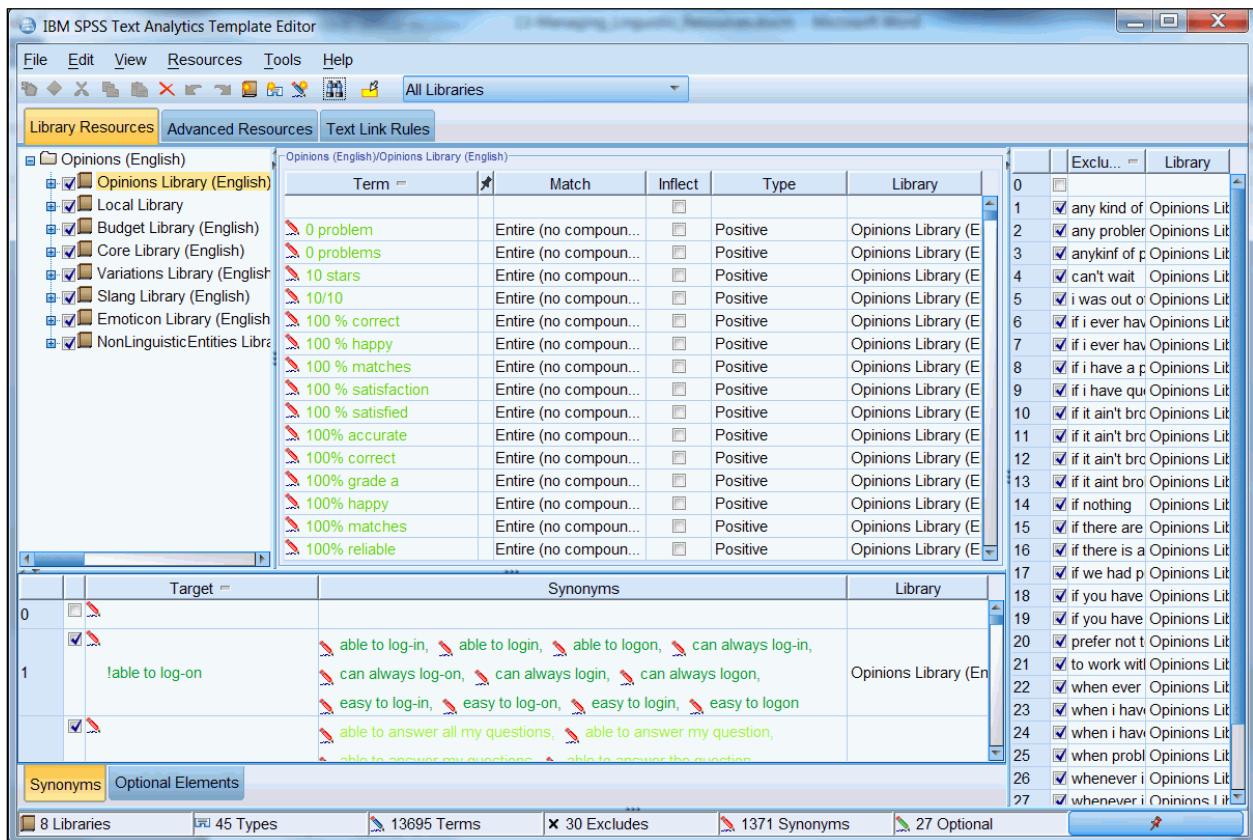
All the current templates for this particular installation of Modeler are displayed, along with the date the template was created, its language, and whether it contains TLA patterns. Templates can be imported or exported from here (see later discussion in this demonstration)

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

4. Click **Opinions (English)** template, and then click **OK**.



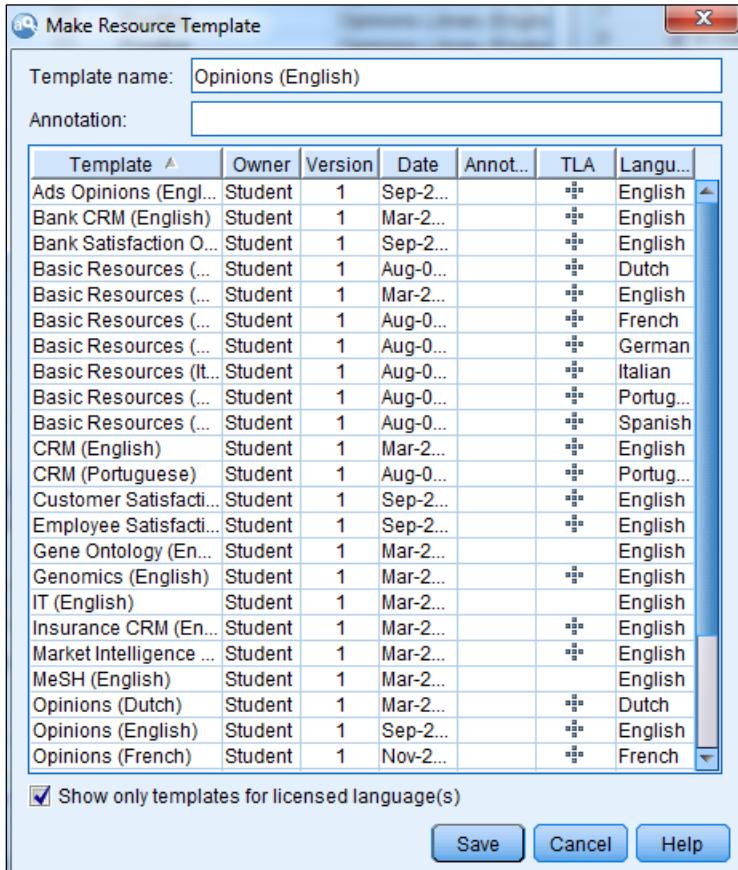
When you select a template and click OK, a Resource Editor session opens. From here, you can make changes to the resources as desired, then save them. Note that there is no list in the upper right corner from which to select different views, specifically Categories and Concepts. That is because this session is not tied to a particular set and its extraction.

In this case, the resources have already been edited and the query modeling node has been updated so there is no need to redo the work here. Moreover, you normally do not want to change to Opinions (English) template anyway, and if you do, you should probably save the changes to a template with a new name.

5. From the **File** menu, click **Close** to exit from the **Resource Editor**.

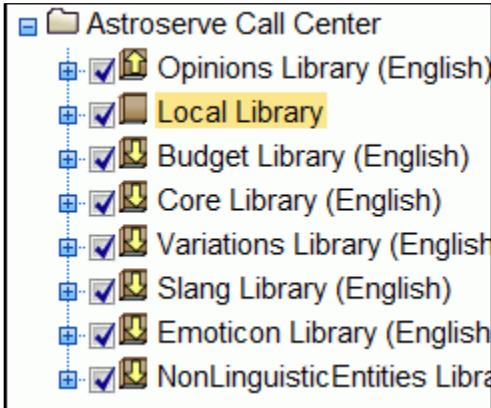
Task 2. Save the Resource Template.

1. Run the **Text Mining** node labeled **query**.
2. After the extraction is complete, switch to the **Resource Editor** view.
3. From the **Resources** menu, click **Make Resource Template**.



The current list of templates for this particular installation of Modeler appears along with the date the template was created and whether it contains TLA patterns.

4. Change the template name to **Astroserve Call Center**, and then click **Save**.
Saving the resources as a template essentially bundles them together as a linked set of libraries a template and advanced resources.
The name of the template from which the resources were loaded changed in the Libraries pane.



Notice that it was not stated that the name of the template you were using changed. This distinction is not semantic. When you are working in the Resource Editor, you are not working on a template; instead, you are working with a copy of the components, or resources, contained in that template that was last loaded. As a result, changes here do not affect the template from which the resources were copied or loaded, unless you resave the template.

Task 3. Use a saved resource template.

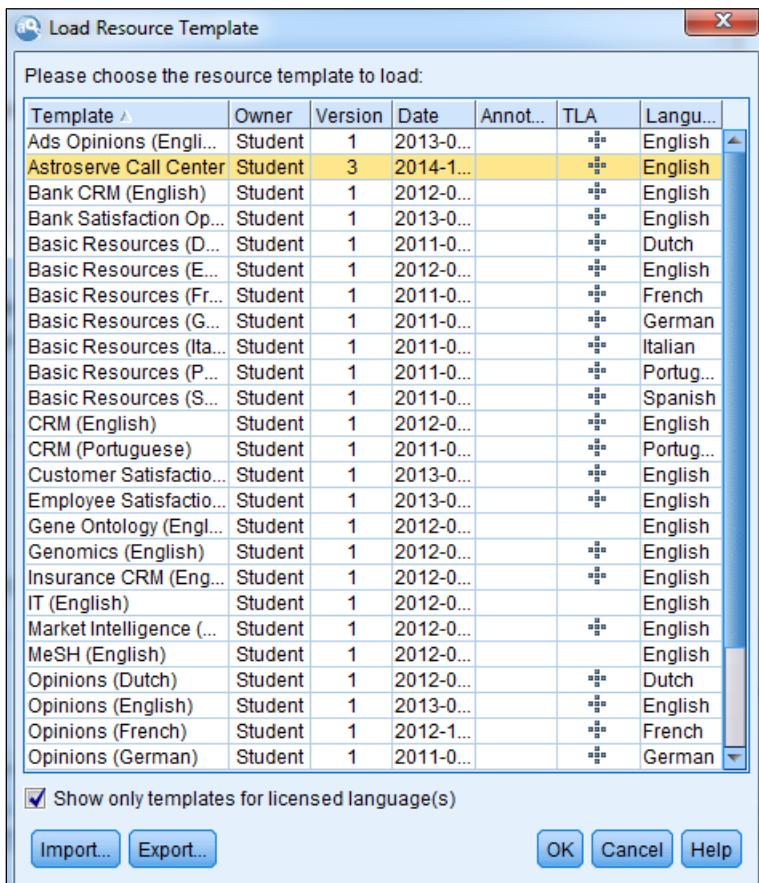
Now that you have saved the linguistic resources as a template, you can load it in Text Mining nodes just the same way you load any of the shipped templates.

1. From the **File** menu, click **Close**, and then click **Close**.

This action keeps the interactive session open but closes the Interactive Workbench window. This is convenient, but it does not release the memory required to manage the session. This is done here so you do not have to re-run the query text mining node.

2. Add a **Text Mining** node to the stream.
3. Connect the **Sample** node to the new **Text Mining** node.
5. Edit the new **Text Mining** node.
6. Beside **Text field**, select **query**.
7. Click the **Model** tab.

8. Click the **Load** button.



The resource template just saved is now available for use. Notice that it has a symbol in the TLA column indicating that it contains TLA patterns. This is because it was created with the resources contained in the Opinions (English) template, which contained TLA patterns.

9. Click the **Astroserve Call Center** template.
 10. Click **OK**, and click **Run**.

11. After the Interactive Workbench opens and the extraction is complete, switch to the **Resource Editor** view.

Target =	Synonyms	Library
0	able to log-in, able to login, able to logon, can always log-in,	
1	can always log-on, can always login, can always logon, easy to log-in, easy to log-on, easy to login, easy to logon able to answer all my questions, able to answer my question, able to answer my question, able to answer the question	Opinions Library (En)

All the additions and changes made to the resources are reflected in the newly loaded template.

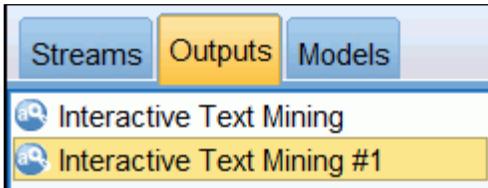
12. From the **File** menu, click **Close**, and then click **Close** again.

Task 4. Review the relationship between saved and loaded templates.

Currently the resources of the Astroserve Call Center template are being used in two text mining nodes. Return to the first open Interactive Workbench session and make a change to the resources.

1. Switch to the **Modeler stream** canvas.

2. In the **Manager** area, click the **Outputs** tab.



3. Double-click **Interactive Text Mining #1**.

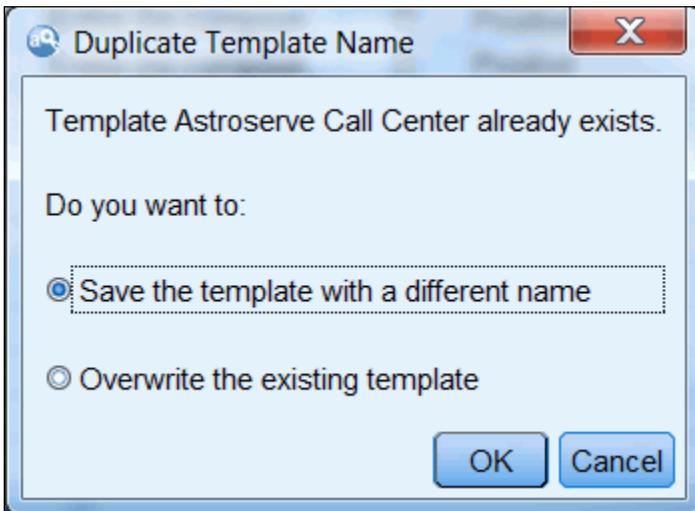
The session reopens, looking identical to the already open session (although the order of the terms may be slightly different because resources are saved and listed in alphabetical order).

You will add another excluded term, "sales", with a wildcard asterisk.

4. Click the **blank cell** in the **Exclude List** column.
5. Add the term **sales *** and then press **Enter**.

	Exclu... ▾	Library
0	<input type="checkbox"/>	
1	<input checked="" type="checkbox"/> sales *	Opinions Lib
2	<input checked="" type="checkbox"/> any kind of	Opinions Lib
3	<input checked="" type="checkbox"/> any problер	Opinions Lib
4	<input checked="" type="checkbox"/> anykinf of p	Opinions Lib
5	<input checked="" type="checkbox"/> can't wait	Opinions Lib
6	<input checked="" type="checkbox"/> i was out o	Opinions Lib

6. From the **Resources** menu, click **Make Resource Template**.
7. Click **Astroserve Call Center**, and then click **Save**.



8. Select **Overwrite the existing template**, and then click **OK**.

The resources have once more been bundled together as the Astroserve Call Center template. Is the change reflected in the second interactive session, since it is using resources from this template? Switch to the other session to see.

9. From the **File** menu, click **Close**, and then click **Close** again.

10. In the **Manager** pane, double-click **Interactive Text Mining**.

The extra excluded term is not included. This is because the resources in the Astroserve Call Center template were loaded before the change was made.

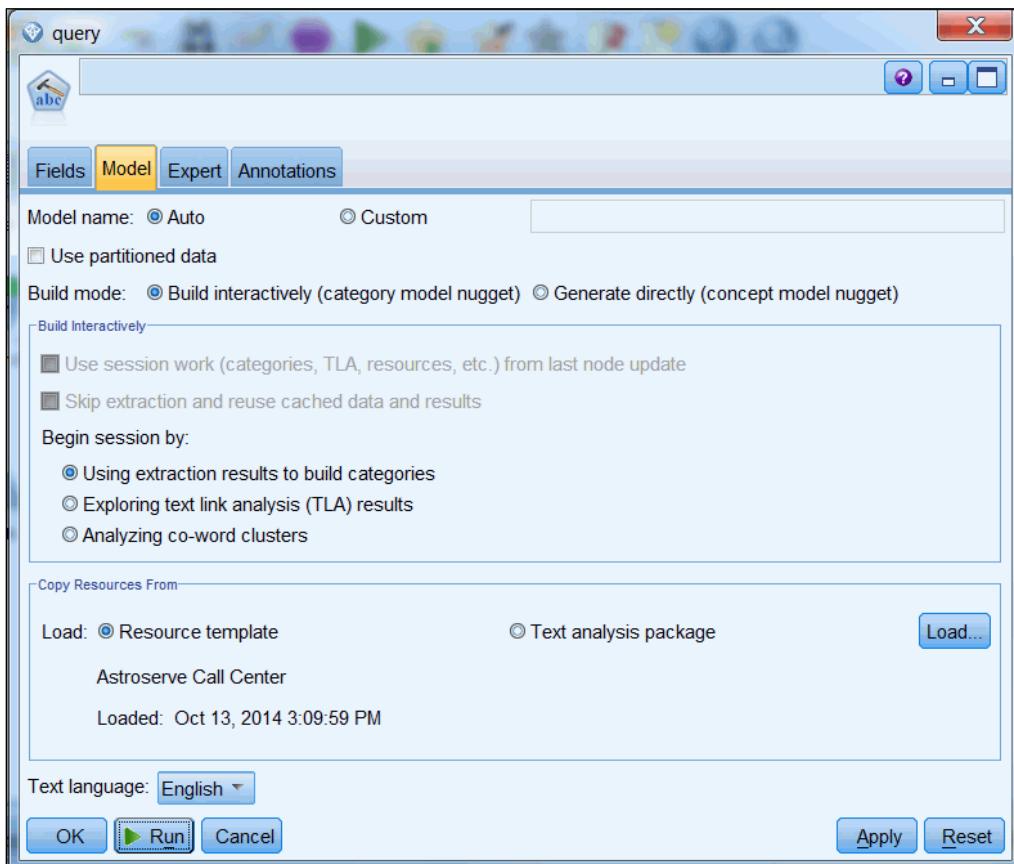
	Exclu...	Library
0	<input type="checkbox"/>	
1	<input checked="" type="checkbox"/>	any kind of Opinions Lib
2	<input checked="" type="checkbox"/>	any problер Opinions Lib
3	<input checked="" type="checkbox"/>	anykinf of p Opinions Lib
4	<input checked="" type="checkbox"/>	can't wait Opinions Lib
5	<input checked="" type="checkbox"/>	i was out o Opinions Lib
6	<input checked="" type="checkbox"/>	if i ever hav Opinions Lib
7	<input checked="" type="checkbox"/>	if i ever hav Opinions Lib

A key point for consideration is this: if you close this interactive session, and re-run the modeling node, will the change in the template be reflected in the resources?

11. From the **File** menu, click **Close**.

12. Click **Exit** so that nothing is saved.

13. Edit the second **Text Mining** node that was added to the stream (be sure to select the correct one).



The dialog box lists the Astroserve Call Center template as being used, or loaded. What this really means, since this node has previously been run, is that the resources from the template were loaded into this node to be used in modeling when you originally clicked the Load button. The resources were taken from the state of the Astroserve Call Center template on that date and time (this information appears for reference and as a reminder of this fact).

To make this clear, you will run the node.

14. Click **Run**.

15. Switch to the **Resource Editor** view.

	Exclu... =	Library
0	<input type="checkbox"/>	
1	<input checked="" type="checkbox"/>	any kind of Opinions Lib
2	<input checked="" type="checkbox"/>	any proble Opinions Lib
3	<input checked="" type="checkbox"/>	anykinf of p Opinions Lib
4	<input checked="" type="checkbox"/>	can't wait Opinions Lib
5	<input checked="" type="checkbox"/>	i was out o Opinions Lib
6	<input checked="" type="checkbox"/>	if i ever hav Opinions Lib
7	<input checked="" type="checkbox"/>	if i ever hav Opinions Lib

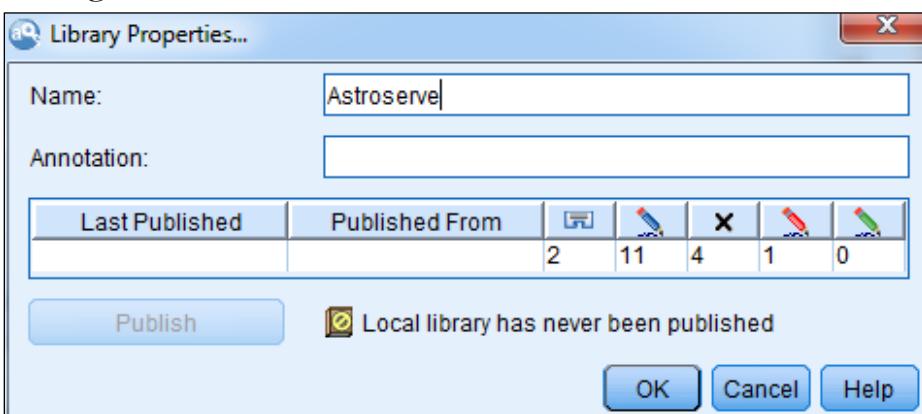
As you can observe, the Excluded terms still do not include **sales***, even though that term is currently in the Astroserve Call Center template. The change in the template is not reflected in the resources in the second modeling node. This is because those resources are loaded only once, when the template was first selected. There is no connection to the saved template that has been loaded, so the resources are not loaded again when the node is re-run.

Thus, to get the benefit of changes to the template, the resource template must be reloaded. This is true even when you are using only one text mining node, and a template is saved in an interactive session begun from that node. (The alternative is to update the modeling node, which adds the resources directly to the node; then a resource template does not need to be re-loaded for those resources).

Task 5. Publish libraries.

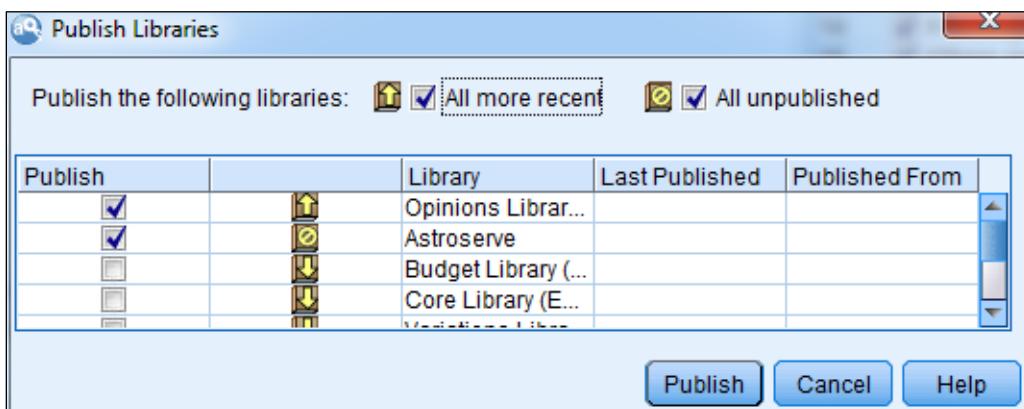
A public version of the Local library cannot be created with its default name. You want to publish our own local library.

1. Right-click the **Local Library** and select **Library Properties**.
2. Change the **Name** to **Astroserve**.



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

3. Click **OK**.
4. From the **Resources** menu, click **Publish Libraries**.

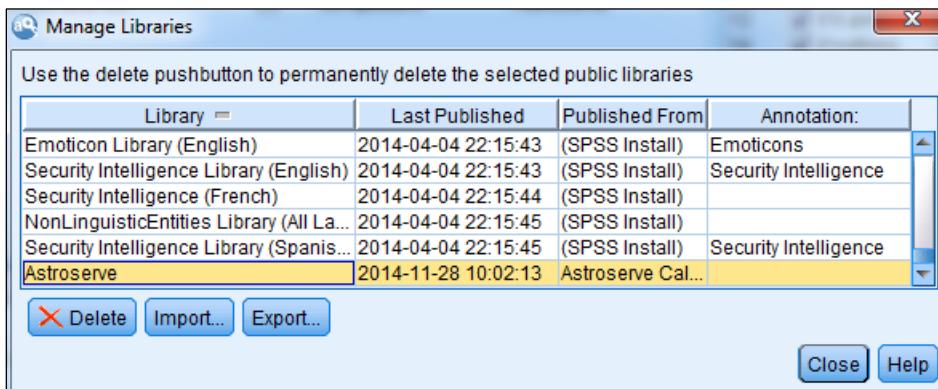


5. Clear the **Opinions Library (English)** check box, and then click **Publish**.
The Astroserve library is now available for other text mining sessions.

Task 6. Share libraries.

Libraries can only be shared with other users after they have been exported to a file. You want to export the Astroserve library.

1. From the **Resources** menu, click **Manage Libraries**.
2. Click the **Astroserve** library.



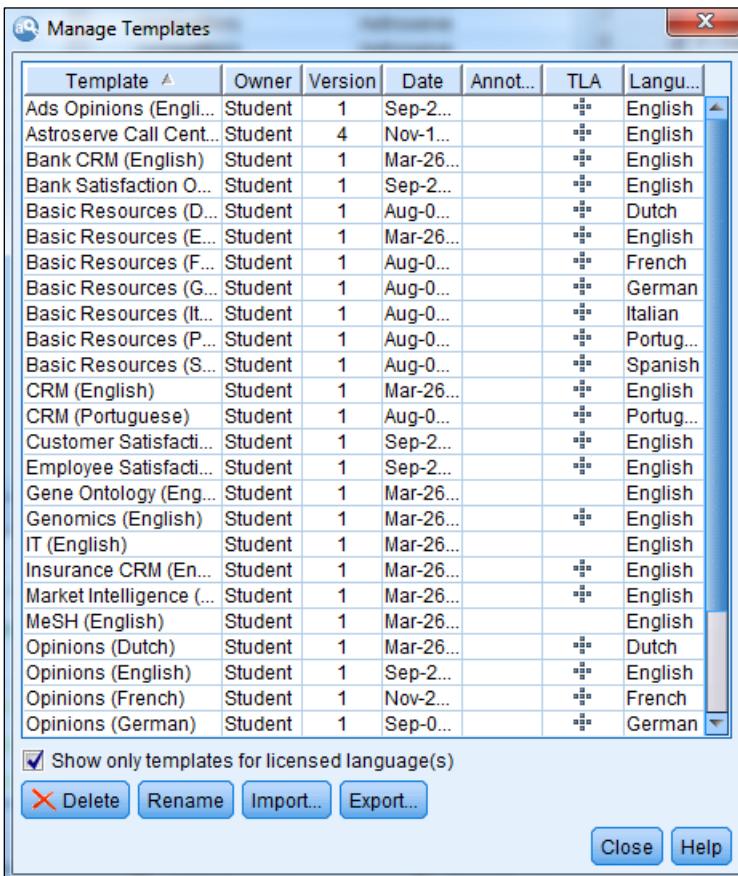
3. Click **Export**.
4. Navigate to **C:\Train\0A105\13-Managing_Linguistic_Resources**, and then click **Export**.
5. Click **Close** to close the **Manage Libraries** dialog.

The name of the exported library in this example is Astroserve.lib. Once this file has been created, you can send it to colleagues.

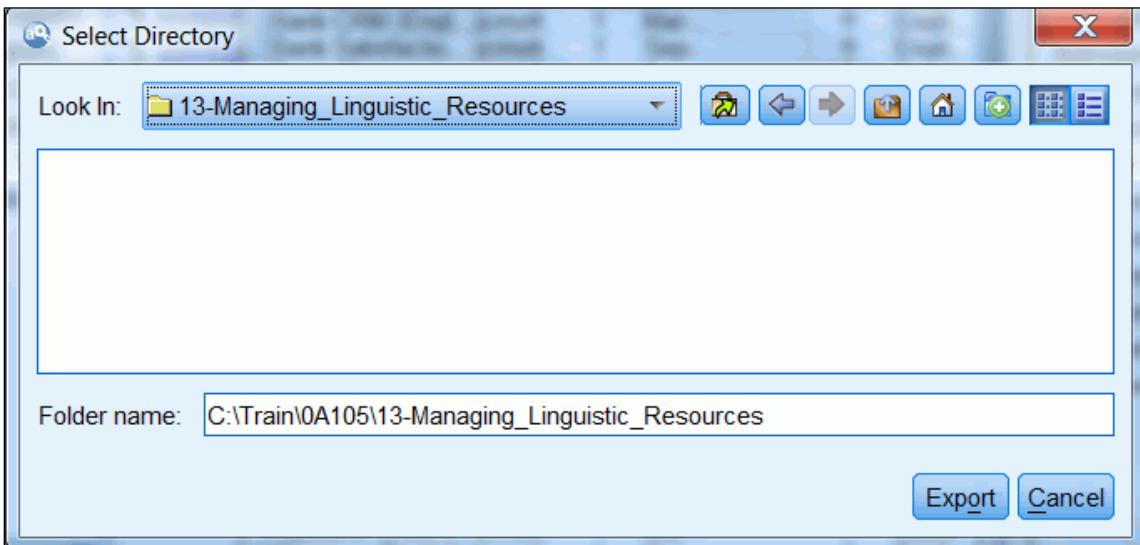
Task 7. Export resource templates.

As with the libraries, there are some basic tasks you can perform to manage templates, such as renaming or deleting them.

- From the **Resources** tab, click **Manage Resource Templates**.



- Click the **Astroserve Call Center** template, and then click **Export**.



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

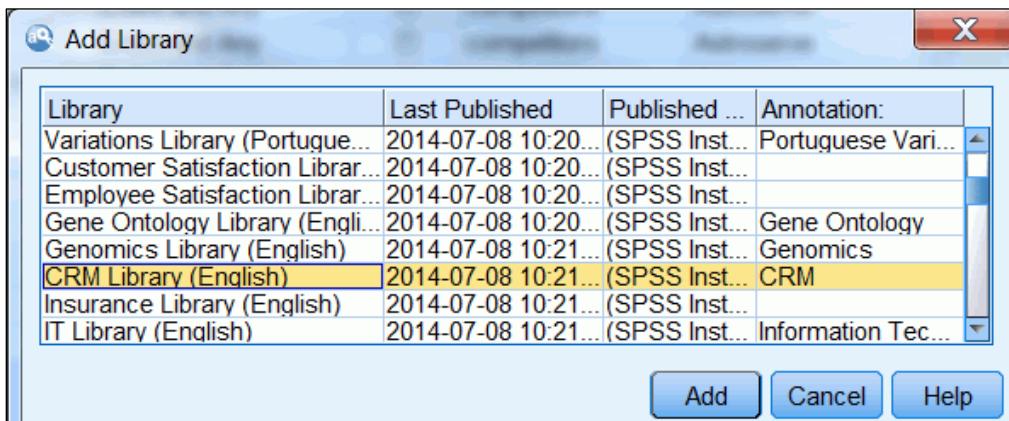
3. Navigate to the **C:\Train\0A105\13-Managing_Linguistic_Resources** folder (if necessary), and then click **Export**.

The file name will be taken from the resource template name. In this case, the file will be called Astroserve Call Center.lrt.

4. Click **Close**.

Task 8. Add a library.

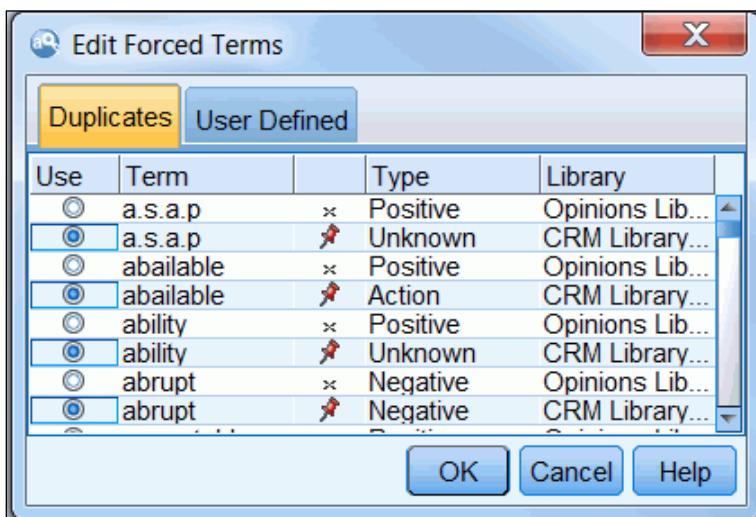
1. From the **Resources** menu, click **Add Library**.
2. Click **CRM Library (English)**.



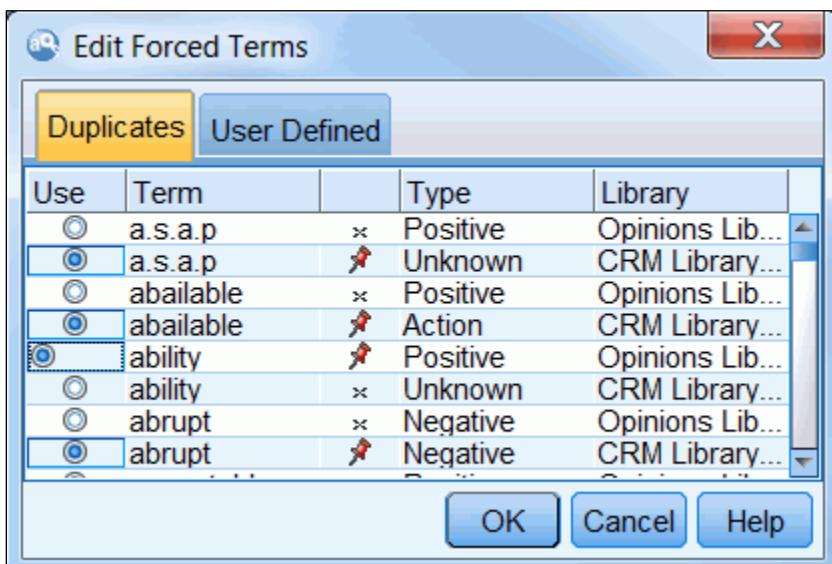
3. Click **Add**.

Rather than seeing the library immediately added to the Library pane, an Edit Forced Terms dialog box opens. This dialog box displays terms that are defined in the library being added and in the existing libraries in the workbench.

Because there must be an unambiguous definition of a term or type, the dialog enables you to resolve these conflicts.



4. Select the radio button to the left of **ability** to force the term to be typed as **Positive**.

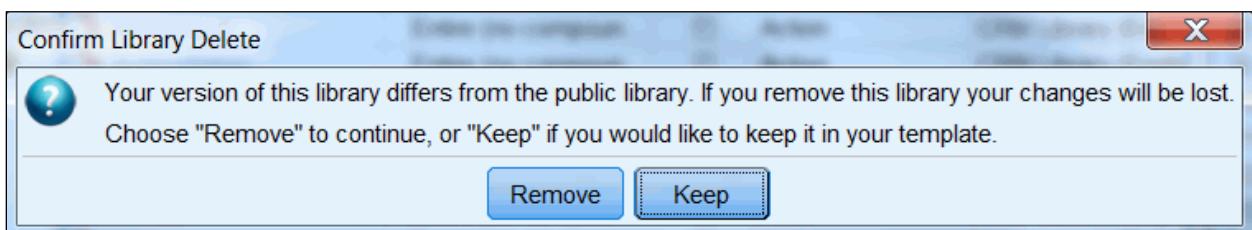


5. Click **OK**.

The pushpin symbol in the term assignment for a type, means that the term "ability" has been forced to this assignment.

You will remove the CRM library.

6. If necessary, from the list on the toolbar, select **All Libraries**.
7. In the Library tree pane, click the **CRM Library**.
8. From the **Edit** menu, click **Delete**.



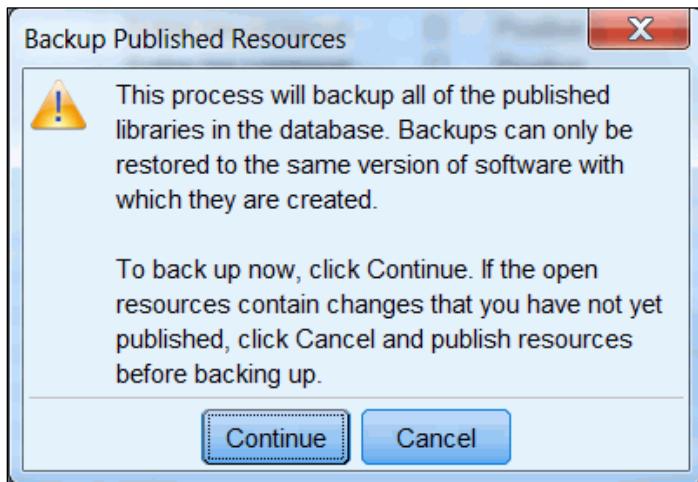
The CRM library is not immediately deleted. A warning message appears that says that the local version of the CRM library differs from the public version (because of the forced terms). If you were to remove the CRM library, then any changes made to it will be lost, because if you reload the public version, it will not have the local edits. Of course, here no real local changes have been made that cannot be duplicated, so you are safe to continue the deletion.

9. Click **Remove**.

Task 9. Back up resources.

You want to back up the Astroserve resources.

- From the **Resources** menu, point to **Backup Tools**, and then click **Backup Resources**.

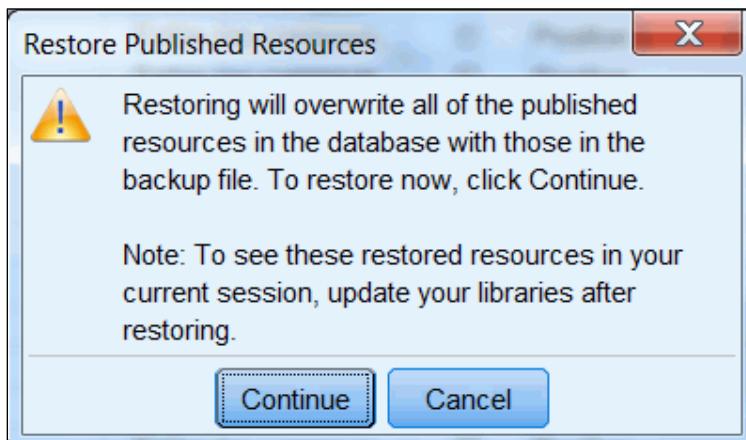


For this exercise, you will not actually go through the process but this is the message you would see if you did want to back up your resources.

- Click **Cancel**.

To restore a database, you would do the following steps.

- From the **Resources** menu, point to **Backup Tools**, and then click **Restore Resources**.



Because you do not have any resources to restore, you will cancel the process.

4. Click **Cancel**.

You can now end the interactive session and exit from Modeler.

5. From the **File** menu, click **Close**, and then click **Exit** to end the Interactive Workbench session.

6. From the **File** menu, click **Exit**, and then click **Exit** to end the Modeler session.

Results:

You have successfully managed your linguistic resources so they can be shared with other people and used in other projects. In addition, you now understand the difference between local versus public libraries.

Apply Your Knowledge

Purpose:

Test your knowledge of the material covered in this module.

Question 1: True or False: Linguistic resources can be shared between projects and users on different computers.

- A. True
- B. False

Question 2: True or False: Local and public libraries are always synchronized.

- A. True
- B. False

Question 3: True or False: It is possible to publish a library called Local Library.

- A. True
- B. False

Question 4: True or False: Changes you make to the shipped templates will affect subsequent projects even if you do not publish them.

- A. True
- B. False

Question 5: True or False: It is possible to assign individual terms more than one type.

- A. True
- B. False

Apply Your Knowledge - Solutions

Answer 1: A. True

Answer 2: B. False

Answer 3: B. False. The Local Library must be renamed before it can be published.

Answer 4: B. False. Changes made to any library remain local unless you publish the library.

Answer 5: B. False. Each term can only be assigned one type. If there is a conflict, you must decide on which type takes precedence.

Summary

- At the end of this module, you should be able to:
 - use the Template Editor
 - save resource templates
 - understand the relationship between templates and loaded templates
 - understand the difference between local and public libraries
 - publish libraries
 - share libraries and templates

© 2014 IBM Corporation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

13-35

Business Analytics software



Workshop 1

Managing Linguistic Resources



© 2014 IBM Corporation

The following file will be used:

- Music Survey with Categories.str - a Modeler stream that reads from a file containing customer likes and dislikes about a portable music player

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Workshop 1: Managing Linguistic Resources

In order to apply the resources you developed to subsequent projects, it is necessary to create public versions of the resources that can be shared with other projects on your own machine, and with other users of IBM SPSS Text Analytics on other computers.

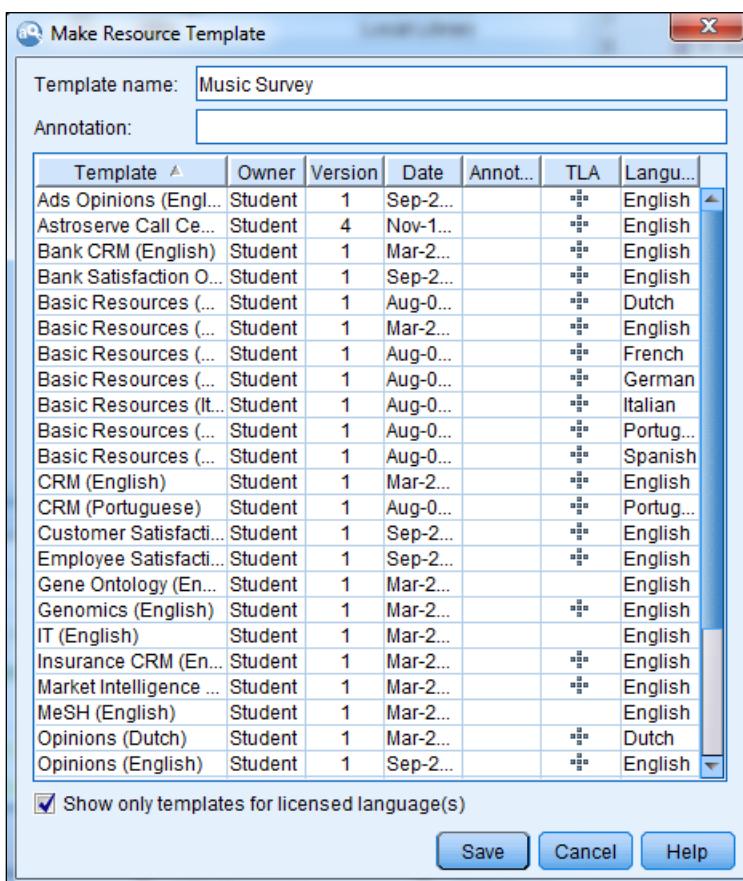
Starting with the Music_Survey with Categories stream (found in C:\Train\0A105\13-Managing_Linguistic_Resources), perform the following tasks:

- Create a Resource Template from the Music Survey with Categories.str stream.
- Create a TAP from the Music Survey with Categories.str stream.
- Rename the local library as Music Survey, and then publish it.

Workshop 1: Tasks and Results

Task 1. Make a resource template from the Music Survey stream.

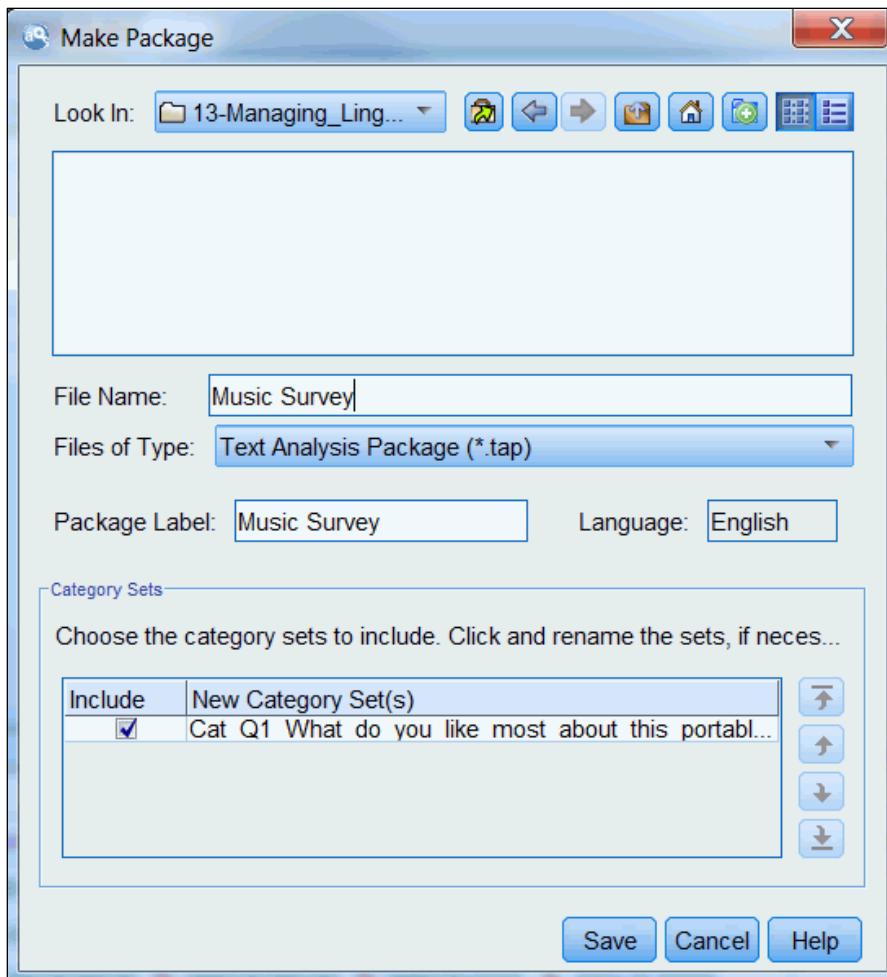
- Open the **C:\Train\0A105\13-Managing_Linguistic_Resources\Music_survey with Categories.str** stream.
- Run the **Text Mining** node.
- In the **Resource Editor**, from the **Resources** menu, click **Make Resource Template**, and then change the template name to **Music Survey**.



- Click **Save**.

Task 2. Create a TAP from the Music Survey stream.

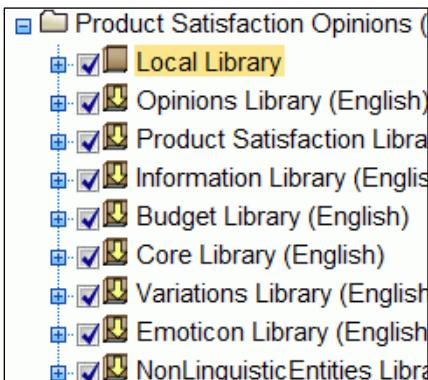
- From the **File** menu, point to **Text Analysis Package**, and then click **Make Package**.
- Navigate to **C:\Train\0A105\13-Managing_Linguistic_Resources**.
- Type **Music Survey** in the **File Name** box.



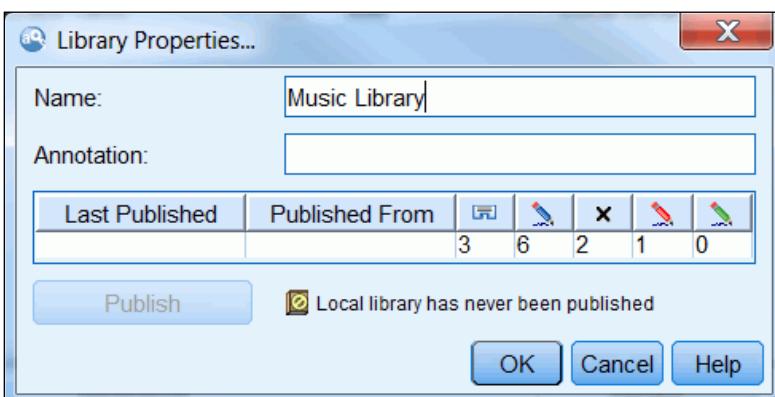
- Click **Save**.

Task 3. Publish a public version of the local library.

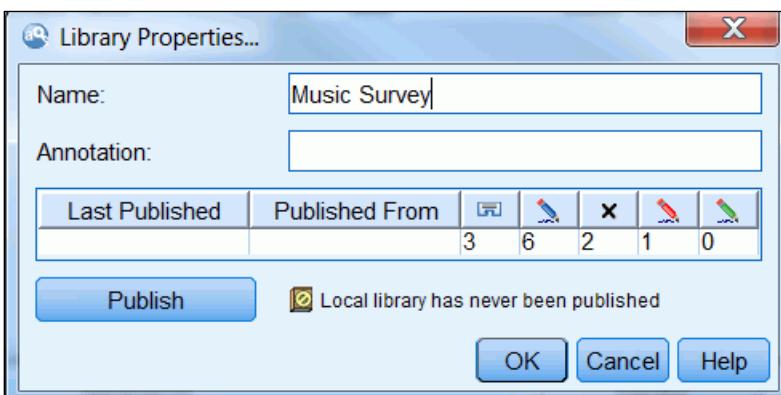
- Right-click **Local Library** and then select **Library Properties**.



- Rename the Local Library as **Music Survey**.

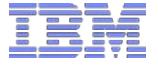


- Click **OK**.
- Right-click the **Music Survey** library, and then click **Library Properties**.



- Publish the library.
- Exit the Interactive Workbench session.
- Exit the Modeler session.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE



Using Text Mining Models

IBM SPSS Modeler Text Analytics (v16)

Business Analytics software



© 2014 IBM Corporation

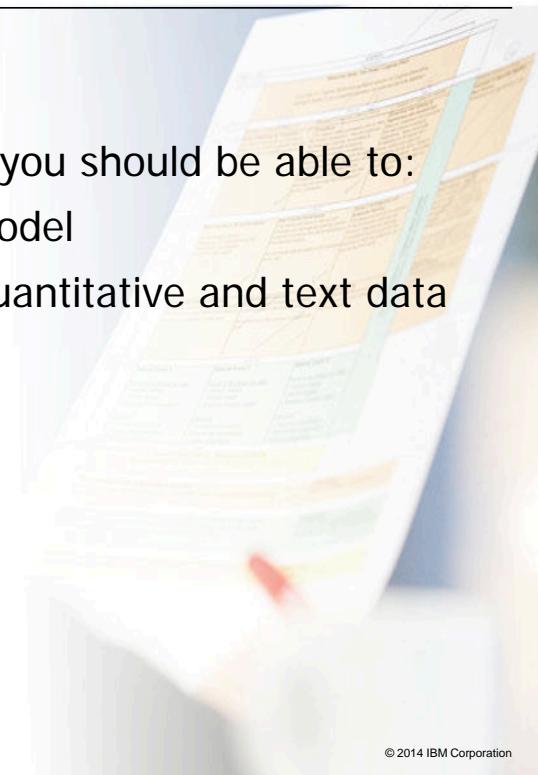
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

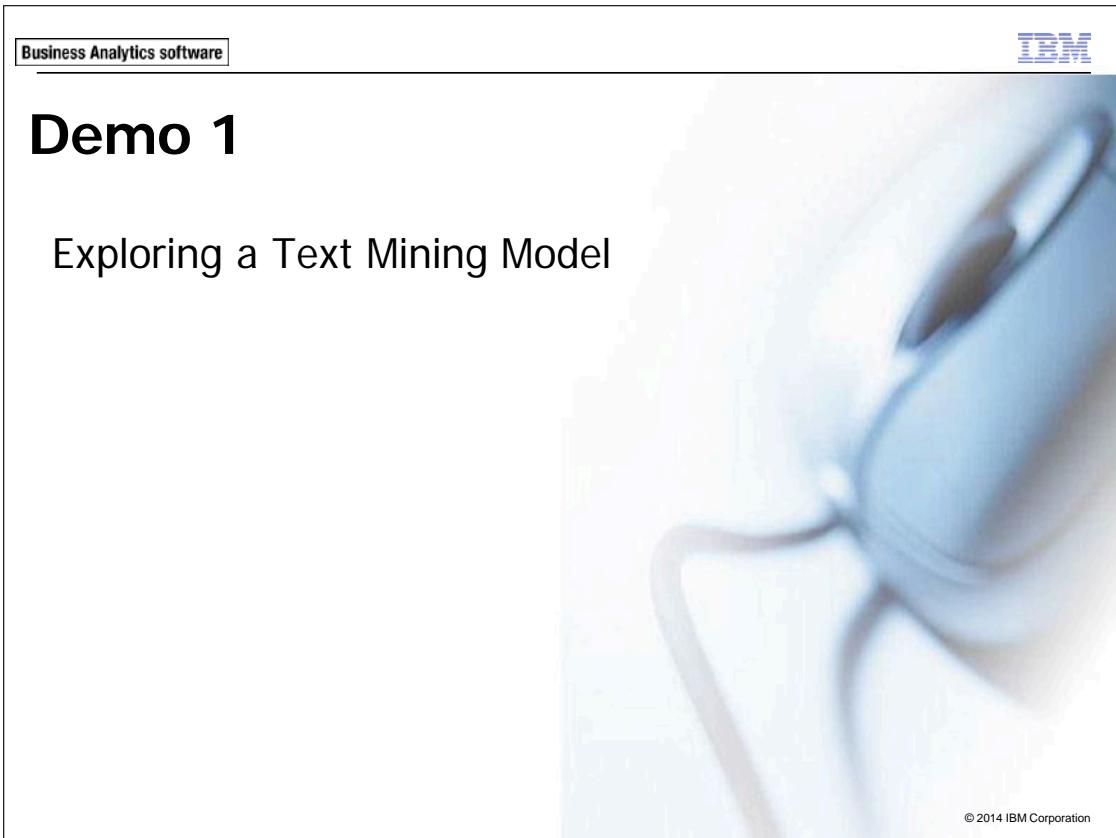
Objectives

- At the end of this module, you should be able to:
 - explore a text mining model
 - develop a model with quantitative and text data
 - score new data



© 2014 IBM Corporation

The final step in a text mining project is to use a model on new data, often in combination with other non-text information. In this module the call center data from May from Astroserve will be scored and combined with the customer database information to make predictions about customer churn. Deployment of text-mining models in the Modeler environment is also briefly discussed.



The following file(s) are used in this demo:

- Using_Text_Mining_Models_demo1_start.str - a Modeler stream contains a generated model with the final categories from the Astroserve call center data from March and April

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Demo 1: Exploring a Text Mining Model

Purpose:

Now that you have created your final set of categories from the Astroserve data, the company executives would like you to summarize your findings. They are interested in finding out what key themes (categories) you found in the data, and the number of records in each category.

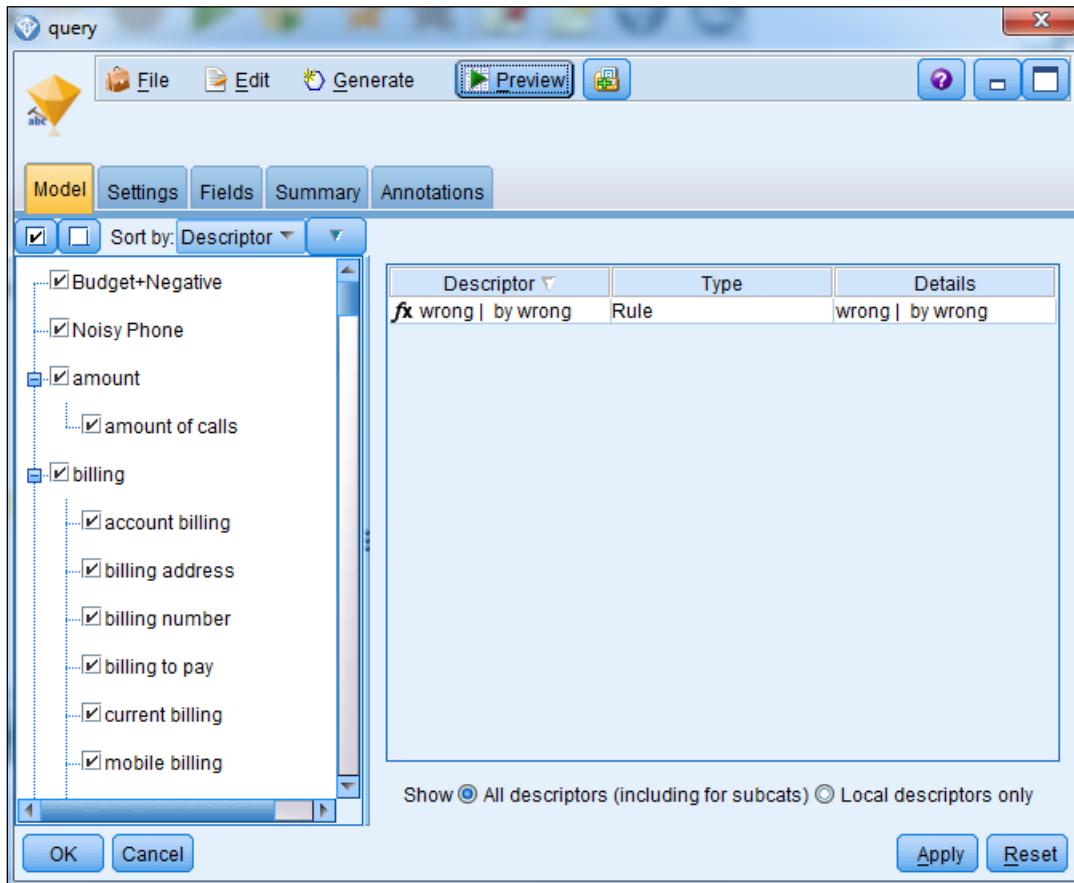
Task 1. Adjust the generated model settings.

1. From the **File** menu, click **Open Stream**.
2. Navigate to **C:\Train\0A105\14-Using_Text_Mining_Models**, and then double-click **Using_Text_Mining_Models_demo1_start.str**.
3. Delete the **25% Random sample** node.

While it is often advisable to use a sample of the data when creating a text mining model, after the model is created, you should use the full data just to make sure that the reports you create from the data are completely accurate.

4. Connect the data file to the generated model **query**.

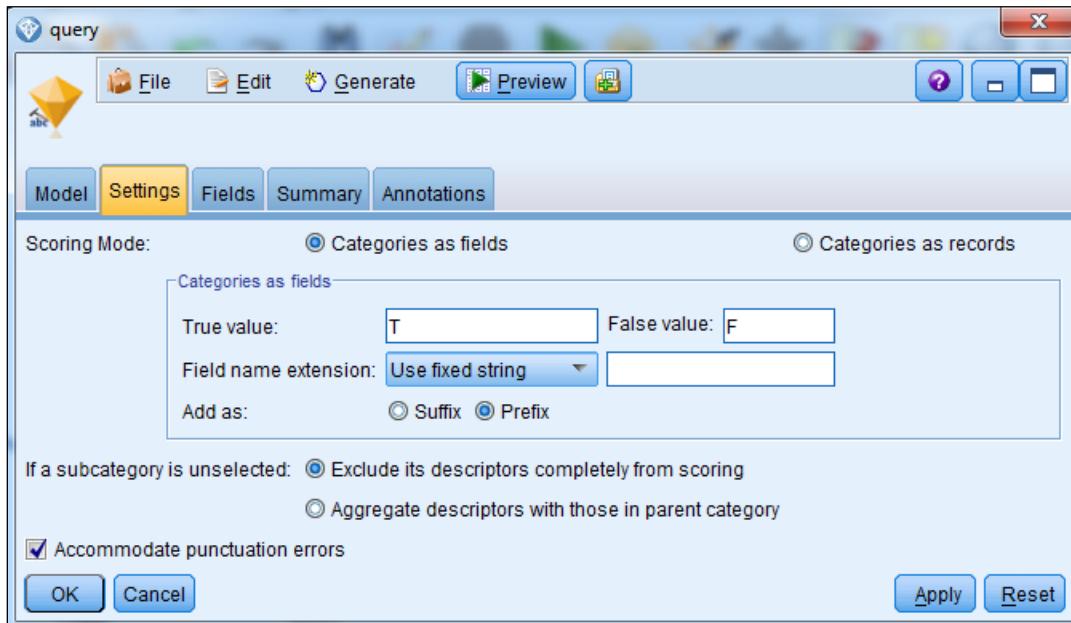
5. Edit the **query** generated model.



By default, all the categories will be used for scoring. Although the categories were created with the expectation that all of them would be helpful in building a predictive model, you may certainly decide that not all of them need to be used or retained. If so, the category can be deselected here with its checkbox.

6. Click the **Settings** tab.

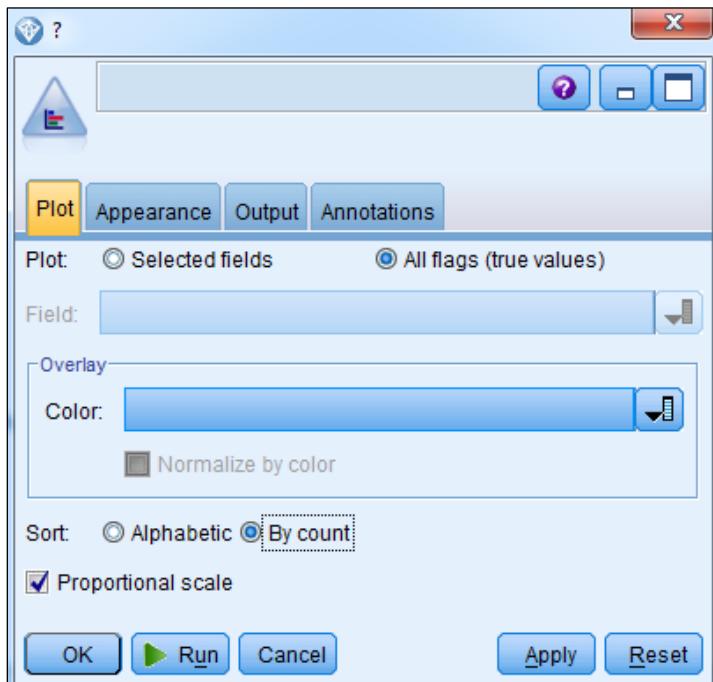
7. Delete the **Field name extension** text.



8. Click **OK**.

Task 2. Create a report from the data.

1. Add a **Distribution** node to the stream.
2. Connect the **generated model** to the **Distribution** node, and then edit the **Distribution** node.
3. Select **All flags (true values)**, and then select **Proportional scale**.
4. Beside **Sort**, click **By count**.



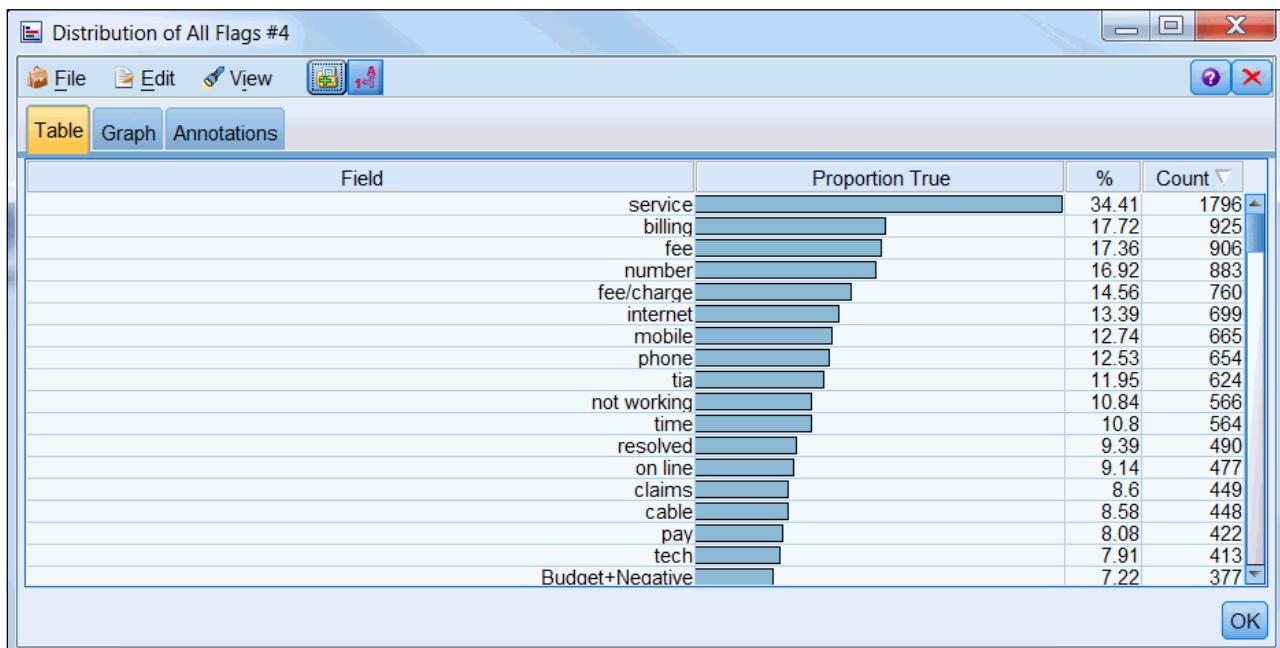
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

5. Click Run.

The results are as follows:



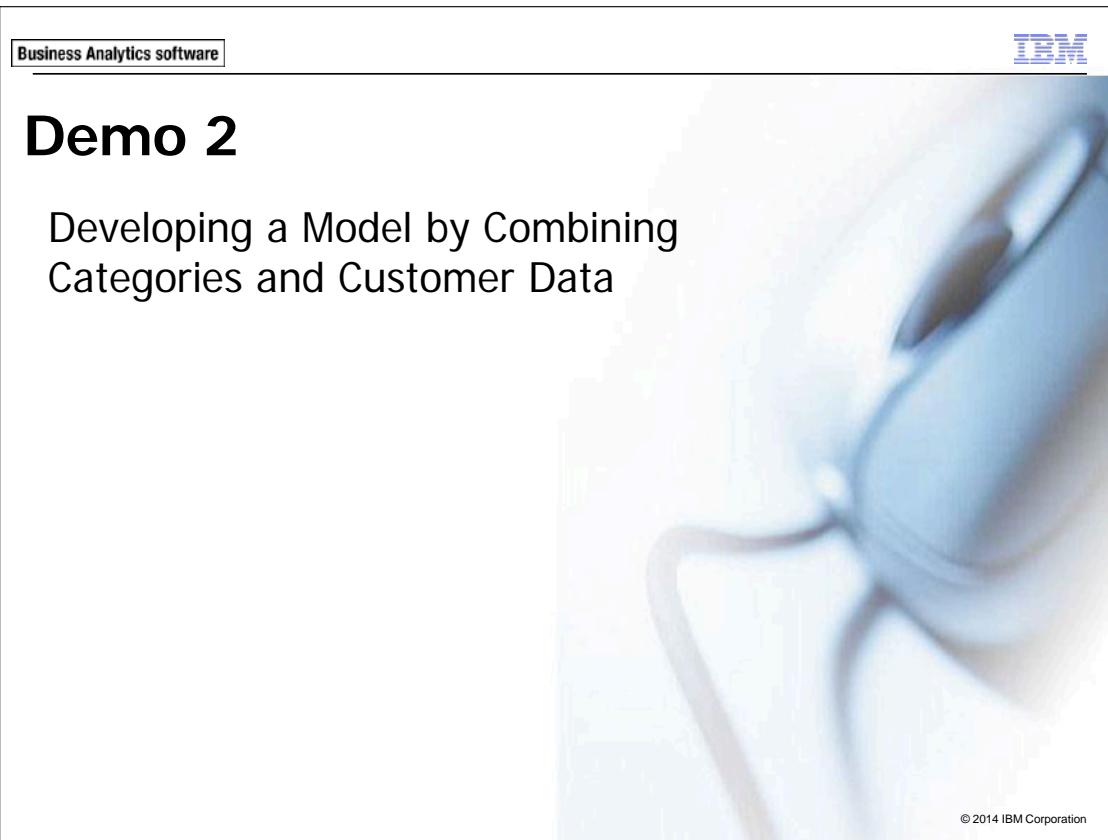
Observe that categories such as "service" and "billing" are high on the list. Also the Budget + Negative category that was constructed from Text Link Analysis occurred quite often.

6. Close the **Distribution node** window.
7. From the **File** menu, click **Close Stream**, without saving.
8. From the **File** menu, click **New Stream**.

Do not close Modeler; leave it open for the next demo.

Results:

You have successfully created a report on the key themes you discovered from the data that was collected by the Astroserve Call Center during the months of March and April.



The slide is titled "Demo 2: Developing a Model by Combining Categories and Customer Data". It features the IBM logo in the top right corner and the text "Business Analytics software" in the top left corner. The background is a blurred image of a person's face. A small copyright notice "© 2014 IBM Corporation" is visible at the bottom right.

The following file(s) are used in this demo:

- Using_Text_Mining_Models_demo2_start.str - a Modeler stream contains a generated model with the final categories from the Astroserve call center data from March and April
- Query.sav - a Statistics file

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

14-9

Demo 2: Developing a Model by Combining Categories and Customer Data

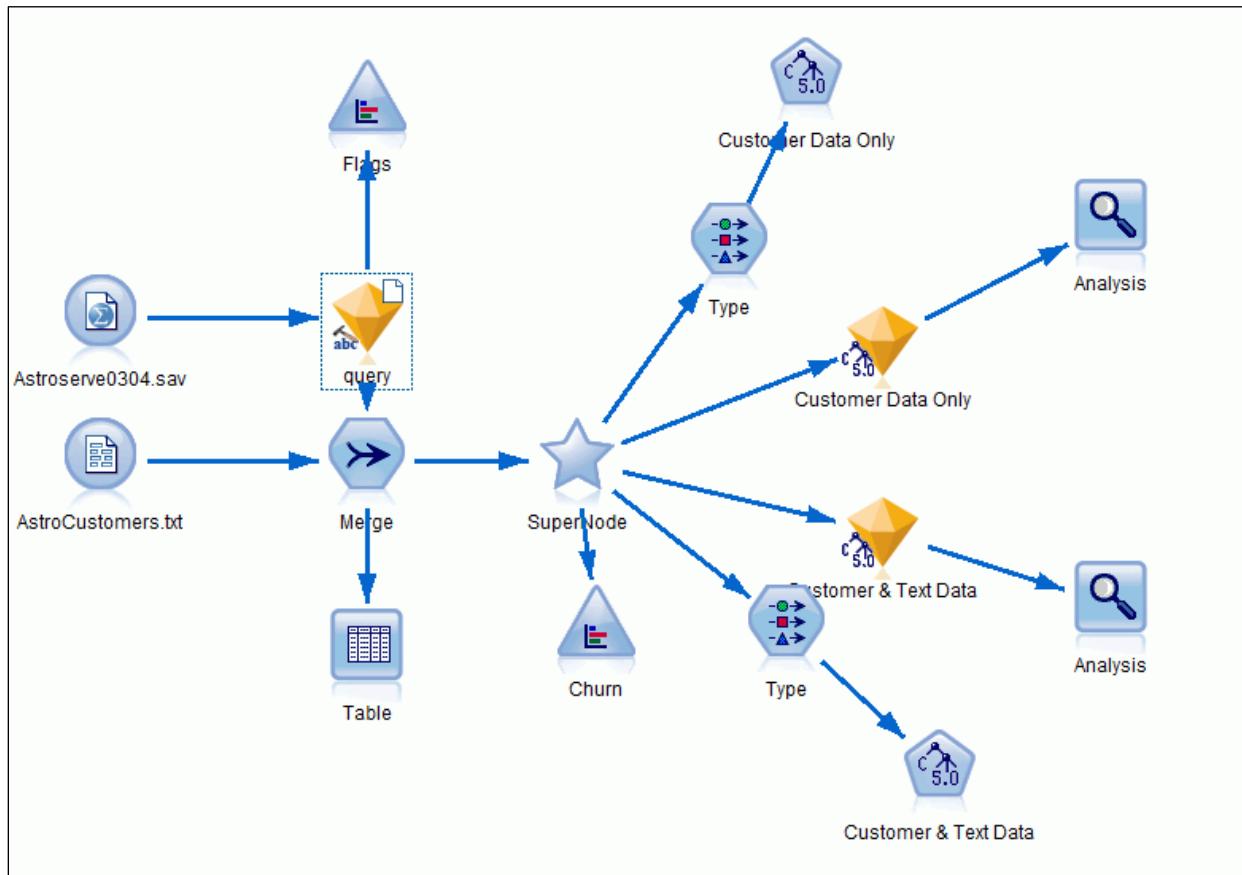
Purpose:

Now that you have identified the key themes you discovered from customer calls, the company is interested to know if the categories that you created from the text data will help to predict customer churn any better than the demographic data already available about the customers.

The Astroserve company wants to use the results of the text mining model to see how issues appearing in the call center records might help to predict future customer behavior, specifically the cancelling of an account (churn). To answer this question, you will develop a decision tree model to predict customer churn with both the text-based categories and other customer data. This requires merging the current data stream with the customer data in the file AstroCustomers.txt, as was done in a previous module. To illustrate how adding text data to a model can improve predictions, models will be created with and without the categories.

To save time because of the many nodes that must be added to the stream, you will use a prepared stream that contains the generated text-mining model.

The stream appears as follows:

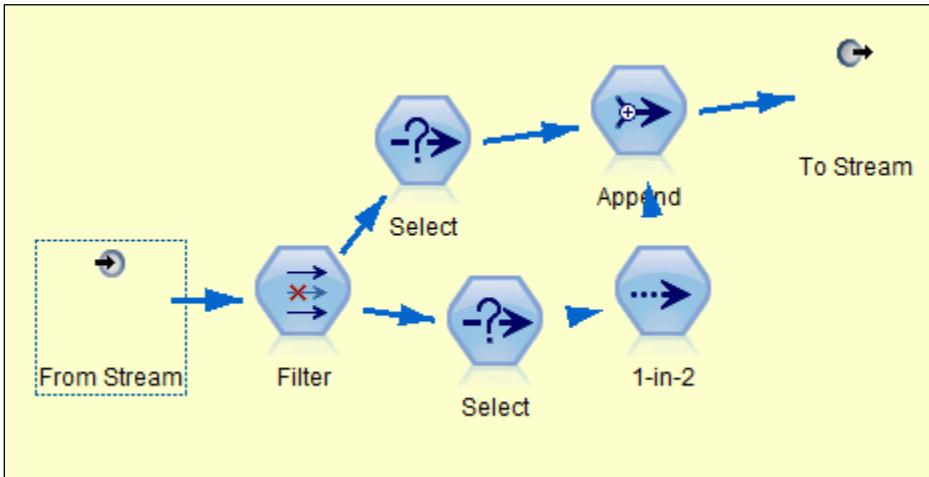


The stream first merges the data file with the customer data. Notice that a cache has been enabled on the generated text mining model query to speed up processing after the first execution.

The customer data fields in the AstroCustomers.txt file include the following, with actual data values in parentheses:

Field Name	Description
Query_ID	Query Identification number from call center data
Cust_ID	Customer Identification number
Priority (Y/N)	Whether customer is a Priority customer
Gender (F/M)	Gender of primary account holder
A_cust (Y/N)	Whether customer has a landline (Astrocomm) account
A_years	Number of years as a landline customer
A_lines	Number of land lines on account
N_cust (Y/N)	Whether the customer has an internet service (Astronet) account
N_years	Number of years as an internet services customer
N_type (B/D)	Type of Astronet account: Broadband or Dial-up
M_cust (Y/N)	Whether the customer has a mobile phone account
M_years	Number of years as a mobile phone customer
M_period (MM/1Y/2Y)	Mobile phone account type: Monthly, one year, two years
Churn (Y/N)	Whether the customer has cancelled all accounts by March or April

The Merge node joins the file together based on Query_ID. The Supernode adjusts the imbalance between the segment that churns (19.5%) and those who do not (80.5%), since an imbalance can cause the model to predict all instances of the largest segment (Churn="N" in this case). You want to ensure that balancing gives the same result each time the stream is run, so the data is balanced by sampling every second record of Churn="N" and all records for Churn="Y".

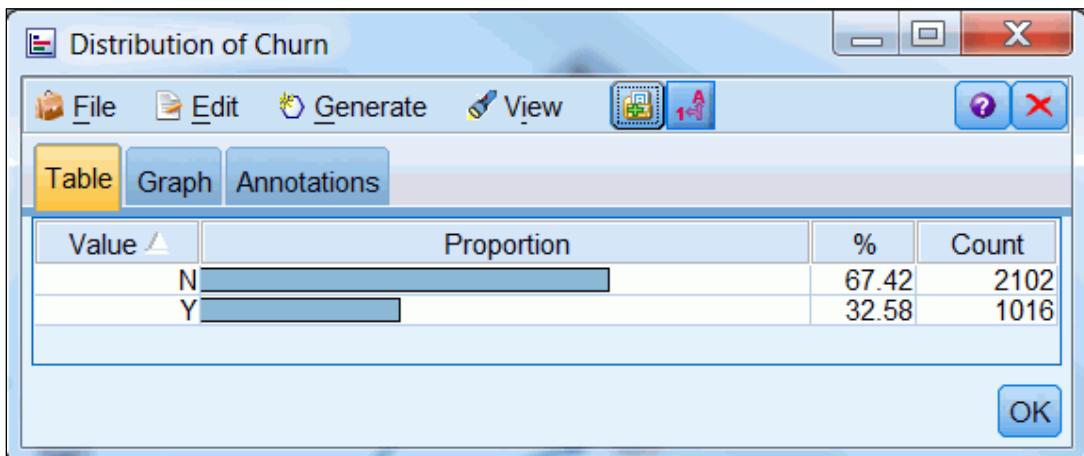


The stream branches into two, with the upper branch selecting all records where a customer churned. The lower branch selects every second record where a customer did not churn. This will yield about a two to one split between non-churners and churners, which is adequate for model development in this example (a 50/50 split would be better but would discard too many of the non-churners). The Append node is then used to add the two sets of records together in the last step.

Task 1. Open the stream.

1. From the **File** menu, click **Open Stream**.
2. Navigate to **C:\Train\0A105\14-Using_Text_Mining_Models**, and then double-click **Using_Text_Mining_Models_demo2_start.str**.
3. Right-click the **model nugget** named **query**, point to **Cache**, and then click **Load Cache**.
4. Navigate to the **C:\Train\0A105\14-Using_Text_Mining_Models** folder (if necessary), select **query.sav**, and then click **Open** to load the cache.

5. Run the **Distribution** node named **churn**.



Decision Trees allow you to develop models that predict or classify future behavior based on a set of decision rules. That is, when data are divided into classes of interest (churned vs. not churned in this example), predictors can be used to build rules that best classify cases with maximum accuracy. The C5.0 decision tree model is one of several available in Modeler and will be used as an example of this kind of predictive modeling.

To predict churn and run a model, the data need to be typed. This means adding a Type node after the Supernode. This ensures that Modeler knows the type of field and values for all the fields created from the text-mining model.

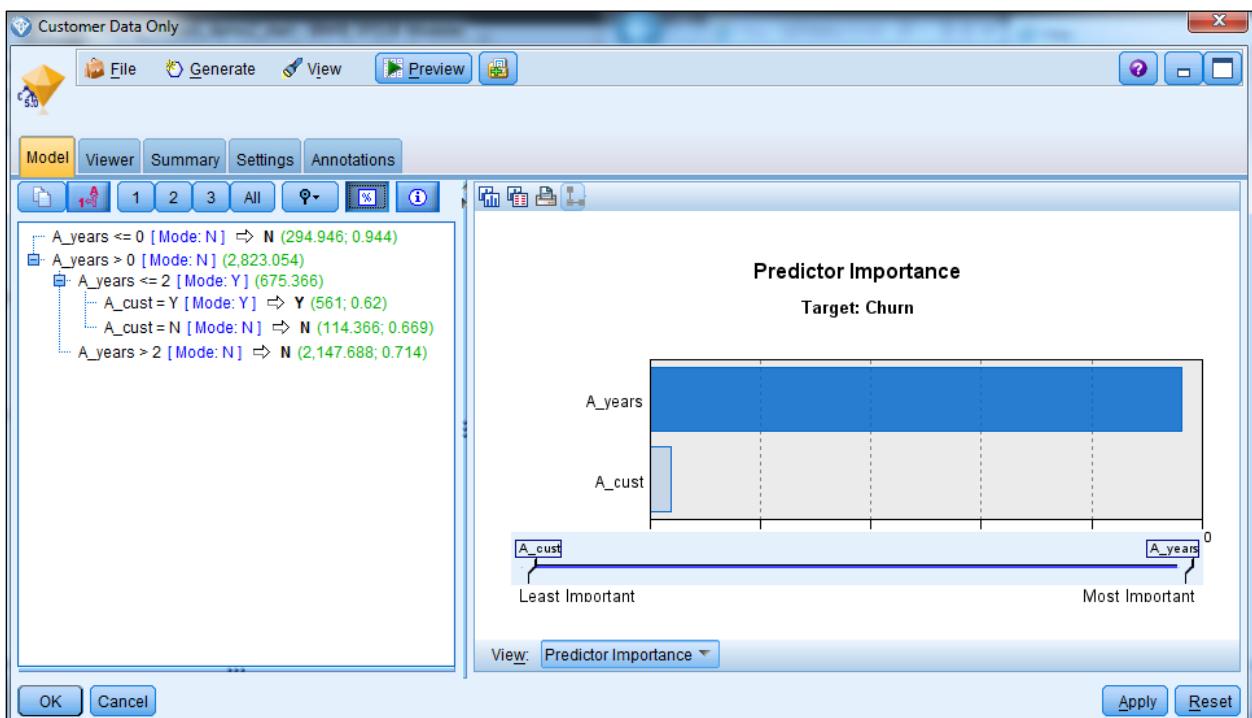
Then use two C5.0 models. The one on the left uses only the fields in the customer database. The one on the right uses those fields and all the categories from the text mining model. These models are set to favor generality so they should perform better on data in the future.

To save time, the models have already been created and then added to the stream. They have appropriate names to differentiate them. Review each in turn.

6. Click **OK** to close the **Distribution of Churn** window.

Task 2. Explore the model created with customer data only.

1. Edit the **Customer Data Only** generated model.
2. On the **Model** tab, maximize the left pane, and then click **All**.
3. Click **Show or hide instances and confidence figures** .



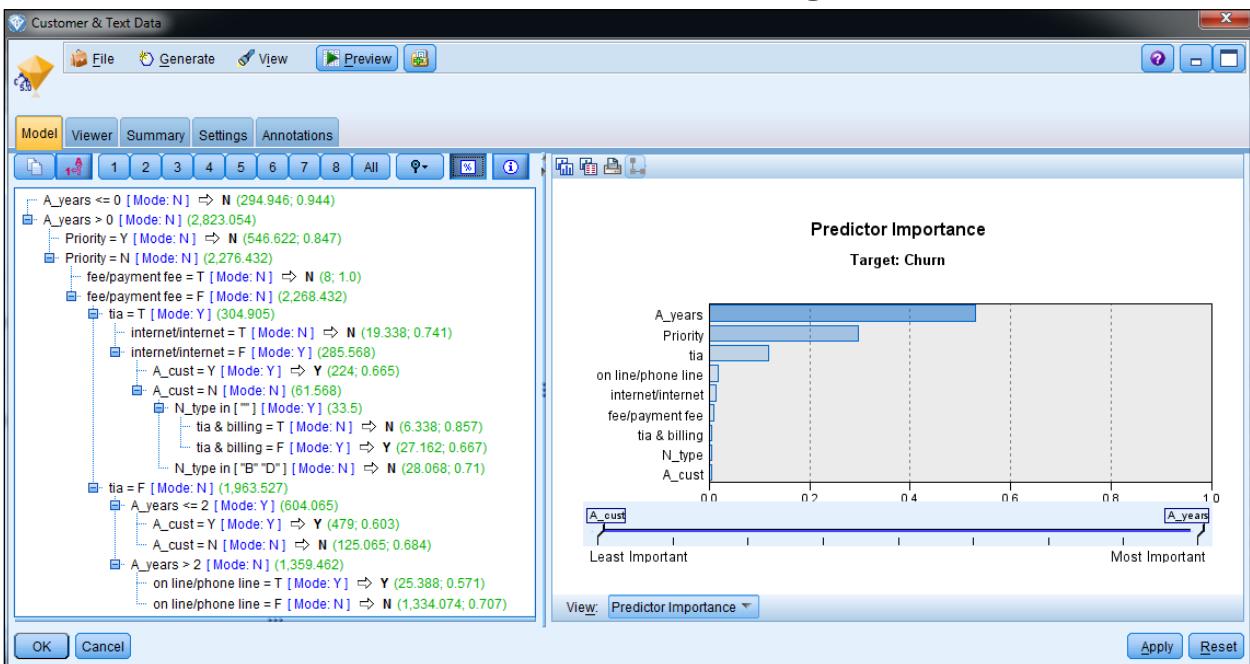
The model is very simple. Only two fields are used in the decision tree, A_years (the number of years as a landline customer), and A_cust (whether the customer has a landline account).

Customers who are very new—who have been with Astroserve less than one year—are in the first line of the tree. They are predicted not to churn (the predicted value of N), and the model is accurate on almost all these records (.944, or 94.4%). Those who are predicted to churn (the branch of the tree with the predicted value of Y) are customers with a landline account who had landline accounts 1 or 2 years.

Now review the model with the added information from the customer call center records.

Task 3. Explore the model created with customer data and text data.

1. Close the **C5.0** generated model.
2. Edit the generated model **Customer & Text Data**.
3. Ensure the **Model** tab is selected, maximize the left pane, and then click **All**.
4. Click **Show or hide instances and confidence figures** .



This model is much more complex. It has eight levels in the tree, and it uses more of the customer fields. The first split is still on the field `A_years`, on new customers versus all others. The other split is on whether the customer was in the "tia" or not, one of the fields from the text mining analysis.

The categories of "tia", "tia & billing", "internet/internet", "fee/payment fee", and on "line/phone line" that were created from the text data are used in the model. Just having a complaint does not necessarily lead to churning; instead, the combination of these two fields and customer fields yields predictions of churning.

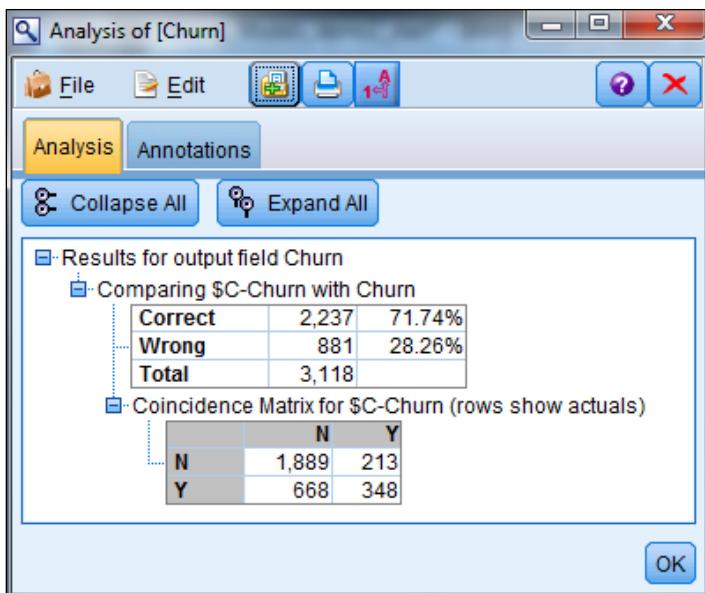
On the average, a more complex tree is likely to lead to a more accurate model. To assess the models, use an Analysis node, which will create a cross-classification matrix to compare the actual values of Churn to the predicted values from the models. The nodes have been annotated with appropriate names for the output.

5. Close the **C5.0** generated model.

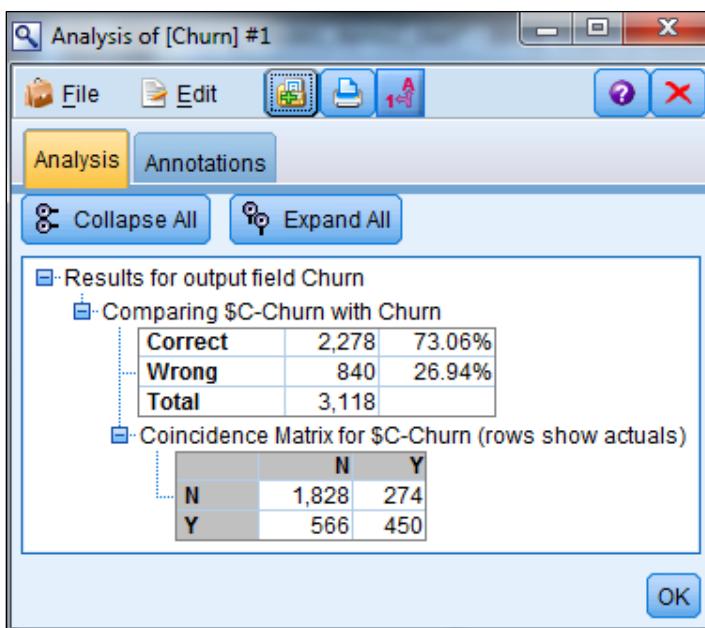
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Task 4. Compare the accuracy of the two models.

- Run the **Analysis** node attached to the **Customer Data Only** model.



- Run the **Analysis** node attached to the **Customer & Text Data** model.



The model without text data is accurate on 71.74% of the records. It is much more accurate on those who are predicted not to churn (the N row) than those will churn. But only about one third of those who will churn are predicted accurately ($348/(348+668)$).

The model with text data is about 2% more accurate (73.06%) overall. It is also promising that this model is much more accurate on those who did churn, with a 44% accuracy ($450/(450 + 566)$).

Thus, including the text data not only improved accuracy, it greatly increased accuracy on the category of most importance to Astroserve.

3. Close both **Analysis** windows.

Task 5. Examine actual records.

While the reports provide convincing evidence of using people's words to predict their behavior, in this case customer churn, it also helps to examine an actual customer records to see if you could have predicted churn just as well from their demographics, or whether it really does help to take their comments into account.

To illustrate, take a look at Query number 386751. According to the data, this is a 4-year landline customer with three phone lines. Based on the demographic data alone, this customer would have predicted that he remains loyal. However, the concepts derived from the text of his query predict otherwise. Here is the actual text:

"TIA level 1 complaint 03/991.cust claims there has been a delay in fixing a no dial tone fault he states he reported on service number. He claims he was advised there is a fault with the cable in the street and alleges a technician attended to the fault on 26/02/03. He claims he was advised a technician would attend again on 28/02/03, and he is extremely dissatisfied as he claims the problem has not been fixed as yet and has made several inquiries to the faults area."

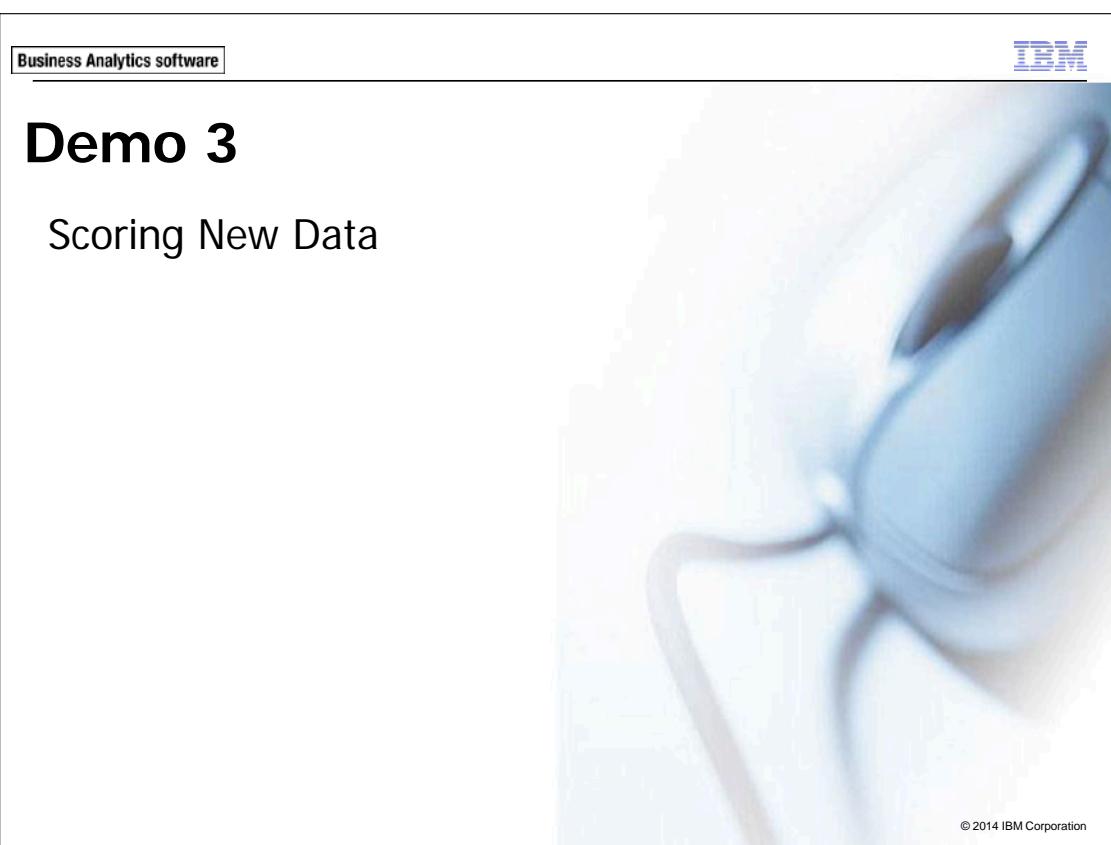
This illustrates the advantage of adding text data to the model.

1. From the **File** menu, click **Close Stream**.
2. From the **File** menu, click **New Stream**.

Do not close Modeler; leave it open for the next demo.

Results:

You have successfully proved that taking Astroserve customer comments into account improves your ability to predict whether or not they will remain loyal.



The slide is titled "Demo 3: Scoring New Data". It features the IBM logo in the top right corner and the text "Business Analytics software" in the top left corner. The background is a blurred image of a person wearing a hard hat and safety glasses. A small copyright notice "© 2014 IBM Corporation" is visible at the bottom right of the slide area.

The following file(s) are used in this demo:

- Using_Text_Mining_Models_demo3_start.str - a Modeler stream contains a generated model with the final categories from the Astroserve call center data from March and April
- query.sav - a Statistics file

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

14-19

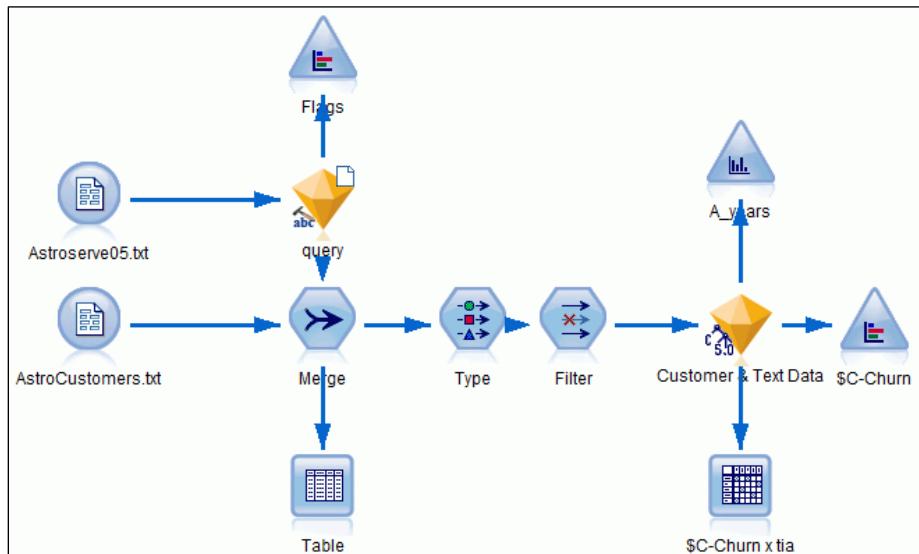
Demo 3: Scoring New Data

Purpose:

In May, Astroserve collected more call center data. The company would like to use the model built from March and April data to predict which customers are likely to churn.

There is a stream created that has most of the necessary nodes to score the May data. The stream, Using_Text_Mining_Models_demo3.start.str, merges the May customer call center data with the customer database file, as was done previously. It then types the data and runs the stream into the generated model that uses both the customer fields and the categories from the text-mining model.

The stream appears as follows:



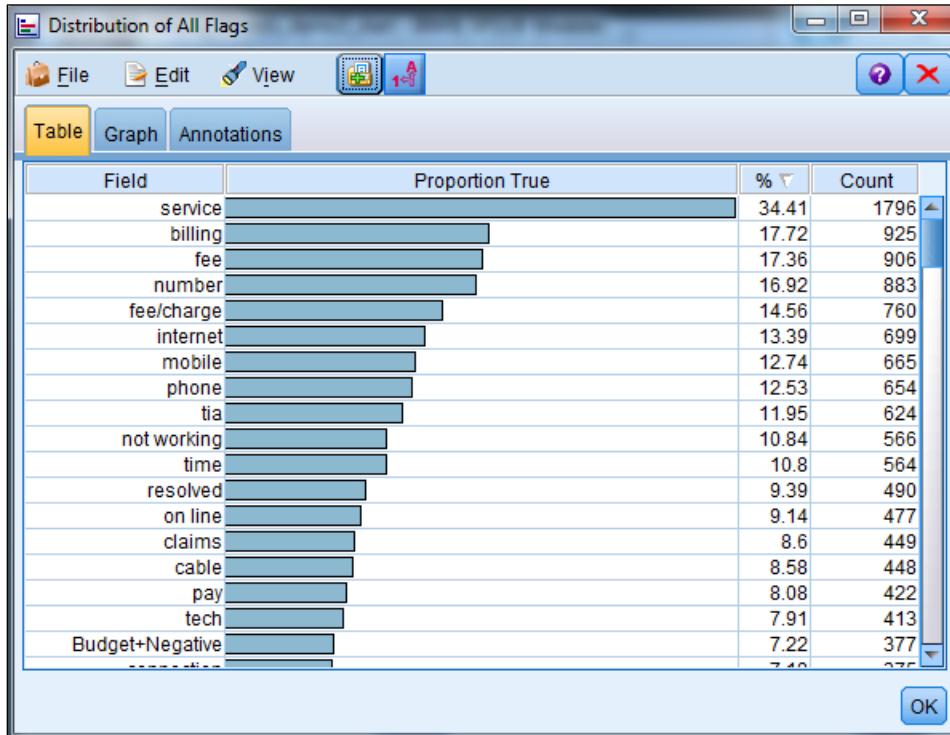
Task 1. Open and prepare the stream for scoring.

- From the **File** menu, click **Open Stream**.
- Navigate to **C:\Train\0A105\14-Using_Text_Mining_Models**, and then open the stream **Using_Text_Mining_Models_demo3_start.str**.
The data for May have the same database fields, except the field Churn is all blank since that information had not yet been recorded.
- Right-click the **model nugget** named **query**, point to **Cache**, and then click **Load Cache**.
- Navigate to **C:\Train\0A105\14-Using_Text_Mining_Models** (if necessary), select **query.sav** and then click **Open** to load the cache.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Task 2. Score the data.

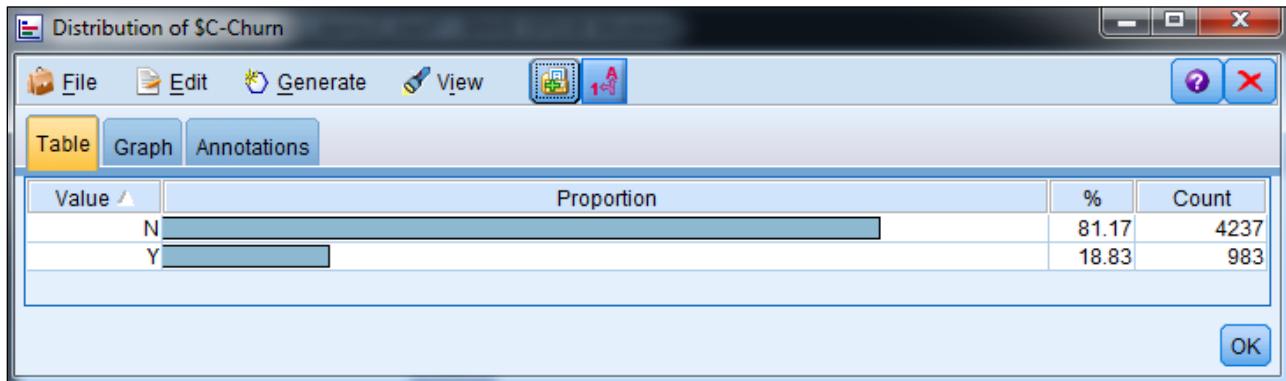
1. Run the **Distribution** node named **Flags**.
2. Click the **%** column heading until the percentages are in descending order.
The results appear as follows:



The top five categories are the same as what was seen previously, so the categories that appear most frequently in May are the same as in March and April.

To see how many of the call center communications generate predictions that those customers will churn, you can produce a bar (Distribution) chart of the field \$C-Churn.

3. Run the **Distribution** node named **\$C-Churn**.

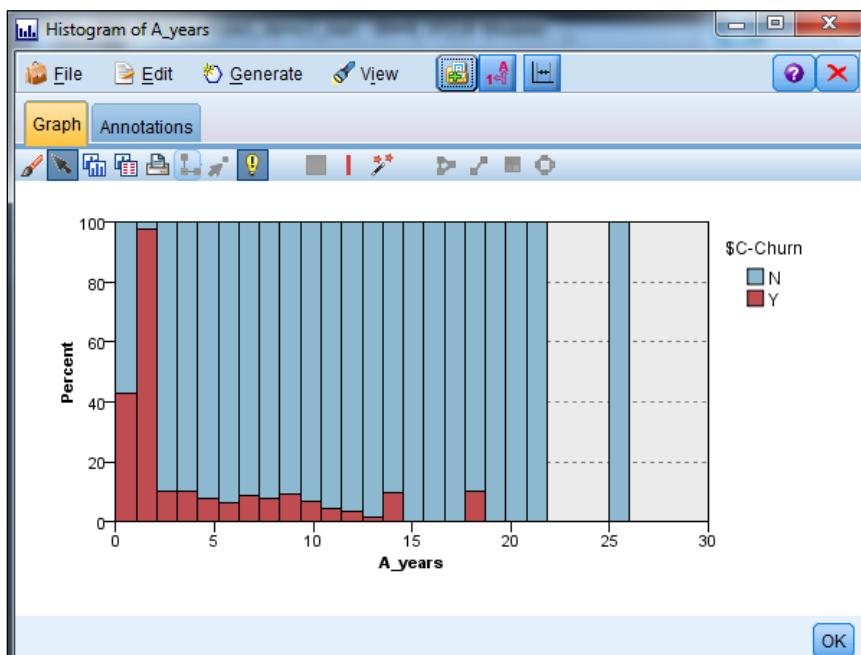


The model predicts that 18.83% of the customers will churn. Recall that in the data for March and April where the churn status is known, 23.22% of the customers churned; thus, these two figures are in good agreement. Any acceptable model should predict about the same amount of churn in the future, all things being equal.

Continuing along with this example, it would be interesting to see the relationship between some of the key variables in the model and predicted churn. The first split in the tree occurred on A_years. That variable is of type Continuous. You can produce a histogram of A_years, with an overlay of \$C-Churn.

4. Close the **Distribution** window.
5. Run the **Histogram** node that is found downstream from the C5.0 modeling nugget named **Customer & Text Data**.

The results appear as follows:

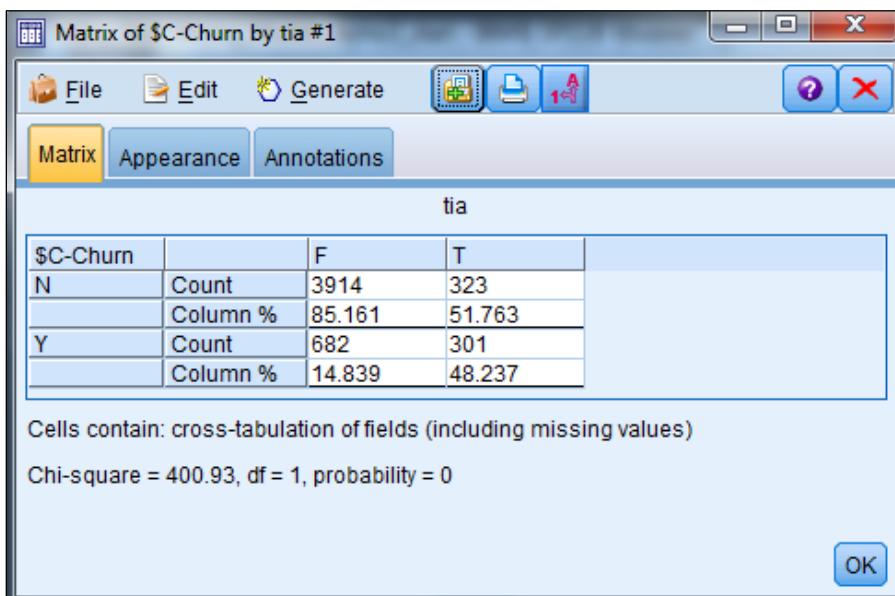


Notice that, as with the March and April data, new customers are not predicted to churn (those with a value of 0 on A_years). But the chance of churning goes up enormously in the first and second years to over 90%. In other words, dissatisfied customers cancel their accounts with Astroserve relatively quickly. But if a customer makes it past the second year, they have a good chance of remaining.

Next notice how one of the key text fields relates to predicted churn, using a tabular report.

6. Close the **Histogram** window.
7. Run the **Matrix** node that is found downstream from the C5.0 modeling node named **Customer & Text Data**.

The results appear as follows:



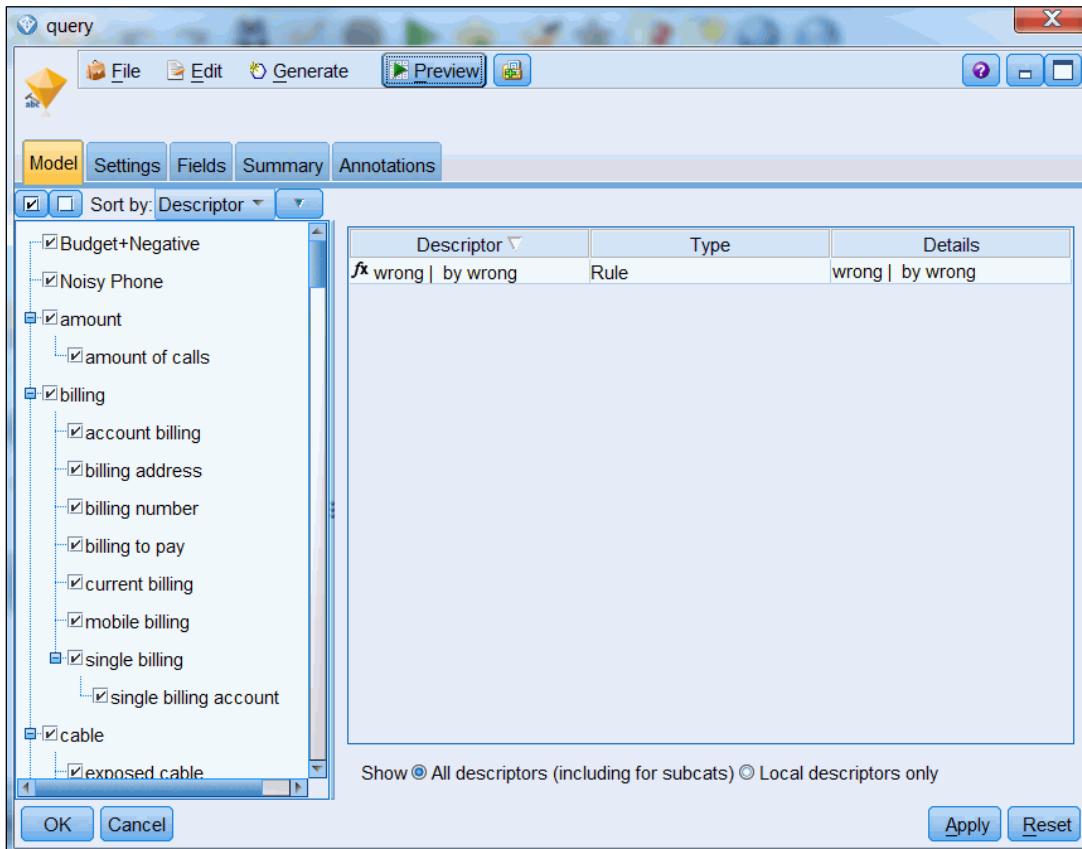
If a customer mentioned tia (T), it appears that they were more than twice as likely to churn than people who did not mention tia (F). This demonstrates the importance of this category to the model's predictions. Astroserve can now produce a list of those customers who are predicted to churn and take preventive action.

Task 3. Perform efficient model deployment.

If you intend to use the Generated Text Mining model in a production environment to score new data on an automatic and periodic basis, you will find it more efficient (and reduce the model size) to edit the generated text mining model so that only the concepts used in the subsequent prediction model are scored (checked).

1. Close the **Matrix** node.
2. Edit the generated **query** mining model.

3. Click the **Model** tab.



All the checked categories will be scored, but only four of the categories were used in the model (this is typical of decision tree models on relatively small samples). There is no reason to create the other categories on new data, as that will significantly increase processing time, so you can uncheck them.

The Text Mining generated model contains all the resources necessary to score new text data. This means that you can use this node, or the stream, on a PC with Text Mining for Modeler, but without the exact resource template or libraries as on the PC on which the model was developed. This flexibility allows the use of a text mining model without having to share libraries and resource templates with colleagues for the scoring of new data.

4. From the **File** menu, click **Exit** to end the Modeler session.

Results:

You have successfully used the text mining model you created from March and April call center data to score the data from Astroserve Customers. According to your results, you predict that 18.5% of these customers will churn.

Apply Your Knowledge

Purpose:**Test your knowledge of material covered in this module.**

Question 1: True or False: On the average, a more complex tree is likely to lead to a more accurate model.

- A. True
- B. False

Question 2: True or False: Any acceptable model should have similar predictions in the future, all things being equal.

- A. True
- B. False

Question 3: True or False: Demographic variables are usually sufficient by themselves for predicting future behavior.

- A. True
- B. False

Question 4: True or False: Scoring is the process of using a model to make predictions about behavior that has yet to happen.

- A. True
- B. False

Question 5: True or False: The data to be scored must have the same fields as the database that was used to create the model.

- A. True
- B. False

Apply Your Knowledge - Solutions

Answer 1: A. True

Answer 2: A. True

Answer 3: B. False

Answer 4: A. True

Answer 5: A. True

Summary

- At the end of this module, you should be able to:
 - explore a text mining model
 - develop a model with quantitative and text data
 - score new data

© 2014 IBM Corporation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

14-27

Business Analytics software



Workshop 1

Using Text Mining Models



© 2014 IBM Corporation

No supporting materials are needed for this workshop. The answers appear on the page following the questions.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Workshop 1: Using Text Mining Models

Applying text mining models to new data is a key component of most text mining projects. This gives you the ability to predict future behavior before it happens. There is no computer exercise for you to perform in this workshop. You will answer the following questions which pertain to the topic of using text mining models.

- Question 1: Although categories were created with expectations that all of them would be helpful in building a predictive model, you may certainly decide that not all of them need to be used or retained.
- A. True
 - B. False
- Question 2: To score new data, you must run the new data through the generated model.
- A. True
 - B. False
- Question 3: If you intend to use the Generated Text Mining model in a production environment to score new data on an automatic and periodic basic, you will find it more efficient to edit the generated text mining model so that only the concepts used in the subsequent prediction model are scored.
- A. True
 - B. False
- Question 4: The Text Mining generated model contains all the resources necessary to score new text data. Therefore, you can use this node without the exact resource template or libraries on which the model was developed.
- A. True
 - B. False

Workshop 1: Tasks and Results

Answer 1: A. True

Answer 2: A. True

Answer 3: A. True

Answer 4: A. True

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

14-30

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.



The Process of Text Mining

IBM SPSS Modeler Text Analytics (v16)

Business Analytics software



© 2014 IBM Corporation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

Business Analytics software

IBM

Objectives

- At the end of this module, you should be able to:
 - explain the steps that are involved in performing a text mining project

© 2014 IBM Corporation

Many data miners who are new to text mining would benefit from a somewhat detailed guide to the text mining process. In this appendix, a basic guide is provided within the context of the Modeler text mining environment.

1. Do data exploration just as with quantitative data. Here, this also means actually reading some of the text data—say 50 to 100 records/documents to get a sense of the data, subjects, organization of text, etc.
2. Either before this or now, discuss the text mining project with the subject matter expert (SME). Get a clear idea of the key goals of the analysis (prediction, understanding), and any issues about the text that may not be apparent from simply reviewing the data.
3. If the data file has more than about 15,000 records/documents, take a sample from the data of no more than this size, even smaller for some of the initial analysis.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

4. Use the Text Mining Modeling node to create a Concepts model. This will allow you to:
 - See which concepts are extracted with the default resources.
 - See which concepts are most frequent and least frequent.
 - See which synonyms are being used.
 - See which text is not being extracted that seems like it could be important.
 - Try at least two resource templates to compare the results. You will have to choose a resource template to use as a basis for the project, so the results here can help you make that choice.
5. Review the extracted concepts with the SME. See which concepts are most promising and which synonyms seem reasonable. Try to discover what text is not extracted but should be.
6. Use the Text Mining Modeling node in Interactive Workbench mode. The same basic set of concepts will be extracted as in step 4.
7. You now must review the results of the extraction very carefully. You must decide which:
 - concepts are important
 - concepts require synonyms
 - misspellings are not corrected by the software
 - concepts are incorrectly grouped together because of the fuzzy grouping algorithm
 - concepts should be grouped under a type
 - text is not extracted but should be by grouping it under a type
 - terms should be excluded because they appear too often but are uninformative

In all of these decisions, you should think about in which library it is best to make a change to the linguistic resources. If you change a library that is shared with other resource templates, such as the Core or Budget libraries, and you publish that library and/or save the template under the original name, then the library will be used by other projects with the changes you made.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

If instead you make changes to the local library (which you will eventually rename), or to a library that you have added to the project (created previously), you will only affect that specific library, or the template that you save under a new name (this is usually the better choice).

The exception can be when you need to change a synonym or type that is already in an existing library that is supplied with the project. What you might do then is to make that change, but be very careful not to publish the library. You can, for example, turn off a synonym definition and then add a synonym definition to the Local library. You will need to note this carefully in a project document so it will be clear how to recreate the linguistic resources.

Another key decision is whether to create a synonym or a type to group concepts together. Generally speaking, you use types because you need to extract text that is not extracted or because you need specific matching choices. For example, you only want to use a concept if the extracted text begins with the concept, or includes the concept anywhere in the extracted text. If instead you simply want to group things and only see a single concept, rather than each listed separately, you can use a synonym definition.

If you use a synonym, then one of the categorization techniques may automatically place the concept in a category. If you use a type, then you are not guaranteed that the type will be a concept, although you can make the top N types into categories. In that situation, you may have to create the category manually.

8. When you have made many changes to the linguistic resources, you re-extract the text and determine whether the changes had the desired effect. Since you are likely to make many changes, you want to make at least a dozen changes before re-extracting to be efficient, but not so many that it will become unclear as to what edits caused what change in the concepts.
9. You repeat this cycle several times—extract, edit linguistic resources, extract, edit—until you are reasonably satisfied with the results. Don't worry about all the concepts that occur only once or twice; unless it is easy to categorize them, it probably isn't worth the effort when building a predictive model. You might also try some of the categorization techniques in the middle of this to see their effect and help decide how to edit the resources.

10. When you are finally done with editing, you then create categories. It is recommended first trying each of the 3 linguistic methods separately to see their effect and understand what each one does with the text. After creating the first categories model, and scoring it to get category frequencies, you look at the following characteristics:
- What are the most frequent and least frequent categories?
 - What concepts are included in each category? Can some be deleted or moved to other categories?
 - Which categories should be retained and which dropped?
 - Look at the unextracted concepts and see which ones can be added to a category. You can add concepts to categories manually, but if you can see a way to edit the resources so it occurs automatically, that is usually preferable (because this means the categorization can be easily redone).

You continue reviewing the categories, going back to make changes to the linguistic resources, re-extracting, creating categories, reviewing, and so forth. You can also change the settings for categorization, such as the number of concepts to use, to see different results. This continues until you are satisfied with the categories.

11. Next, consider creating categories based on links between concepts. You can do this manually or with text link analysis:

- Manually: You may want to create a rule that says that you want a category whenever peanut butter and jelly occur together anywhere in the same record or document (these concepts would already have to be extracted to create the rule). You select both categories and then create the rule (from the menus, click Categories / Add & Rule to Categories).
- Text Link Analysis: This procedure will automatically find links between text based on supplied text link macros for several of the resource templates, such as the Opinions (English) template used in this course. These links take into account how close the concepts are in a record so that it is more likely that the link refers to something meaningful. After running the analysis, you then review the results, finds links that are interesting, and adds them as categories. These will supplement the categories created in the previous step. You should not have to edit the regular linguistic resources for the Text Link Analysis.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

12. When done, generate a text-mining model from the menus. Place it in the stream and examine the model. This means first running a Distribution node for all the flag fields, but also using other nodes, such as the Matrix node, to see the relationship between the categories and other fields in the original data. You can also use a Web Graph node to look at relationships between categories, or even other fields.
13. When you have chosen the final set of concepts, filter out the ones that are not needed, and then output the data in an appropriate format, or merge the categories with other data sources to create one data stream that you can use to create a predictive model. Because you are likely to have many flag fields, neural net models are not the first choice. Decision trees, stepwise regression, or stepwise logistic regression (depending on the type of target field) are often models used with text data.
14. As with any modeling exercise, you should have both training and validation data sets to develop and then test the models. Do not redo the text mining on the validation data; just apply the text-mining model, and then the predictive model, to see how accurate the model is on the validation data.
15. Categories from a model can also be used to cluster the records or documents, or to create association models (such as the Apriori node).

Business Analytics software

IBM

Summary

- At the end of this module, you should be able to:
 - explain the steps that are involved in performing a text mining project

© 2014 IBM Corporation

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

A-8

© 2010, 2014, IBM Corporation

This guide contains proprietary information which is protected by copyright. No part of this document may be photocopied, reproduced, or translated into another language without a legal license agreement from IBM Corporation.

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

*IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.*

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE