

Text Analytics Using SAS® Text Miner

Course Notes

Text Analytics Using SAS® Text Miner Course Notes was developed by Terry Woodfield. Additional contributions were made by Rich Perline, Russell Albright, and James Cox. Editing and production support was provided by the Curriculum Development and Support Department.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Text Analytics Using SAS® Text Miner Course Notes

Copyright © 2011 SAS Institute Inc. Cary, NC, USA. All rights reserved. Printed in the United States of America. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

Book code E2046, course code LWDMTX51/DMTX51, prepared date 29Sep2011. LWDMTX51_001

ISBN 978-1-61290-120-6

Table of Contents

Course Description	vi
Prerequisites	vii
Chapter 1 Introduction to SAS® Enterprise Miner™ and SAS® Text Miner	1-1
1.1 Data Mining and Text Mining.....	1-3
Demonstration: Text Mining the Federalist Papers	1-39
Exercises.....	1-62
1.2 Working with Data Sources	1-63
1.3 Using SAS Enterprise Miner and SAS Text Miner.....	1-77
Demonstration: Text Mining SAS Course Descriptions	1-95
Exercises.....	1-107
1.4 Chapter Summary	1-108
1.5 Solutions	1-111
Solutions to Exercises	1-111
Solutions to Student Activities (Polls/Quizzes).....	1-111
Chapter 2 Overview of Text Analytics	2-1
2.1 Data Preparation for Text Analytics.....	2-3
Demonstration: Using the Text Import Node	2-10
2.2 Forensic Linguistics	2-13
Demonstration: Stylometry for Forensic Linguistics	2-18
2.3 Information Retrieval.....	2-21
Demonstration: Retrieving Medical Information.....	2-25
Exercises.....	2-33
2.4 Text Categorization.....	2-34

Demonstration: Categorizing Reports in the ASRS	2-39
Exercises.....	2-48
2.5 Chapter Summary	2-51
2.6 Solutions	2-52
Solutions to Exercises	2-52
Solutions to Student Activities (Polls/Quizzes).....	2-60
Chapter 3 Algorithmic and Methodological Considerations in Text Mining.....	3-1
3.1 Methods for Parsing and Quantifying Text.....	3-3
3.2 Quantifying Concepts Using Latent Semantic Analysis	3-24
3.3 Chapter Summary	3-34
3.4 Solutions	3-35
Solutions to Student Activities (Polls/Quizzes).....	3-35
Chapter 4 Applications of Text Mining to Pattern Discovery	4-1
4.1 Text Mining in Warranty Analysis	4-3
Demonstration: Text Analytics for Warranty Analysis.....	4-7
4.2 Processing and Categorizing Documents.....	4-25
Demonstration: Text Categorization for Identifying Potential Fraud Cases	4-33
4.3 Association and Sequence Discovery in Text Analytics	4-53
Demonstration: Association Discovery of Terms	4-56
Exercises.....	4-58
4.4 Chapter Summary	4-59
4.5 Solutions	4-60
Solutions to Exercises	4-60
Solutions to Student Activities (Polls/Quizzes).....	4-67

Chapter 5 Applications of Text Mining to Predictive Modeling	5-1
5.1 Predictive Modeling with SAS Enterprise Miner	5-3
5.2 Using Adjustor Notes to Predict Recovery Potential in Insurance Claims	5-17
Demonstration: Predicting Workers' Compensation Recovery Potential.....	5-20
5.3 Text Categorization via Predictive Modeling	5-25
Demonstration: Text Categorization of the ASRS Data.....	5-28
Exercises.....	5-41
5.4 Chapter Summary	5-42
5.5 Solutions	5-43
Solutions to Exercises	5-43
Solutions to Student Activities (Polls/Quizzes).....	5-44
Appendix A Index	A-1

Course Description

This course covers the functionality of SAS Text Miner software, which is a separately licensed component available for SAS Enterprise Miner. SAS Text Miner enables you to uncover underlying themes or concepts contained in large document collections; automatically group documents into topical clusters; classify documents into predefined categories; and integrate text data with structured data to enrich predictive modeling endeavors.

To learn more...



For information on other courses in the curriculum, contact the SAS Education Division at 1-800-333-7660, or send e-mail to training@sas.com. You can also find this information on the Web at support.sas.com/training/ as well as in the Training Course Catalog.



For a list of other SAS books that relate to the topics covered in this Course Notes, USA customers can contact our SAS Publishing Department at 1-800-727-3228 or send e-mail to sasbook@sas.com. Customers outside the USA, please contact your local SAS office.

Also, see the Publications Catalog on the Web at support.sas.com/pubs for a complete list of books and a convenient order form.

Prerequisites

Before attending the course, you should have experience using SAS Enterprise Miner to do pattern discovery and predictive modeling, or you should have completed the Applied Analytics Using SAS® Enterprise Miner™ course.

Chapter 1 Introduction to SAS® Enterprise Miner™ and SAS® Text Miner

1.1 Data Mining and Text Mining	1-3
Demonstration: Text Mining the Federalist Papers	1-39
Exercises	1-62
1.2 Working with Data Sources	1-63
1.3 Using SAS Enterprise Miner and SAS Text Miner	1-77
Demonstration: Text Mining SAS Course Descriptions.....	1-95
Exercises	1-107
1.4 Chapter Summary.....	1-108
1.5 Solutions	1-111
Solutions to Exercises	1-111
Solutions to Student Activities (Polls/Quizzes)	1-111

1.1 Data Mining and Text Mining

Preliminary Remarks

Text analytics is a very general field of study. Sources disagree on the formal definition of text analytics and of text data mining. The purpose of this course is to teach you how to solve analytic problems that include access to relevant textual data.

Access to real business data is always problematic. Because text fields often contain confidential information, access to business data that includes text is even more difficult. Even when real data is available, detailed and explicit descriptions of troublesome events can be disturbing. Most data sets used in this course are publicly available. However, some data sets contain explicit and potentially disturbing entries as well as references to specific companies and brand names. *All data used in this course is modified in some way.* Modifications include the following:

- deletion of sensitive entries
- deletion of potentially embarrassing or misleading entries
- editing or deletion of entries with named individuals or business organizations
- editing of text fields having obscure or confusing references
- resolution of ambiguities that might be incorrect
- modification or deletion of entries to promote educational goals

Because of these modifications, the data should not be used for any purpose other than education. All publicly available data sets are introduced with a reference to the source of the actual data. You should acquire data from the source if you want to use the data for business or scientific reasons.

Objectives

- Describe text analytics and define text data mining (text mining).
- Describe how SAS Enterprise Miner is used for data mining and text mining.
- Briefly describe concepts related to document processing.
- Illustrate text mining with simple examples.

Text Analytics

- *Text analytics* includes **applications** and **algorithms** for turning text into data and analyzing the data using **statistical methods** and **natural language processing**.
- Previously text analytics was nearly synonymous with text data mining, but evolved to include natural language processing techniques for **extracting topics** and **summarizing content**.
- *Text data mining* (in the remainder of this course, referred to simply as *text mining*) matured and currently includes algorithms from **natural language processing** and **machine learning**. Text mining is classified as a subset of text analytics.

4

This course focuses on the use of SAS Enterprise Miner and SAS Text Miner. SAS has a rich set of text analytic products. Visit **support.sas.com** for information about the latest text analytic offerings. Other courses discuss topics related to products such as SAS Enterprise Content Categorization.

Text Analytics

- Concept Extraction
- Summarization
- Categorization
- Sentiment Analysis
- Content Management
- Ontology Management

5

Sentiment Analysis: Example

Good, but **a little outdated**. I bought the **ACME Z37** as my first digital compact P&S camera. I had it for a couple of weeks, until mine had a '**lens error**' that basically **made the camera inoperable** (it was stuck open). It might've been due to batteries running low, but I tried another set (which I now think was also low).

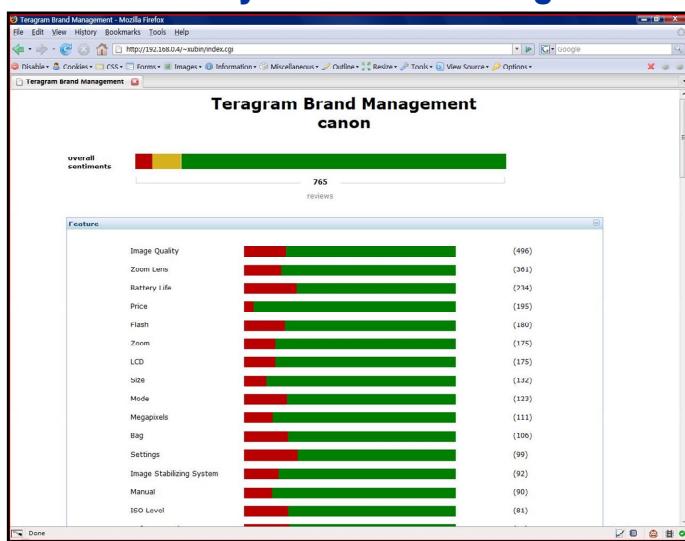
The picture quality from the **Z37** was **very good**, a **bit of barrel distortion** was noticed in the wide angle and shooting tall skyscrapers (noticed by the curve along the side of the frame where the buildings are supposed to be straight). Another gripe I had with the camera was **how slow the auto-focus was**. It would basically go through the whole range of focus every time I pressed the shutter half-way and then some. This became more **annoying** the more I used it.

Eventually a lot of my **pictures came out blurry**, including outdoor overcast days with 3x optical zoom. Basically anytime there's zoom & less than ideal lighting, I would have to have rock steady hands to get non-blurry pictures. **Overall it is a good camera** if you can **overlook the issues I mentioned**.

6 Camera review with positive and negative comments

Sentiment analysis categorizes a document as being positive, negative, or neutral about a specific topic. In the above review, positive associations are depicted in blue, and negative associations are highlighted in red. (In gray scale, match the shading associated with the words positive and negative in the caption.)

Sentiment Analysis: Brand Management



7

Sentiment analysis permits visualizing of the sentiments about features of a product. In the above example related to a compact digital camera, red is bad and green is good. Overall, the sentiment is favorable to the camera. (In gray scale, red appears as the darker shade on the left of each bar.) The example uses SAS Sentiment Analysis software, which is addressed in a separate course.

Text Analytics

- Concept Extraction
 - Summarization
 - Categorization
 - Sentiment Analysis
 - Content Management
 - Ontology Management
- }
- Text Mining

8

While text analytics applications often share methodologies and technologies, the field is partitioned into primary areas of focus. Text mining deals with more general or abstract applications, such as information retrieval and text categorization. The distinction is often subtle, and as a discipline, text mining is somewhat ill-defined. To facilitate understanding, you can think of text mining as “what SAS Text Miner does.” Obviously, SAS Text Miner does not include every text mining algorithm or methodology, but this course addresses the text mining applications that can be performed using SAS Text Miner.

Text Mining: Definition

Text mining is the process of **discovering** and **extracting** meaningful patterns and relationships from text collections.



9

The above definition implies that text mining focuses on pattern discovery. Because the topics, themes, and concepts derived from text mining can be used as inputs to a predictive model, text mining includes the two major components of data mining, namely **pattern discovery** and **predictive modeling**.

Some references that struggle with defining text mining as a discipline instead choose to describe what text mining is *not*. For example, text mining is not natural language processing. Text mining does not give you the ability to have your computer read and understand documents. To understand text mining, it helps to understand the types of problems that text mining addresses.

Text Mining

Text mining has the following characteristics:

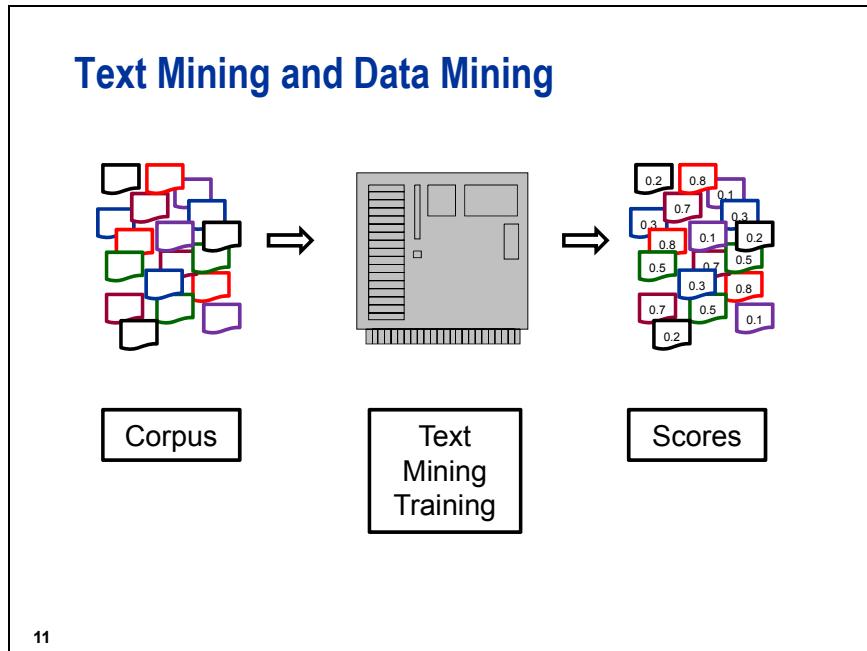
- operates with respect to a **corpus** of documents
- uses a **dictionary** or **vocabulary** to identify relevant terms
- accommodates a variety of **metrics** to quantify the contents of a document within the corpus
- derives a **structured vector*** of measurements for each document relative to the corpus
- uses **analytical methods** applied to the structured vector of measurements based on the goals of the analysis, for example, groups documents into segments

* Some text mining methods use a structured **matrix**.

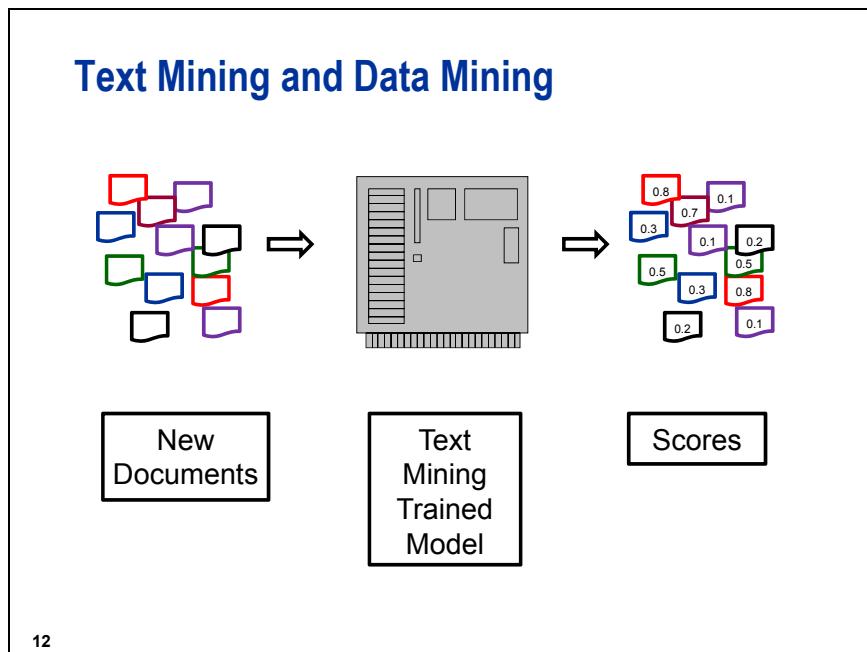
10

The concept of a dictionary can be thought of as a *vocabulary*. The vocabulary of a person is almost as good as DNA in identifying the person. The document collection has a vocabulary that is the union of all of the vocabularies exhibited by each document. Consequently, text mining references will use **dictionary** or **vocabulary** to refer to the collection of terms that are used in the analysis. Terms not in the dictionary are ignored, except possibly for use in determining the relative frequencies of terms in each document. Zipf's Law, discussed in a later chapter, helps to identify terms in a dictionary that should be included in an analysis. SAS Text Miner refers to the derived dictionary or vocabulary to be used for an analysis as the *start list*.

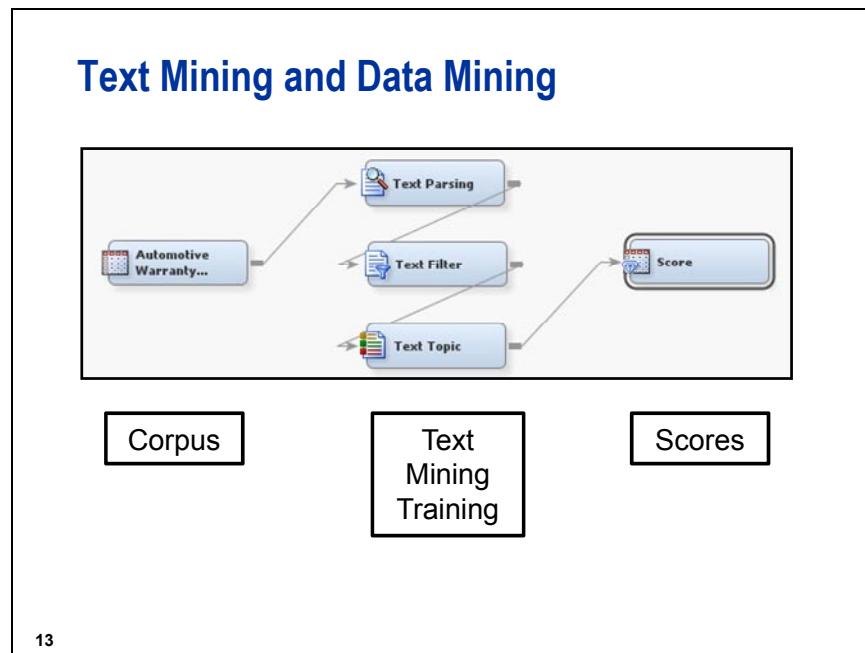
One way to distinguish text mining from other areas of text analytics is to examine the nature of text that is analyzed. Text mining works with a collection of documents. The collection can be dynamic, in that documents can be added to the collection. You can use the collection to train a model, and you can apply the model to new documents coming into the collection. The solutions coming from text mining will be relative to the document collection. Any insight into a single document in the collection will be relative to the entire collection, and will be unrelated to any document not in the collection. New documents coming into the collection will be scored relative to how the new documents compare to the documents in the collection. If a new document contains a new concept, then text mining will be ignorant of this new concept until that document is employed in training.



Text mining training can be exclusive to SAS Text Miner and not rely on any other SAS Enterprise Miner node. Alternatively, SAS Text Miner nodes export data, and this data can be imported into pattern discovery and predictive modeling nodes of SAS Enterprise Miner. Thus, a trained model can be obtained by using a combination of SAS Text Miner nodes and SAS Enterprise Miner nodes. While many commercial text mining products have strong text analytics capabilities, most lack data mining capabilities beyond text analytics. The ability to score new documents using a decision tree or a neural network presents new opportunities to improve text mining outcomes – for example, making it easier to improve accuracy for predictive models.



As new documents appear, they can be scored using the model trained on the original corpus. Eventually, the model can be updated by being re-trained on the corpus with the new documents added.



13

The above process flow shows how SAS Enterprise Miner can be used to implement a text mining solution. The Text Parsing, Text Filter, and Text Topic nodes are part of SAS Text Miner.

Most data mining professionals recognize how a predictive model scores new data. However, some professionals might be unaware that unsupervised learning models also produce scores, and new data can be scored using the model that was produced using unsupervised learning methods. For example, the Text Cluster node divides a document collection into mutually exclusive clusters. A new document is scored by calculating the probability of membership in each cluster, and then it is assigned to the cluster associated with the highest probability.

As suggested above, data mining is often described with respect to two general application areas: pattern discovery (unsupervised learning) and predictive modeling (supervised learning). Specific examples of these two abstract applications are in this course, and the two abstract areas are examined in the general context of data mining.

Text Mining Application Areas

- Stylometry (determining authorship)
 - Are documents created by more than one author? (Pattern discovery)
 - Who wrote a given document? (Prediction)
- Document categorization
 - Do the documents separate naturally into different categories? (Pattern discovery)
 - Can you assign a new document to a subject matter category? (Prediction)
- Information retrieval
 - Which documents are most relevant for a given information request? (Pattern discovery and prediction)

Stylometry is a specific application that is rarely described in detail in text mining references; perhaps because it is one of the easier applications of text mining. Most text mining references focus on document categorization and information retrieval.

Text Mining Application Areas

- Anomaly detection
 - Are there any unusual documents in the collection? (Pattern discovery and prediction)
 - What makes a document unusual? (Pattern discovery)
- Forensic linguistics
 - Can you identify the author of a manifesto? (Prediction)
 - This application area applies stylometry to crime investigation, and is related to anomaly detection for crime prevention.

15

Anomaly detection is one of the newer applications of text mining. *Forensic linguistics* can be thought of as a special case of stylometry.

Text mining application areas are not mutually exclusive. Experts differ in how they characterize the applications of text mining. For example, Berry and Castellanos (2008) divide text mining into the following general areas: clustering, document retrieval and representation, e-mail surveillance and filtering, and anomaly detection. Konchady (2006) considers the areas of information extraction, search engines, searching the Web, clustering, categorization, summarization, and [automating] question and answer applications. Feldman and Sanger (2007) address categorization, clustering, and information extraction.

1.01 Multiple Answer Poll

Select the type (or types) of text data sets that you are interested in analyzing.

- a. open-ended survey responses
- b. call center requests for information
- c. customer support requests for help
- d. formal technical documents
- e. informal voice or written communication
- f. e-mail and Internet communication
- g. World Wide Web content
- h. Domain-specific text, such as patient medical reports and descriptions of warranty claims
- i. other

17

To illustrate text mining, consider a problem in categorizing open-ended responses to a survey question.

Example: Open-Ended Survey Question

A company conducts annual surveys to solicit information about employee satisfaction.

Text responses are to an open-ended question about the quality of communication within the organization.

The next chapter addresses how to collect documents in a corpus into a single SAS table using the Text Import node. For most demonstrations in this course, the task of creating the document data set has already been performed. The open-ended survey questions originated as data extracted from a Web-based application. Employees went to a company Web page and responded dynamically to survey questions. The original responses were captured as individual HTML files. A Text Import node was used to process the individual HTML files to produce a single SAS table.

Example: Open-Ended Survey Question

The screenshot shows the 'Interactive Filter Viewer' window. At the top, there's a search bar and buttons for 'Apply' and 'Clear'. Below that is a table titled 'Documents' with columns 'TEXT' and 'ID'. The table contains 15 rows of survey responses. Underneath this is another table titled 'Terms' with columns 'TERM', 'FREQ', '# DOCS', 'KEEP ▼', 'WEIGHT', 'ROLE', and 'ATTRIBUTE'. This table lists various words from the survey with their frequency, document count, keep status, weight, role, and attribute.

19

continued...

The Text Filter node provides a mechanism for viewing each document. By default, you get the first part of the document, but you can use the Toggle Show Full Text option and view the complete document.

Example: Open-Ended Survey Question

The screenshot shows a table with two columns: 'TEXT' and 'ID'. The 'TEXT' column contains 15 survey responses, and the 'ID' column contains their corresponding IDs. The responses are identical to those shown in the previous screenshot.

20

continued...

Dilbert is a comic strip character who represents an office worker in a dysfunctional organization. Many of the responses to the survey question seem to be negative. You could flag each response using a Likert scale, for example, 1=Negative, 2=Slightly Negative, 3=Neutral, 4=Slightly Positive, and 5=Positive, and then train a predictive model to classify the responses. While this can be done using SAS Text Miner and SAS Enterprise Miner, a customized version of this approach is part of SAS Sentiment Analysis.

Example: Open-Ended Survey Question

Identify responses related to e-mail.

TEXT ▲	TEXTFILTER_SNIPPET
Better spam filter for email.	... spam filter for email ...
How about corporate email guidelines? Like eliminating emails that reply-to-all	... How about corporate email
How can a company rely so much on computers and email, when Microsoft	... on computers and email ,
I am all for not killing trees, but could we also try to save a few electrons as	... the company propaganda
I do not read corporate propaganda emails if they are longer than a few	... read corporate propaganda
I hate getting email from my boss's blackberry. Most can wait until	... I hate getting email from my
I have to live in the Outlook personal folders because I am always deleting or	... or moving old emails ...
I like being able to email my boss when PERSONAL_PRONOUN is traveling.	... being able to email my boss
I swear to God, if you send me one more EXPLETIVE , email copying your	... one more EXPLETIVE
I wish managers would grow a pair and come speak to me face to face. Do	... hide behind an email ...
If you could teach my boss how to use email and how to use Google, I could	... how to use email and how to
Make people use email.	... Make people use email ...
My boss uses email to document everything. Talking would be better. I get	... My boss uses email to
My job is to answer the phone. The phones barely ring with email and texting	... barely ring with email and
Network is always down. No internet when you need it. Outlook server cannot	... so no email access ...
Our biggest problem is that most people do not know how to use email	... how to use email effectively ...
Outlook inbox is always filled with trash. Seems like I_T_ could do something	... do something about email .
Quantify email quantity versus email quality, and provide management	... Quantify email quantity
Some of our security features block emails from legitimate clients.	... security features block
This push to put everything on the Internet is just whiz-bang-gizmo and has	... People are sending emails to
Too many spam emails, even within company.	... Too many spam emails ,
We need more archival storage for older emails.	... storage for older emails ...
We should eliminate all of the automatic notification emails, and instead have	... the automatic notification
We should replace emails with detailed revenue breakdowns with terse	... We should replace emails
We should replace most of the internal corporate wide emails with status	... internal corporate wide

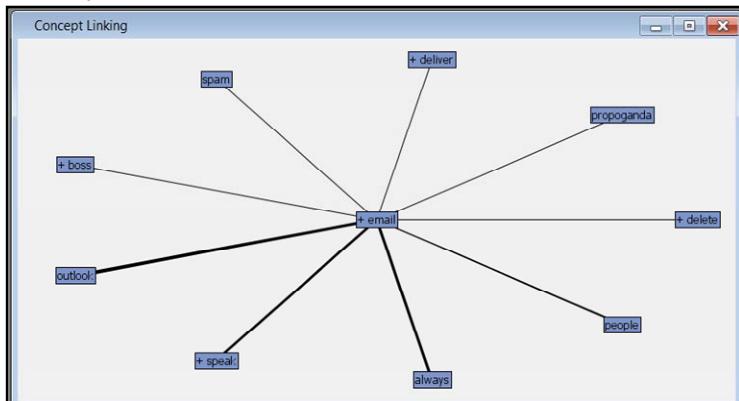
Filter on the term **email**.

continued...

The term *filter* refers to the action of selecting documents that have a specified set of characteristics, such as having one or more specified terms. However, a popular perception of a filter is actually what is called a *Boolean filter*. The Text Filter node uses a more sophisticated technique that has been described as *vector space matching*, a technique closely related to *latent semantic indexing*, which will be discussed later. This technique can produce documents that have zero occurrences of the keyword used in the query. How is this possible? Using sophisticated natural language processing algorithms, the Text Filter node “learns” that, for example, “cats and dogs” is a phrase that describes a weather condition, not domesticated animals, as in, “It was raining cats and dogs.” Thus, a query using the term “precipitation” might actually produce the document “It was raining cats and dogs.” Of course, false positives are possible, and a query on “dogs” might produce, “A heavy rainfall washed away evidence of the crime.” Because everything is relative to the corpus, results of a filter query will depend on terms and phrases used in the collection. If no document mentions “raining cats and dogs,” then SAS Text Miner cannot discern that “dogs” might be related to weather.

Example: Open-Ended Survey Question

Identify responses related to e-mail.



What other terms are used when employee responses use the term **e-mail**?

22

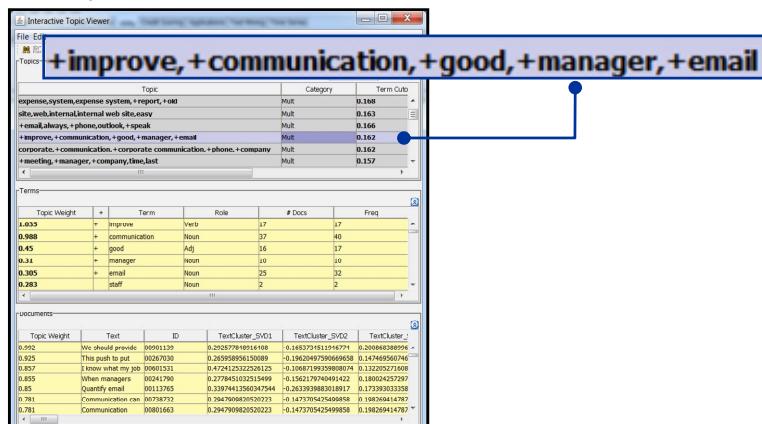
continued...

SAS Text Miner supplies a variety of tools for identifying terms that are related, based on their co-occurrences within documents or their higher order associations within documents. For example, if two terms tend to appear in documents having a third term, then the two terms have a second-order association through the third term even if they rarely co-occur in the same document. *Concept linking* is a particular scheme for identifying higher order associations. Concept linking is a technique from social network analysis. An original application of concept linking depicts the organization chart of a company with lines of varying thickness between executives and managers representing the strength of communication between company leaders.

SAS Text Miner can identify topics, themes, or concepts within a document collection. You can also provide a custom definition of a topic and permit SAS Text Miner to determine whether the topic is represented in the collection.

Example: Open-Ended Survey Question

Identify responses related to e-mail.



Are there any underlying topics, themes, or concepts related to the term **e-mail**?

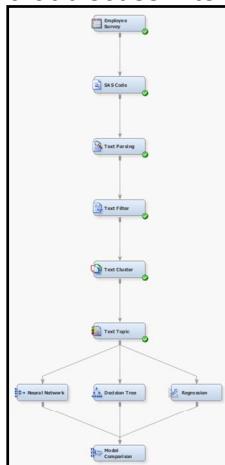
23

continued...

If you can tag each document with a relevant label, then you can use the tag as a categorical target variable and build a predictive model to score new documents.

Example: Open-Ended Survey Question

Can you train a model to automatically extract responses that discuss Internet-based communication?



Which model produces the best accuracy?
Is the accuracy adequate for implementing the model? (**One in five are misclassified.**)

Fit Statistics			
Selected Model	Model Node	Model Description	Valid: Misclassification Rate
Y	Neural	Neural Network	0.204993
	Tree	Decision Tree	0.235217
	Reg	Regression	0.247043

continued...

The above example shows the results for three predictive models, with a neural network model chosen as the winner. A neural network model classifies a document with respect to content related to Internet communications with 80% accuracy.

Example: Open-Ended Survey Question

Can documents be separated into mutually exclusive categories?

Cluster ID	Descriptive Terms
1	+bring +effort damn executive productive wrong management time
2	+fire +profit +report god head +speak back _persons_name -
3	+boss +decide +eliminate +office projects wait hate +look
4	'communication problems' +problem action caused eliminated problems solve fix
5	'analog technology' analog digital forcing technology required solutions replace
6	'internal web site' 'web site' +client +explain does easy site internal
7	'communications paradigm' +internal communication' +competitor competitors paradigm communications customers required
8	'dead workers' dead executives worked companies dilbert
9	+agent computers +agent computer application online info
10	+head brought division great personality weedy work years
11	+hope +require +spend phone marketing day people 'corporate communications'
12	'cell phones' phones cell internet bosses people effectively home
13	'corporate position' guidelines news position better corporate staff sales
14	'expense system' +allow +delete expense limits propaganda reports +few
15	+idea organization sha software improved +improve +communication english
16	'internet connection' +bit always connection empowered ideas talking improvement
17	+happy +hear gossip press rumors run negative managers

25

The Text Cluster node derives groups of documents that can satisfy analytic objectives related to information retrieval and text categorization. A set of descriptive terms helps you decide what each cluster contains.

Because SAS Text Miner is a component of SAS Enterprise Miner, a software solution that implements data mining, a review of data mining is in order.

Data Mining

“Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.”

– Hand, Mannila, and Smyth (2001)

“Data mining is the process of posing queries to large quantities of data and extracting information, often previously unknown, using mathematical, statistical, and machine learning techniques.”

– Thuraisingham, Khan, Awad, and Wang (2009)

26

As the above two definitions illustrate, getting professionals to agree on the definition of data mining is challenging. The first definition emphasizes analysis, whereas the second definition emphasizes querying. Statisticians might have a problem with the second definition because statistical techniques are not tools for querying, but rather they are tools that are *applied* to queries. A query is simply a derived data table obtained from operations performed on one or more source data tables. Statistical analysis can lead to refined queries, and proper queries are essential to satisfying an analytic objective.

Phrases such as “unsuspected relationships” are potentially misleading. Verifying a suspected relationship is perfectly acceptable. However, Hand, Mannila, and Smyth (2001) imply that the confirmatory analysis of hypotheses or conjectures (suspected relationships) often falls in the realm of the analysis of experimental data rather than observational data. Experimental data derives from controlled experiments that are designed to test hypotheses of interest, whereas observational data derives from a process that is unrelated to the analysis of the data, such as a business process that collects operational data.

Simply stated, ***data mining is statistics for (large) observational data sets***. Because the statistical sciences include the study of algorithms implemented on digital computers (Thisted 1988), techniques from computing science designed for the efficient processing of massive data sets are included. However, many contributions to the methods of statistical analysis of observational data are from non-statisticians. Consequently, most discussions of data mining emphasize the interaction between multiple disciplines, including statistics, machine learning, artificial intelligence, and cognitive science. Furthermore, specific disciplines, such as psychometrics, contribute an abundance of techniques for the visualization of data, a cornerstone of any data mining solution.

Some data mining references claim to be general, but instead focus on a small subset of data mining, such as constructing queries that result in hypercubes of data. Online analytical processing (OLAP) constructs hypercubes that facilitate rapid querying of data sets. Some references are misleading. OLAP represents a tiny portion of the field of data mining, but a reference might imply that OLAP *is* data mining. A common refrain from data mining professionals is that, “Data mining is ***not*** querying and reporting.” For a good comprehensive reference about data mining, see Hand, Mannila, and Smyth (2001).

Data mining can be separated into two application areas.

Data Mining: Two General Goals

- Pattern Discovery (Unsupervised Learning)
 - Identify naturally occurring groups (classification*).
 - Derive convenient segments (clustering).
- Prediction (Supervised Learning)
 - Input variables are associated with values of a target variable.
 - Derive a model or set of rules that produces a predicted target value for a given set of inputs.

* Classification with a target variable is prediction.

27

1.02 Multiple Choice Poll

What type of text mining will you perform?

- a. pattern discovery
- b. prediction
- c. both pattern discovery and prediction
- d. not sure

29

Pattern Discovery

The Essence of Data Mining?



***“...the discovery of interesting,
unexpected, or valuable
structures in large data sets.”***

– David Hand

30

...

Pattern Discovery

The Essence of Data Mining?



***“...the discovery of interesting,
unexpected, or valuable
structures in large data sets.”***

– David Hand

***“If you’ve got terabytes of data, and you’re
relying on data mining to find interesting
things in there for you, you’ve lost before
you’ve even begun.”***

– Herb Edelstein

31

A corollary to Edelstein’s quote is, “If you don’t know what you are looking for, then you probably won’t find it.”

Data Mining: Signal versus Noise

Predictive Modeling

- Target = Signal + Noise
- Signal = Systematic Variation = Predictable
- Noise = Random Variation = Unpredictable

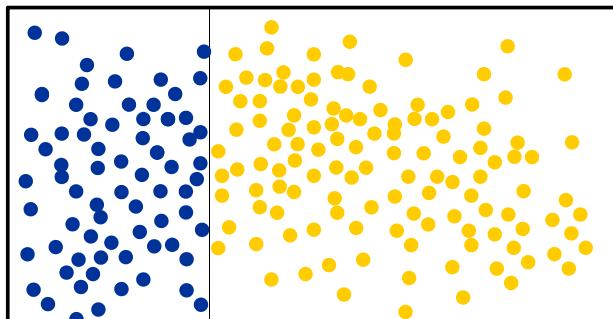
32

Users new to the world of analytics often have a naïve notion about noise. Science fiction movies include computers and robots that speak and understand human languages. Television police dramas have detectives making perfect predictions about where crimes will occur. The reality is that noise permeates existence. Rarely will predictions be exactly correct. You might have an expectation that after you master text mining you can perfectly predict customer behavior based on responses to an online survey. This expectation is unrealistic.

Psychologists know that human beings might react differently to the exact same stimulus if sufficient time elapses between exposures. On Monday when you are hungry at lunchtime, you eat a sandwich. Yet on Tuesday when you are hungry, you opt for a salad. This tendency for different outcomes to occur with similar inputs is attributed to noise, something that is unpredictable. You can predict with almost certainty that you will eat lunch next Thursday, but you cannot predict what you will eat with the same certainty. (Of course, if you bet someone a million dollars that you will eat a spinach salad next Thursday, then you will almost certainly eat a spinach salad!) Analytic experts expect errors in prediction related to noise, so methods are developed to minimize error in the presence of noise. You cannot judge the success or failure of text mining simply by observing whether you are 100% accurate. Instead, you must compare the accuracy achieved by adding text mining to your toolkit to the accuracy you experienced before using text mining.

A quick example illustrates how expectations should be set relative to the environment where the predictions will be used. Without data mining, an insurance company finds sufficient evidence of fraud to pursue legal action in about 2 out of 100 cases investigated. After adding data mining and text mining to the process, about 10 in 100 cases result in legal action. The company is still wrong 90% of the time, but the company identifies 5 times as many fraud cases using analytics.

Pure Separation = Pure Signal = No Noise



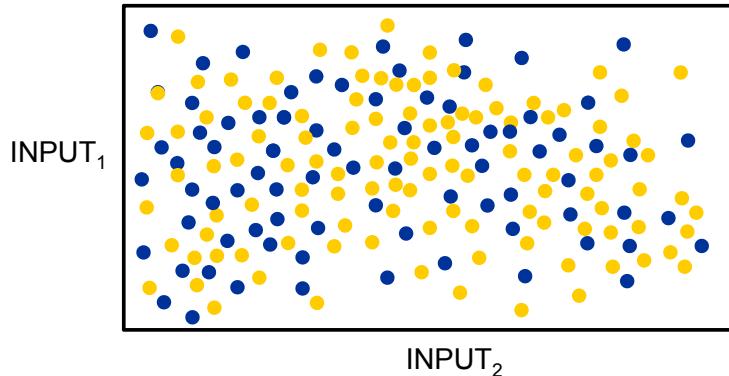
INPUT

Target: Primary Outcome=● Secondary Outcome=●

33

The above graphic illustrates the pure signal situation. You should never expect to see this in practice. Unfortunately, many of those who are new to text mining are disappointed when the methods do not perfectly categorize documents with 100% accuracy.

No Separation = Pure Noise = No Signal

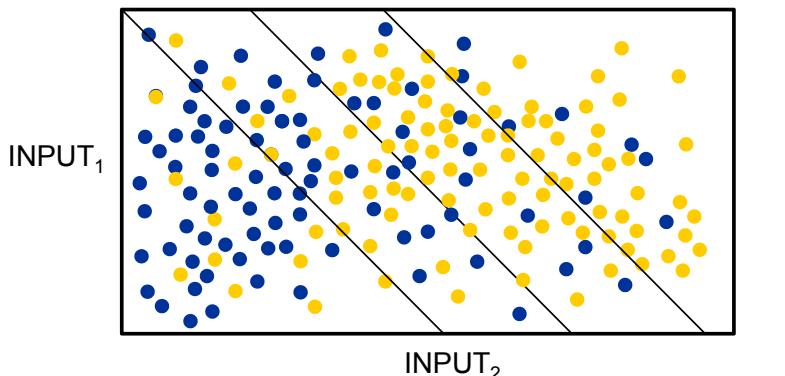
INPUT₁INPUT₂

Target: Primary Outcome=● Secondary Outcome=●

34

At the other extreme is the pure noise situation. This situation is more common than you might like. While pure signal is essentially impossible, pure noise can actually occur in practice.

Some Separation = Signal + Noise

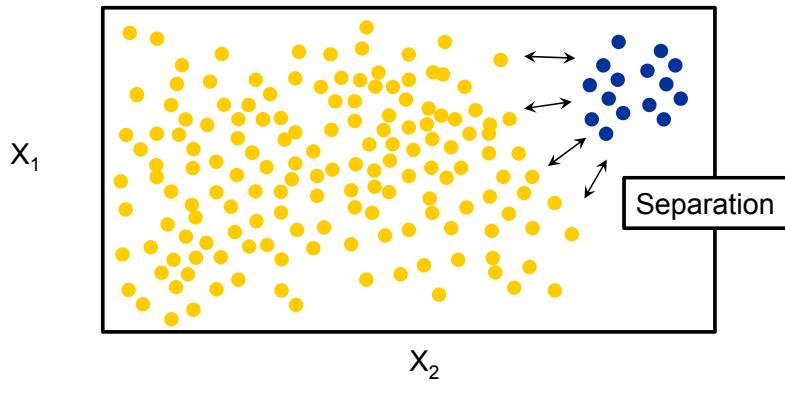


Target: Primary Outcome=● Secondary Outcome=●

35

The most common situation in practice is a mixture of signal and noise. You can predict more accurately than randomly guessing, but how well you predict depends on whether data is dominated by systematic variation or random variation.

Unsupervised Classification: Fraud Cases?

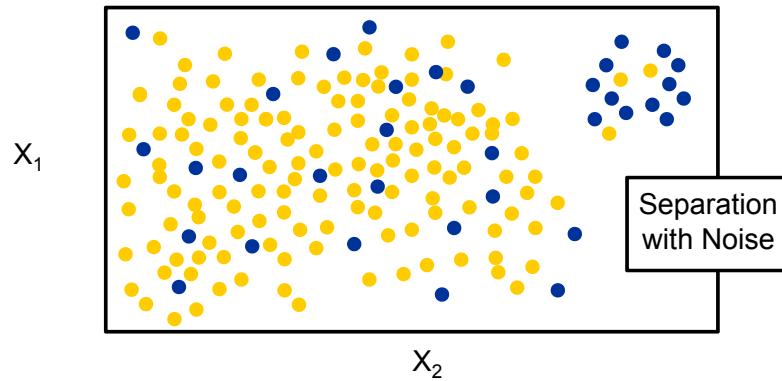


X₁=Distance to Physician X₂=Ratio BI/PD

36

When no target variable is available, you can still investigate whether a natural separation occurs in the data with respect to the analytic objective. For example, fraud cases are often unusual in higher dimensional space because the human beings that commit fraud have difficulty controlling outcomes so that they look normal in many dimensions.

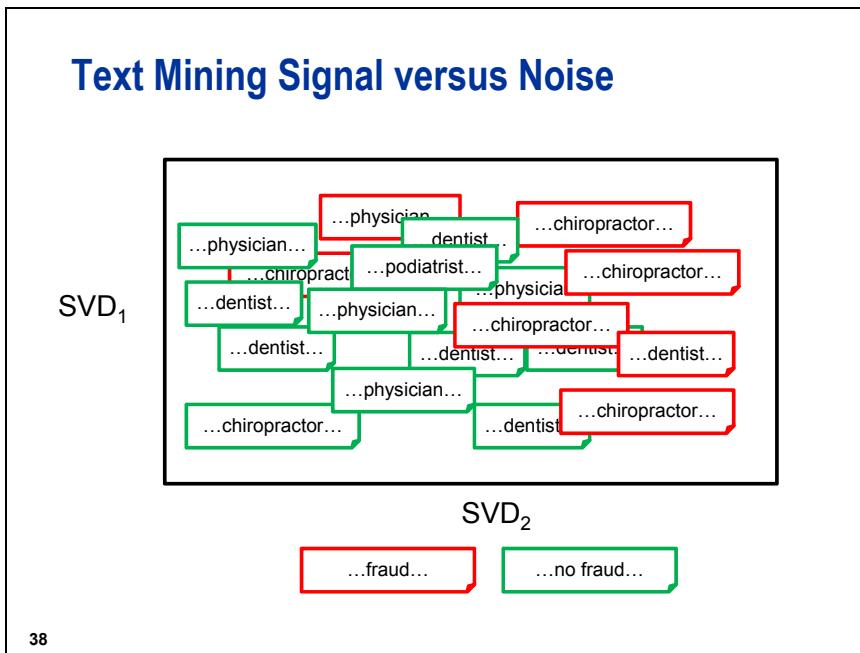
Unsupervised Classification: Actual Fraud



X_1 =Distance to Physician X_2 =Ratio BI/PD

37

Even with good separation, noise is usually present. For the fraud example, most of the claims with a long distance from claimant to physician and a high ratio of bodily injury to property damage costs are fraudulent, but a few are legitimate cases. Other fraud cases are not so separated, perhaps because the fraudulent physician just happened to have a practice near the claimant's home.



A string of fraud rings operating in Southern California in the 1990s had the common elements of a lawyer, a chiropractor, and a recruiter. The recruiter would approach people getting unemployment benefits and tell them that they could get workers compensation benefits from their previous employers. The recruiter would refer a candidate to an unscrupulous lawyer, who would schedule treatments with a chiropractor, a partner in the fraud ring. After a few weeks of treatments, the lawyer would file a claim for three to five times the chiropractor bills, a fairly common practice in insurance litigation. Claims adjusters would often receive training in fraud prevention. When news of the fraud rings was disseminated, a claims adjuster might add a comment to the adjuster notes when unusual activity involving claimant representation by a lawyer and incoming chiropractor bills became known. The above slide illustrates the notes that mentioned chiropractors tend to be color coded to indicate fraud, while notes that mentioned other medical professionals, like dentists, tended to be legitimate.

Text Mining Signal versus Noise

Document ID	Marine Mammal Words	Land Mammal Words	Topic Target
1	3	0	1
2	5	0	1
3	2	0	1
4	8	0	1
5	0	1	0
6	0	5	0
7	0	3	0
8	0	7	0

Perfect Separation: Pure Signal/No Noise

39

Some document collections are well separated. The hypothetical example above shows eight documents, with four describing marine mammals, and the remaining four describing land mammals. A topic related to marine mammals is easily derived from the collection.

Text Mining Signal versus Noise

Document ID	Marine Mammal Words	Land Mammal Words	Topic Target
1	3	1	1
2	5	3	0
3	2	0	1
4	8	2	1
5	0	1	0
6	1	5	0
7	2	3	1
8	3	7	0

Good Separation: High Signal/Noise Ratio

40

With the same topic and analytic objective, another document collection has documents that might mention a heterogeneous set of animals. You still get good separation, but noise creeps in due to the fact that a document can include multiple topics.

Text Mining Signal versus Noise

Document ID	Marine Mammal Words	Land Mammal Words	Topic Target
1	3	1	1
2	5	3	0
3	2	0	0
4	8	2	1
5	0	1	1
6	3	3	1
7	2	3	0
8	6	7	0

Poor Separation: Low Signal/Noise Ratio

41

Finally, the above example shows that if you have a collection of documents that mention many topics in a non-systematic way, then trying to classify documents into categories will be difficult.

Assessing Text Mining Results

Objectives

- Information Retrieval (IR) Outcomes
 - True Positives: You want to retrieve the information that was requested.
 - False Positives: You want to avoid retrieving information that is unrelated to the request.
 - False Negatives: You want to avoid omitting information related to the request.
 - True Negatives: You do not want to retrieve information that is irrelevant.
- Text Categorization Outcomes
 - Same general outcomes as IR

42

Information retrieval (IR) and text categorization can be assessed the same way that you often assess binary response models, by looking at statistics based on correct classification (true positives/negatives) and incorrect classification (false positives/negatives).

Assessing Text Mining Results

Notation

ND=Number of documents

NS=Number of documents selected (retrieved)

TP=Number of true positives

FP=Number of false positives

TN=Number of true negatives

FN=Number of false negatives

		Action	
		Retrieved	Omitted
Result	Contain Info	TP	FN
	Absent Info	FP	TN
	Column Totals	NS	ND-NS

43

Assessing Text Mining Results

Misclassification Rate=Fraction misclassified

$$\text{_MISC_=}(FP+FN)/ND$$

Precision=Fraction of retrieved documents that are relevant

$$\text{Precision}=TP/(TP+FP)$$

Recall=Fraction of relevant documents that are retrieved

$$\text{Recall}=TP/(TP+FN)$$

		Action	
		Retrieved	Omitted
Result	Contain Info	TP	FN
	Absent Info	FP	TN
	Column Totals	NS	ND-NS

44

Assessing Text Mining Results

Other Measures

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN})$$

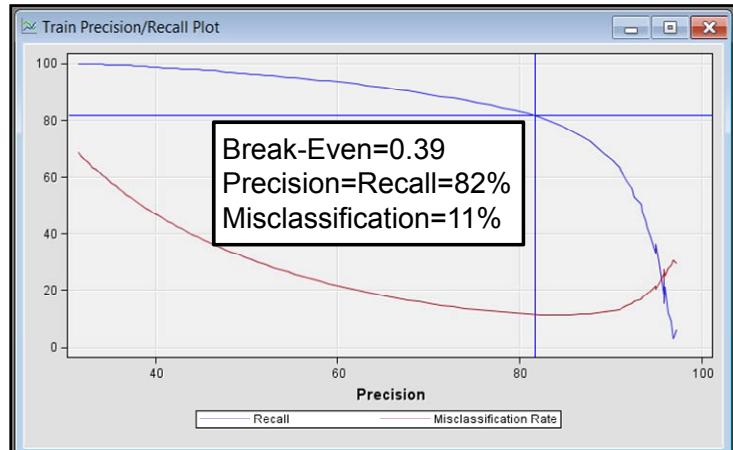
$$\text{Specificity} = \text{TN}/(\text{FP} + \text{FN})$$

Cutoff: A model scores each document relative to the category/query of interest. A cutoff is established so that documents having a score above the cutoff are retrieved and those having a score below the cutoff are omitted.

Break-even point: Value of the cutoff that produces the same precision and recall statistics.

45

Precision and Recall



If the posterior probability of category membership exceeds 0.39, classify the document into the category.

46

General predictive modeling software has statistics such as misclassification rate and plots such as the receiver operating characteristic (ROC) curve that depend on misclassification rates. However, statistics such as precision and recall are less common and are primarily employed in text mining related to text categorization.

Text Mining: Pattern Discovery Examples

- Automotive warranty early warning systems - potentially fatal manufacturing problems exposed through the analysis of warranty claims – for example, the Transportation Recall Enhancement, Accountability and Documentation (or TREAD) Act
- Marketing studies using online communities - consumer perceptions – for example, Communispace (www.communispace.com)
- Anomaly detection - identifying target behavior without a target variable – for example, Aviation Safety Reporting System (ASRS)

47

As described earlier, Berry and Castellanos (2008) place clustering and anomaly detection in separate parts of their text mining reference, but careful reading of the book shows that there is substantial overlap in the areas. Clustering techniques can be applied to all three examples listed above.

Warranty Analysis

Analytic Objective:

- Automatically assign claims to categories
- Identify claims that can potentially pose a serious safety threat

Corpus:

- Text descriptions of warranty problem
- Some descriptions also mention the solution

48

continued...

Warranty Analysis

Sample Entries

Popping noise when turning front end all tight call tech line no help, replaced Right front strut drive no noise.

Check front speakers cut in and out, check speakers, check antenna clean and tighten, front speakers connections.

Brakes make sq noise when applied and a thumping noise on turns rear shows and backing plate dry lube steering bolt loose secur.

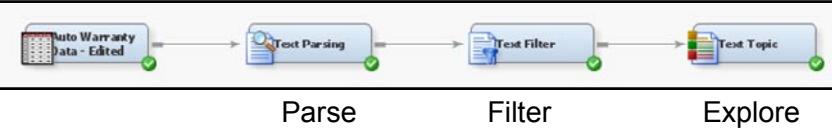
49

continued...

Warranty claims exhibit misspellings and nonstandard abbreviations.

A typical analysis first parses the documents into words, phrases, and entities. Then the dictionary of parsed items is filtered to try to include only words, phrases, or entities that are useful for the analysis. Documents can also be filtered out of the flow. Finally, naturally occurring topics are identified and described using well-chosen keywords.

Warranty Analysis



50

continued...

The Text Topic node finds topics or concepts in the document collection. When ten topics are selected, two topics exhibit warranty problems with braking and steering, which could potentially pose serious safety threats. Topics 3 and 7 below should be investigated for the potential for screening important warranty cases for scrutiny by experts.

Warranty Analysis

Topic ID	Topic
1	check,+light,engine,check engine light,+code
2	rear,spoiler,rear spoiler,+crack,hatch
3	+brake+light,+crack,third,+third break light
4	inop,power,+outlet,+fuse,+power outlet
5	fuel,gauge,fuel gauge,working,+unit
6	+pull,+vehicle,left,right,+car
7	+align,+wheel,+clean+steer,+pull
8	center,+install,+rattle,+clip,console
9	rear,seatbelt,+lock,right,rack
10	+stay,battery,+charge,+light,+start

Specifying 10 Topics

51

continued...

Often domain knowledge can be combined with the derived topics to construct topics more in line with a specific goal. The user can investigate and modify topics to suit project needs. The Interactive Topic Viewer in the Text Topic node enables you to explore custom and derived topics as well as modify topic definitions.

Warranty Analysis

The screenshot shows the "Interactive Topic Viewer" window with three main sections:

- Topics:** A table listing derived topics with their weights, categories, term cutoffs, and document cutoffs. Topics include "un,+light,+cone,+engine,abs", "+fuel,+tank,+leak,gas,+pump", "+fire,+catch,+vehicle,+park,+start", and "+transmission,+gear,+shift,+slip,+go".
- Terms:** A table showing terms with their topic weight, role, document count, and frequency. Terms include "fire" (Noun, 759 docs, 820 freq), "catch" (Verb, 434 docs, 435 freq), "vehicle" (Noun, 5171 docs, 5739 freq), "park" (Verb, 429 docs, 438 freq), and "start" (Verb, 824 docs, 855 freq).
- Documents:** A table listing repair documents with their topic weights, descriptions, claim IDs, mileage at repair, repair amounts, repair dates, and sales dates. Examples include "While vehicle was R000097685 16956 552.0 1997-11-26 1997-09-20", "Vehicle was parked, R000074771 919 38.0 1997-03-20 1993-11-07", and "Vehicle caught fire in T000135333 4506 153.0 1999-01-07 1998-12-24".

Interactive Topic Viewer

52

Text Mining: Prediction Examples

- Stylometry - identifying authorship
- Rapid retrieval of potentially beneficial medical information
- Detecting insurance fraud
- Automatic classification of safety reports
- Automatic classification of survey responses into negative versus positive categories

53

Stylometry

Stylometry: Determining authorship

Example: Authorship of the Federalist Papers

Analytic Objective: Determine who wrote 12 unattributed essays.

Author	Number of Essays
Alexander Hamilton	51
James Madison	14
John Jay	5
Hamilton and Madison	3
Unattributed	12
Total	85

54

continued...

Stylometry

Mosteller, F. and D. L. Wallace. 1964. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. New York: Springer-Verlag.

Using a dictionary of 70 words, classified the 12 unattributed essays as authored by Madison

Fung, Glenn. 2003. *Disputed Federalist Papers: SVM and Feature Selection via Concave Minimization*.

Using a dictionary of three words, classified the 12 unattributed essays as authored by Madison using a separating hyperplane

The stylometric investigation of the Federalist Papers by Mosteller and Wallace represents one of the first implementations of a text mining solution to a prediction problem. The fact that this problem can be solved in a few hours today by one researcher compared to about one year with a team of researchers and assistants in the early 1960s illustrates the incredible advances in computer performance and computer software over the past five decades.

Alexander Hamilton, James Madison, and John Jay wrote a series of essays in 1787 and 1788 to try to convince the citizens of the state of New York to ratify the new constitution of the United States. These essays are collectively called *The Federalist: A Collection of Essays*. Copies of the papers in a variety of formats can be found at avalon.law.yale.edu/subject_menus/fed.asp or www.constitution.org/fed/federa00.htm.

Of the 85 essays, 51 are attributed to Hamilton, 14 to Madison, 5 to Jay, and 3 to Hamilton and Madison jointly. The 12 remaining essays can be attributed only to Hamilton or Madison. Mosteller and Wallace (1964) used Bayesian statistical techniques to provide evidence that Madison wrote all 12 of the essays of unknown authorship. (The essays in question are numbers 49 through 58 and numbers 62 and 63.)

Stylometry

Corpus:

- 85 essays (For analysis, 77 essays = 51 Hamilton + 14 Madison + 12 Disputed)

Unique words:

- 8,752 (more than 190,000 words in the corpus)

56

continued...

The classification of the 12 unattributed essays spawned a competition of sorts, with data mining experts trying to perfectly classify the essays using as few words as possible. Fung (2003) uses a support vector machine (SVM) algorithm to perfectly separate the essays. His text mining inputs are relative counts of the three words *to*, *upon*, and *would*.

Stylometry

Fung (2003) Solution

Metrics and structured vector:

- Relative frequency of dictionary terms per 1000 terms in essay; three-word dictionary

Analytic Method:

- Separating hyperplane (Support Vector Machine)

$$0.5368to + 24.6634upon + 2.9532would = 66.6159$$

57

continued...

Stylometry

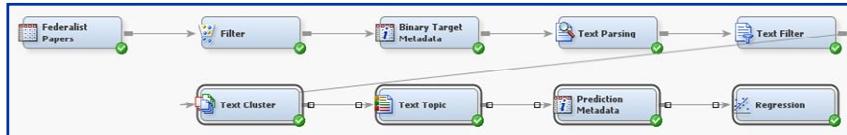
SAS Enterprise Miner/SAS Text Miner Solution

Metrics and structured vector:

- Inverse document frequency of dictionary terms; singular value decomposition (SVD); 5,000+ word dictionary; 29 SVD dimensions

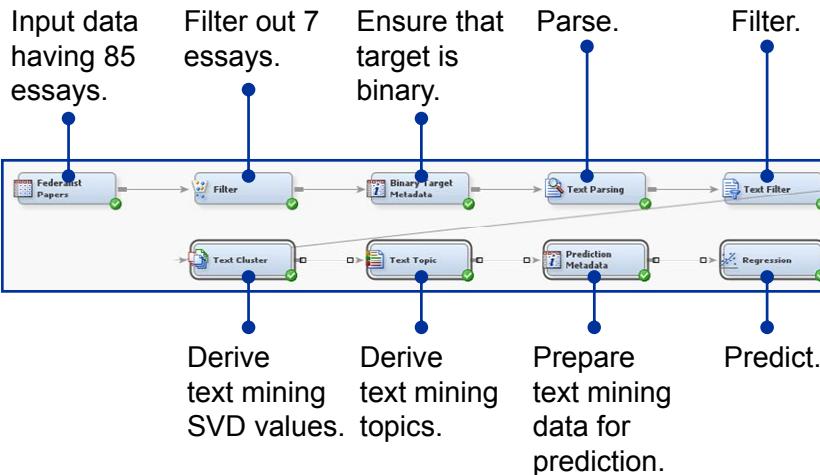
Analytic Method:

- Logistic regression

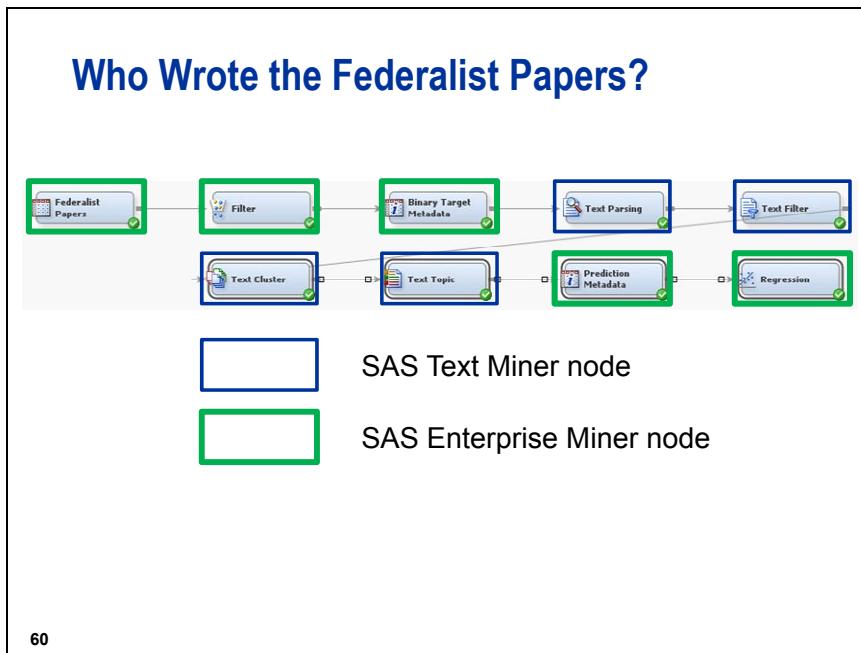


58

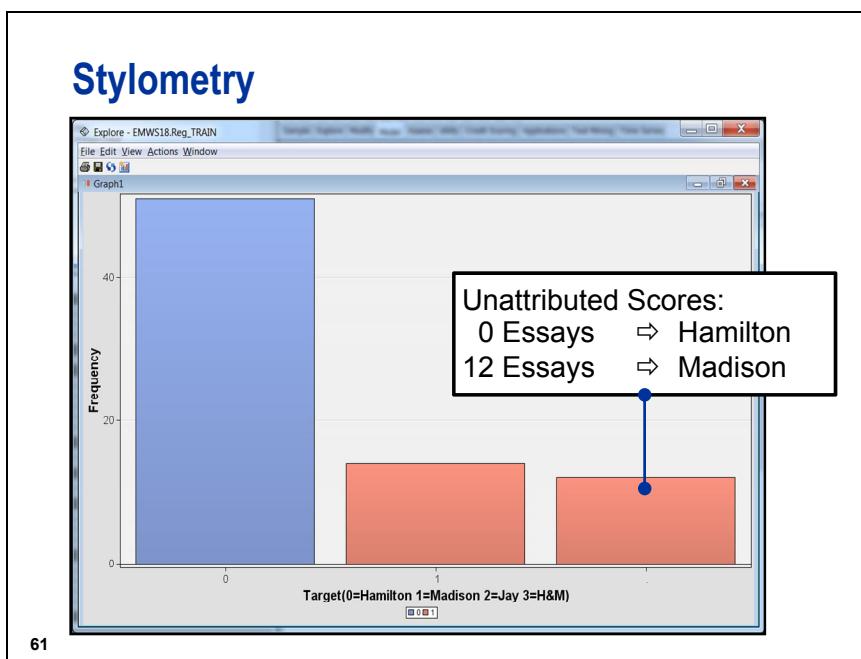
Who Wrote the Federalist Papers?



59



A typical process flow will have a mixture of SAS Text Miner nodes and SAS Enterprise Miner nodes.



The text mining approach produced results that are consistent with other analyses of the Federalist Papers. The general consensus is that Madison wrote all 12 unattributed essays.

Pedagogy versus Practice

The success of text mining the Federalist Papers was somewhat fortuitous. In practice, you often need to experiment with different options in the Text Miner nodes to try to optimize some measure of success, such as obtaining the lowest percentage of misclassified documents. Experimentation requires the use of validation and test data to prevent *overfitting*. Overfitting is the result of excessive experimentation that results in overly optimistic results that could not be achieved in practice when the true answer is unknown.

The results provided by other researchers can start with the answer and work toward getting results consistent with the answer. There is little to be gained by publishing failures. Mosteller and Wallace could not start with the answer because they were investigating what the answer should be. It is possible that many success stories for classifying the Federalist Papers are merely exercises in overfitting to obtain a preconceived conclusion.

In practice, you must verify the stability of a model by using holdout samples. A model that successfully extrapolates the holdout data can then be applied to score data to determine an appropriate classification. The Federalist Papers analysis should have separated the 12 contested essays as score data, constructed an acceptable model, and then scored the 12 essays using this model. One problem with the train/validation approach is that 85 essays is too small a sample to work with. Fung (2003) employs cross validation techniques to overcome the scarcity of data.

The Federalist Papers provide a useful educational exercise, but the analysis provided is certainly not representative of an analysis that should be used in practice. The forensic linguistics example given in a later chapter provides a practical solution to a problem that is valid in practice as well as being useful as an educational tool.



Text Mining the Federalist Papers

This demonstration illustrates how to use the Text Miner node to identify the author or authors of the 12 unattributed Federalist Papers.

- ✍ Some of the demonstrations in this course are “pre-cooked”; that is, a project diagram already exists in the course data that contains the complete demonstration. The instructor might use the pre-cooked version, or might choose to re-create the diagram using a different name. Creating and running a diagram can consume valuable class time.
- ✍ SAS Enterprise Miner requires the user to associate an analysis data set with metadata. The Data Source Wizard guides you through creation of the metadata table. After the metadata table has been created, the data is available as an input data source and can be placed in a SAS Enterprise Miner diagram. Most data sets in this course have already been set up as SAS Enterprise Miner input data sources. Demonstration instructions usually contain a preliminary step requesting, “If necessary, create an input data source for the demonstration data.” Chapter 2 provides details about creating data sets for text collections and setting up input data sources.

Access to SAS Enterprise Miner depends on the nature of your computing environment, whether attending a SAS classroom offering, participating in a Live Web class, or receiving instruction on-site at your organization. These course notes assume that you are using a workstation version of SAS Enterprise Miner with course data stored in the following folder:

Course Data Folder: D:\workshop\dmtxt51

Course SAS Program Folder: D:\workshop\dmtxt51\sassrc

1. Write the locations for your environment below:

Your Course Data Folder: _____

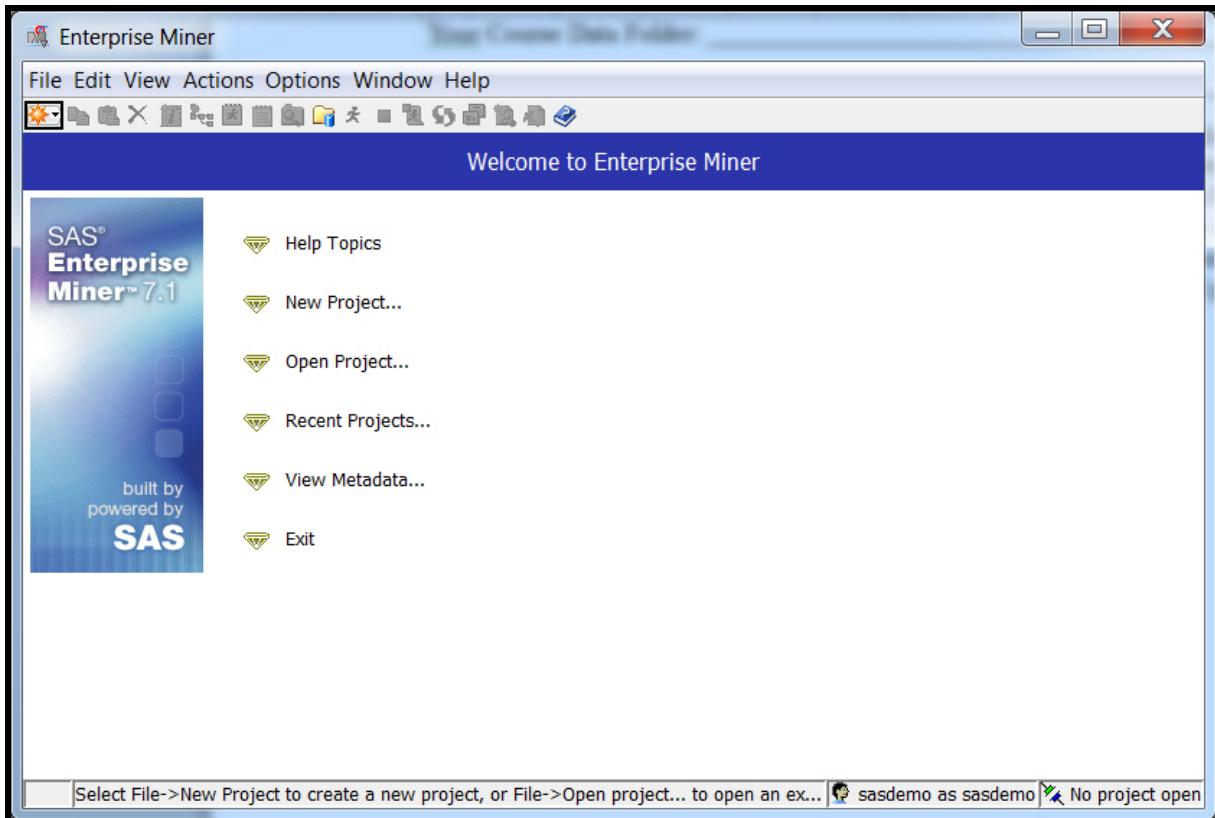
Your Course Program Folder: _____

You can access SAS Enterprise Miner in many different ways. For example, the Web-based access method uses a Web browser entry point. Some courses use a virtual lab setting to provide access to SAS Text Miner software. The server platform for the virtual lab contains both client and server installations on a single computer. The course data resides on a hard drive accessed directly by the server, that is, the data do not travel through a network connection. You might have to connect through some Web-based application to access the virtual lab. Your instructor will provide complete instructions along with the necessary logon credentials.

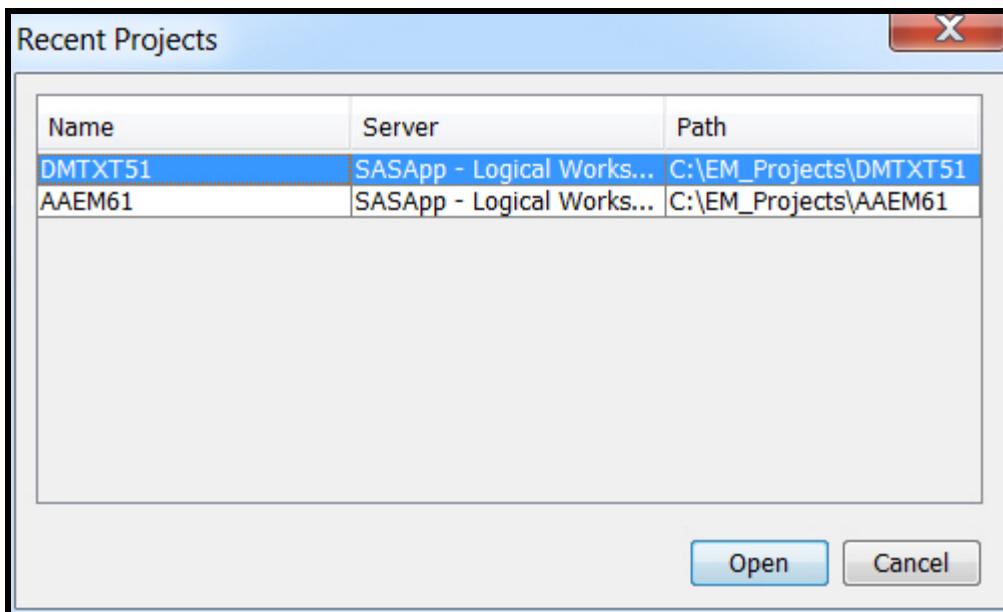
Following is a typical logon window for SAS Enterprise Miner 7.1.



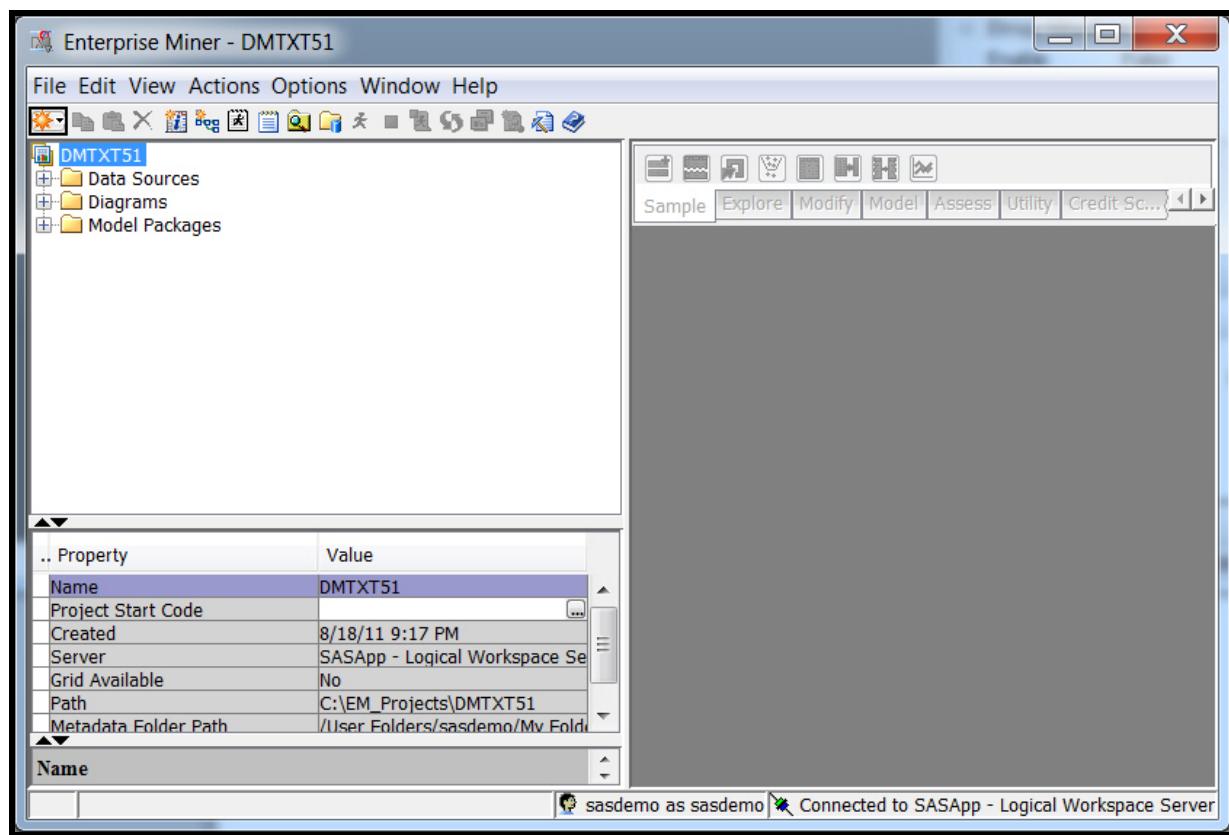
- To access the logon window, you typically select **Start** ⇒ **Programs** ⇒ **SAS** ⇒ **Analytics** ⇒ **SAS Enterprise Miner**. Usually, a shortcut for SAS Enterprise Miner is also on the desktop.
2. For most course environments, including those with a virtual lab, the logon fields will be populated, and you can just click **Log On**. If necessary, type your user ID and password and click **Log On**. The welcome window is shown below.



3. The typical SAS classroom configuration has an established set of SAS Enterprise Miner projects. The project for this course is DMTXT51. The project can be accessed by selecting **Recent Projects**. To create a new project, select **New Project**.
4. Select **Recent Projects**. The projects that you see might differ from those shown below. Select the **DMTXT51** project, and then click **Open**. You might be instructed to pick a different project.



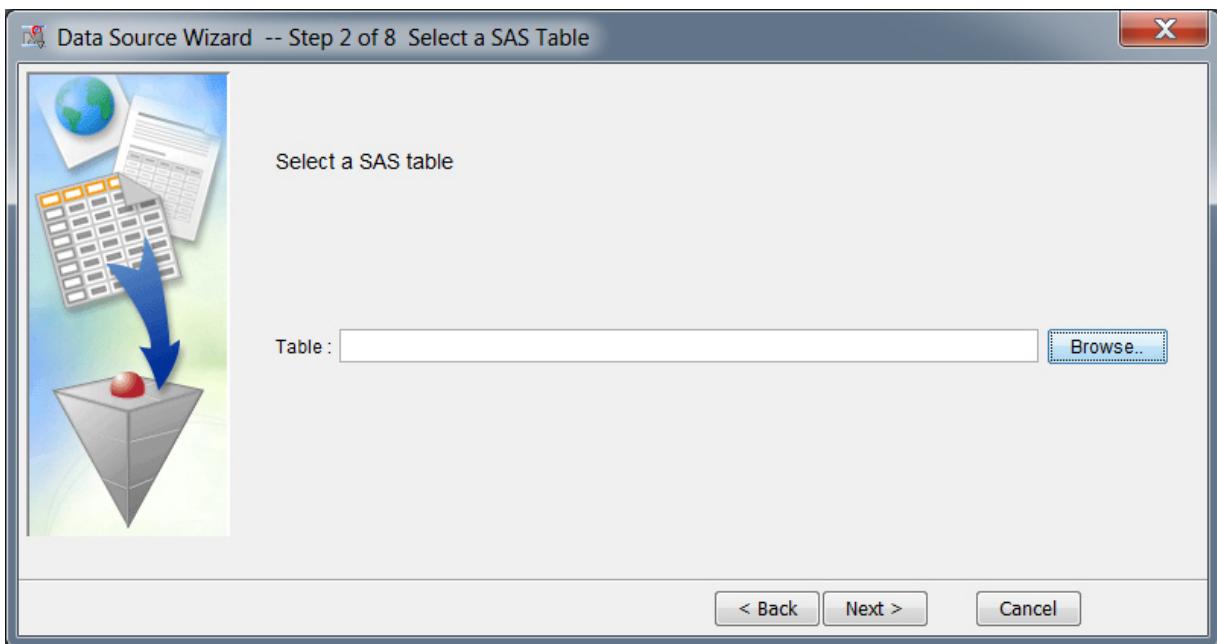
The DMTXT51 project opens.



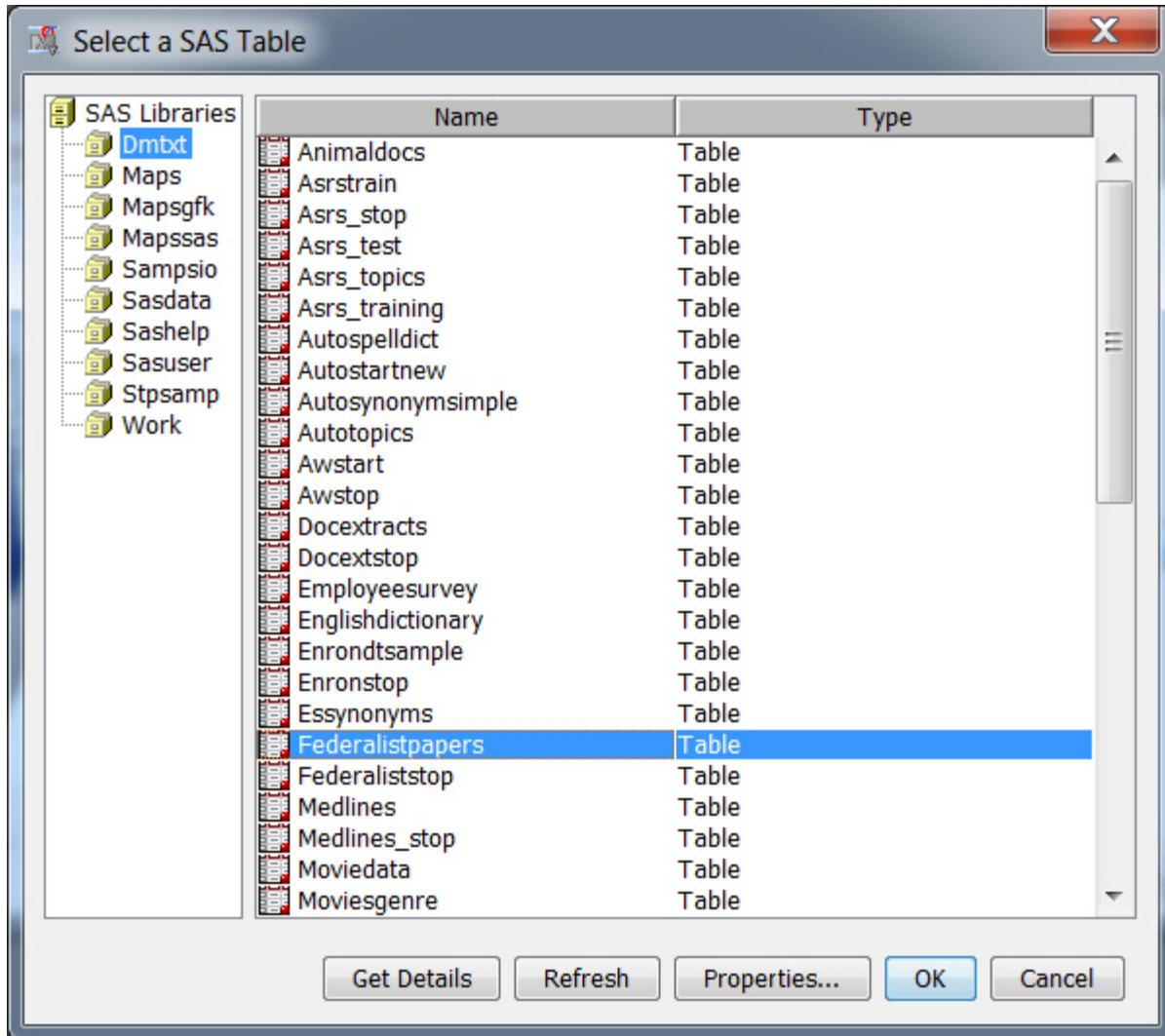
5. As mentioned above, most data sources have already been processed for use by SAS Enterprise Miner. For illustration, the steps for creating a data source are included, but you can skip them if the data source already exists. Select **File** \Rightarrow **New** \Rightarrow **Data Source**. The Data Source Wizard appears.



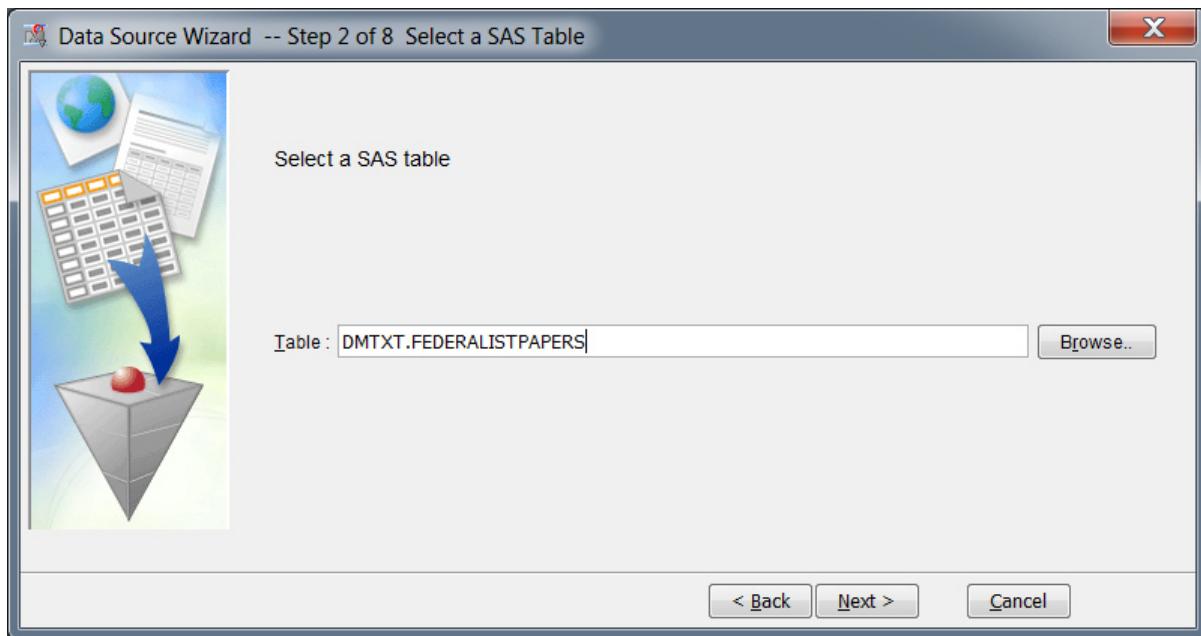
6. Click **Next**.



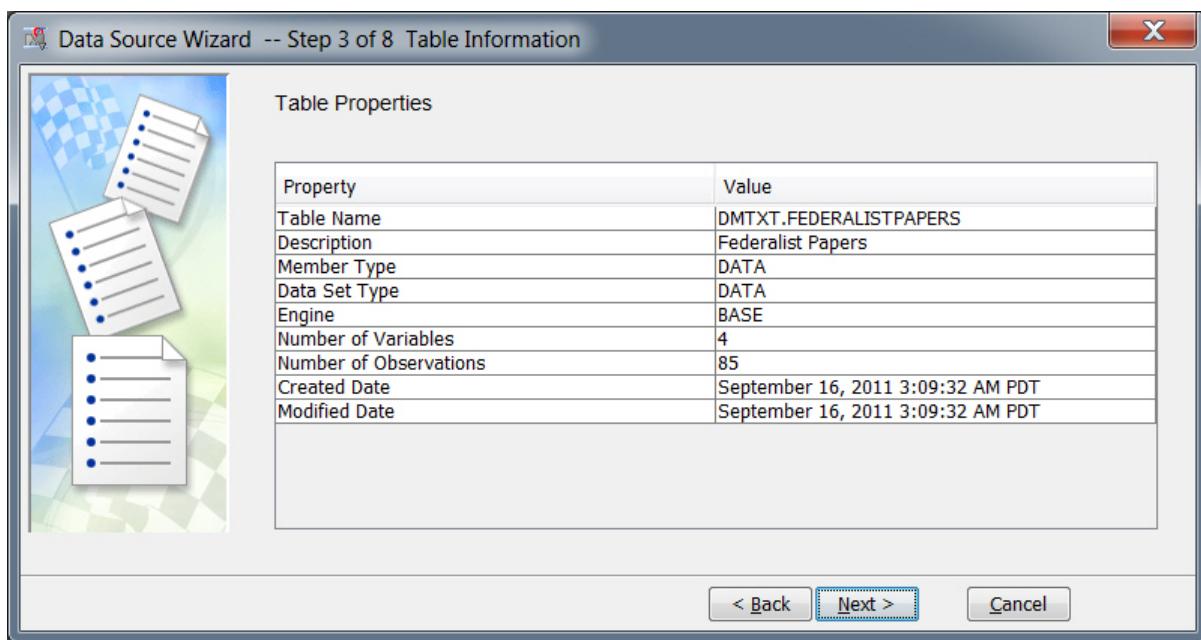
7. Click **Browse**.



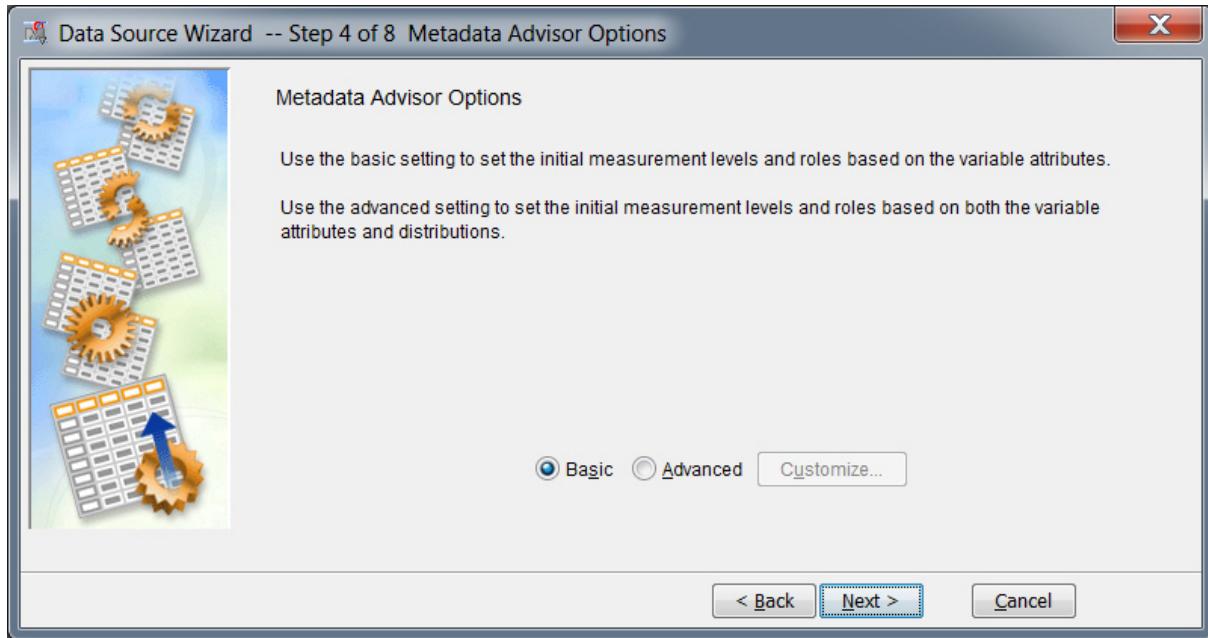
8. Select the table **Federalistpapers** from the **Dmtxt** library. Click **OK**.



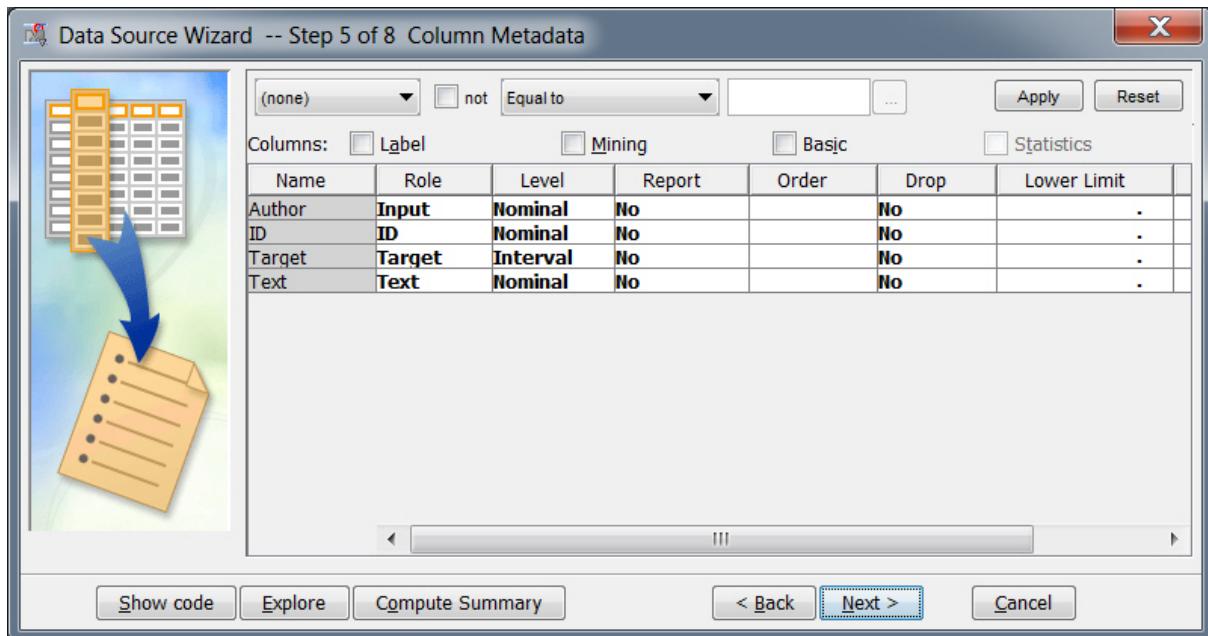
9. Click **Next**.



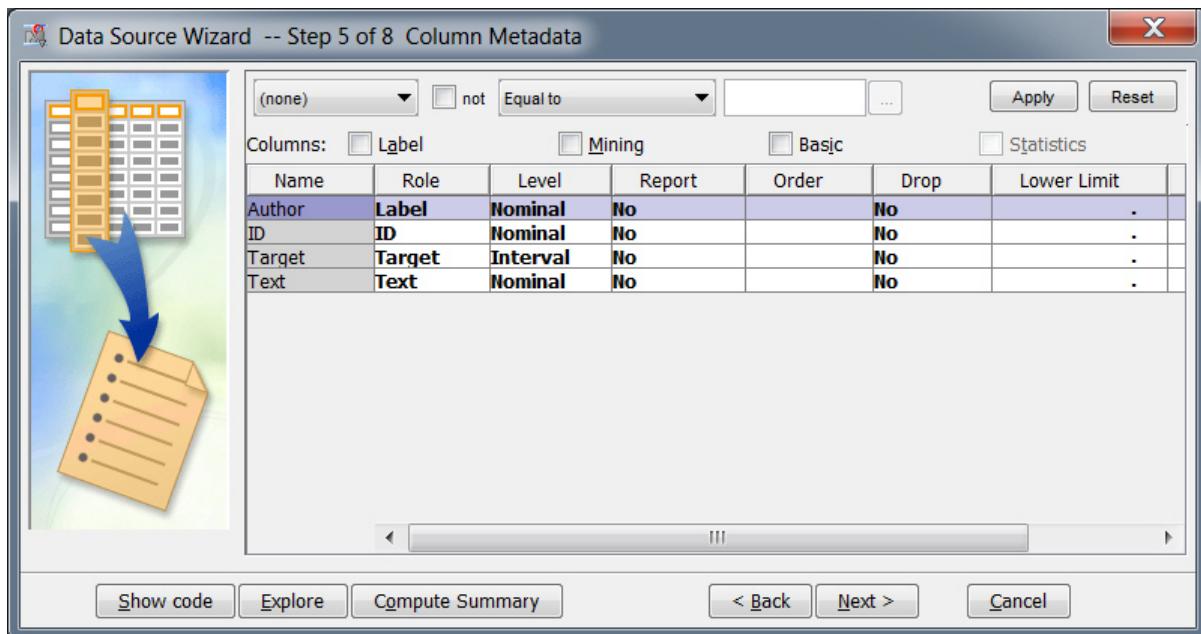
10. Click **Next**.



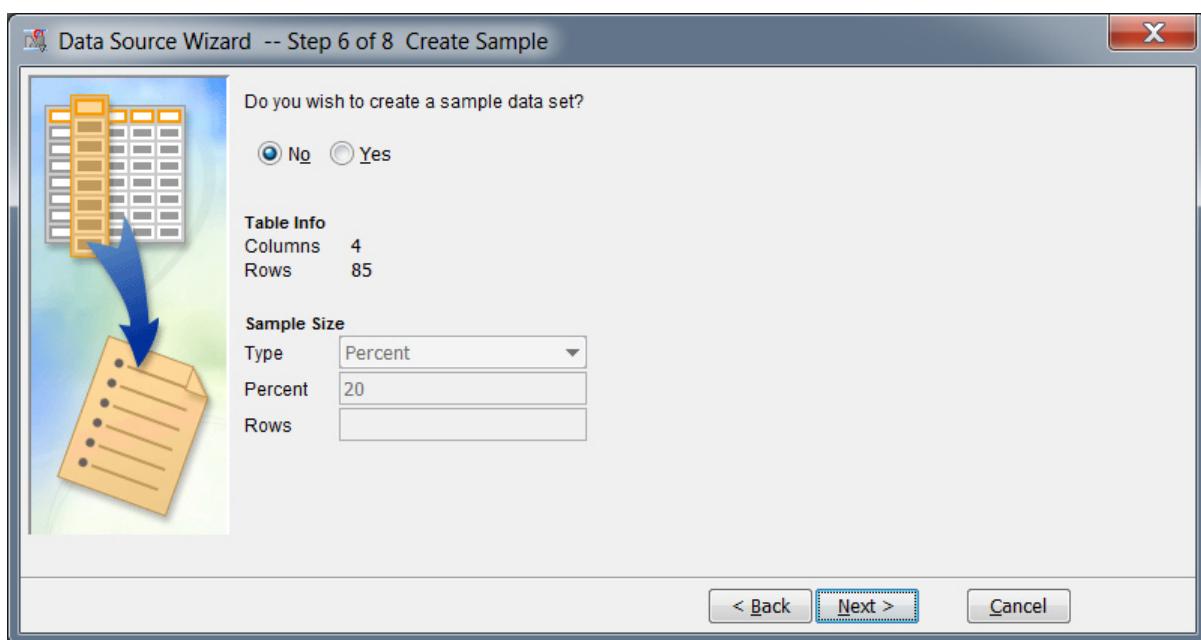
11. Click **Next**.



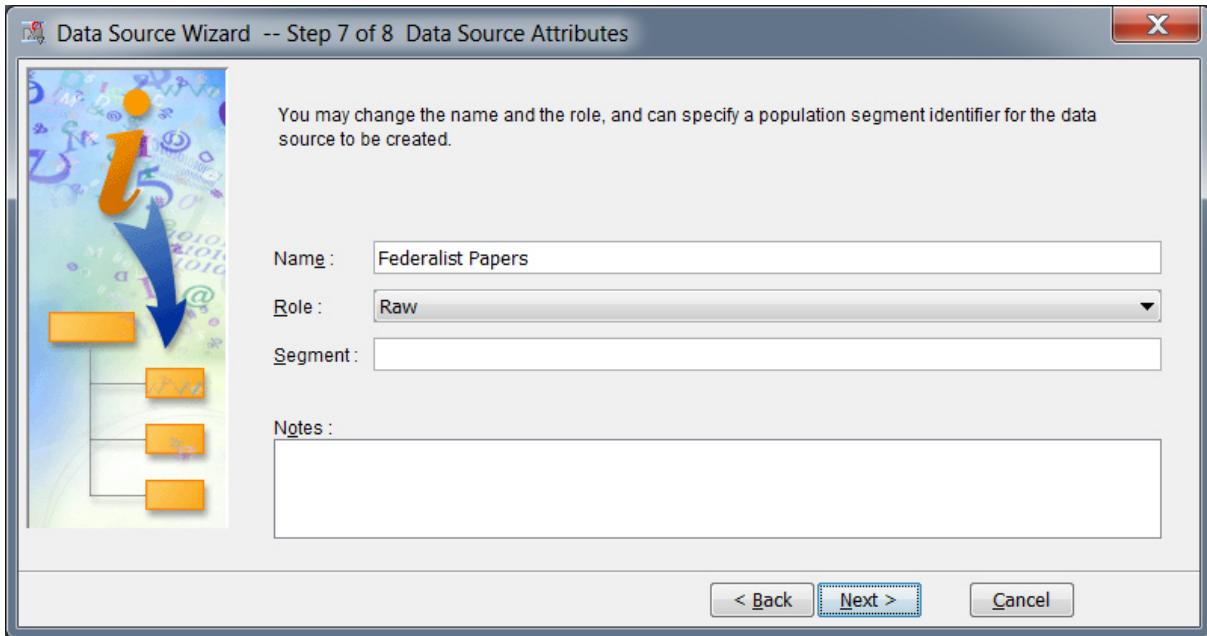
12. You need to change the role of the variable **Author** from **Input** to **Label**. Select the row that **Author** is in and then right-click the word **Input**. Select **Label**.



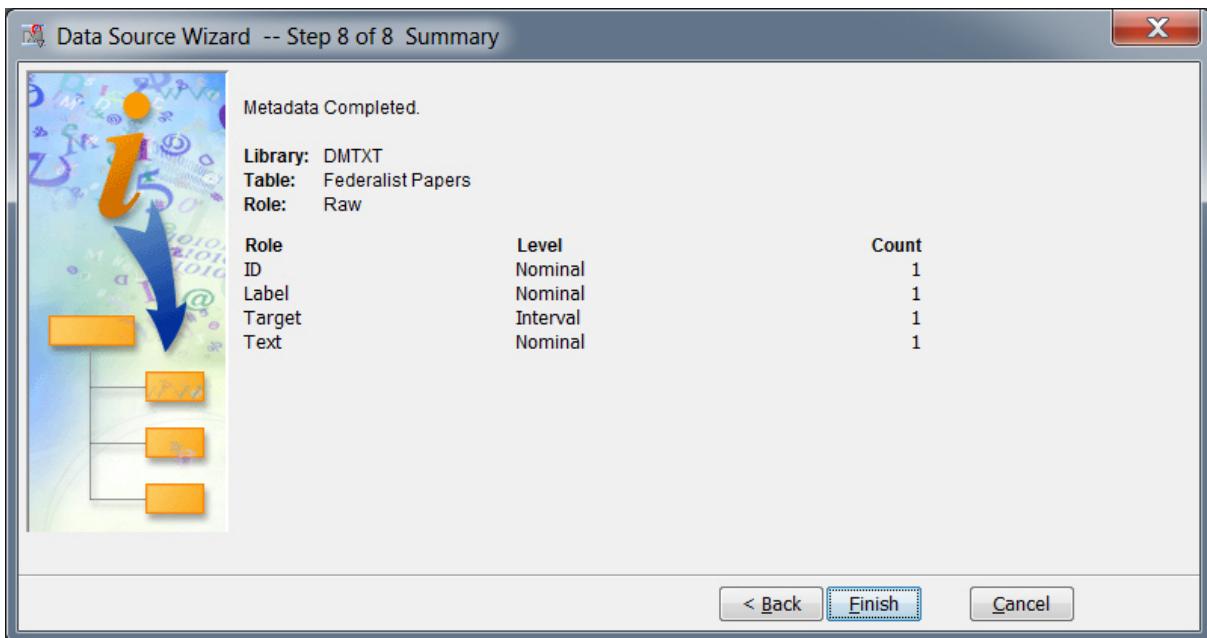
13. Click **Next**.



14. Click **Next**.

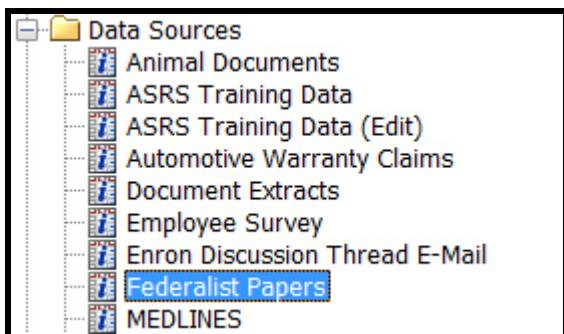


15. Click **Next**.

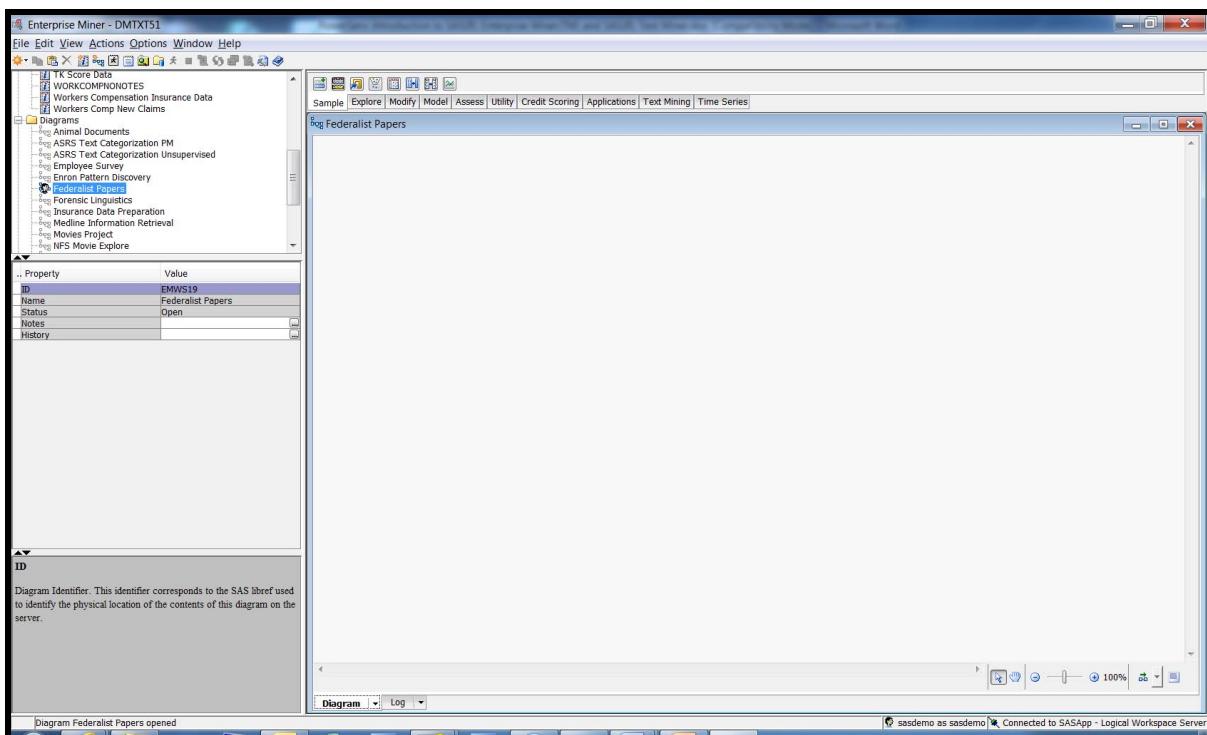


16. Click **Finish**.

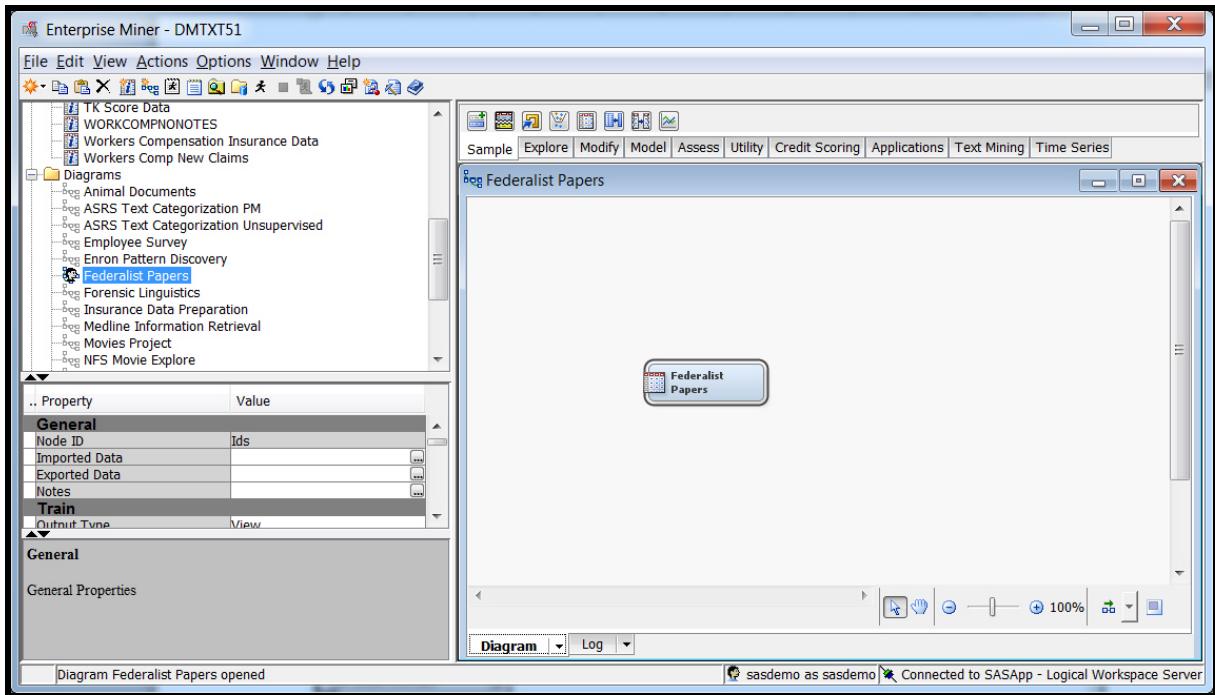
The Federalist Papers data source is now available for your diagrams.



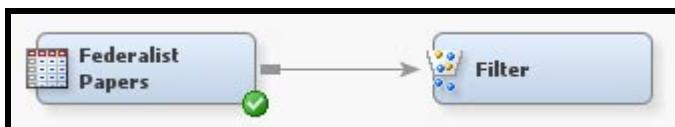
17. Select **File** \Rightarrow **New** \Rightarrow **Diagram**. Name the diagram **Federalist Papers**. Click **OK**.



18. Click the **Federalist Papers** data source, hold down the mouse button, and drag the data source into the diagram. Release the mouse button when the data source is in the diagram.



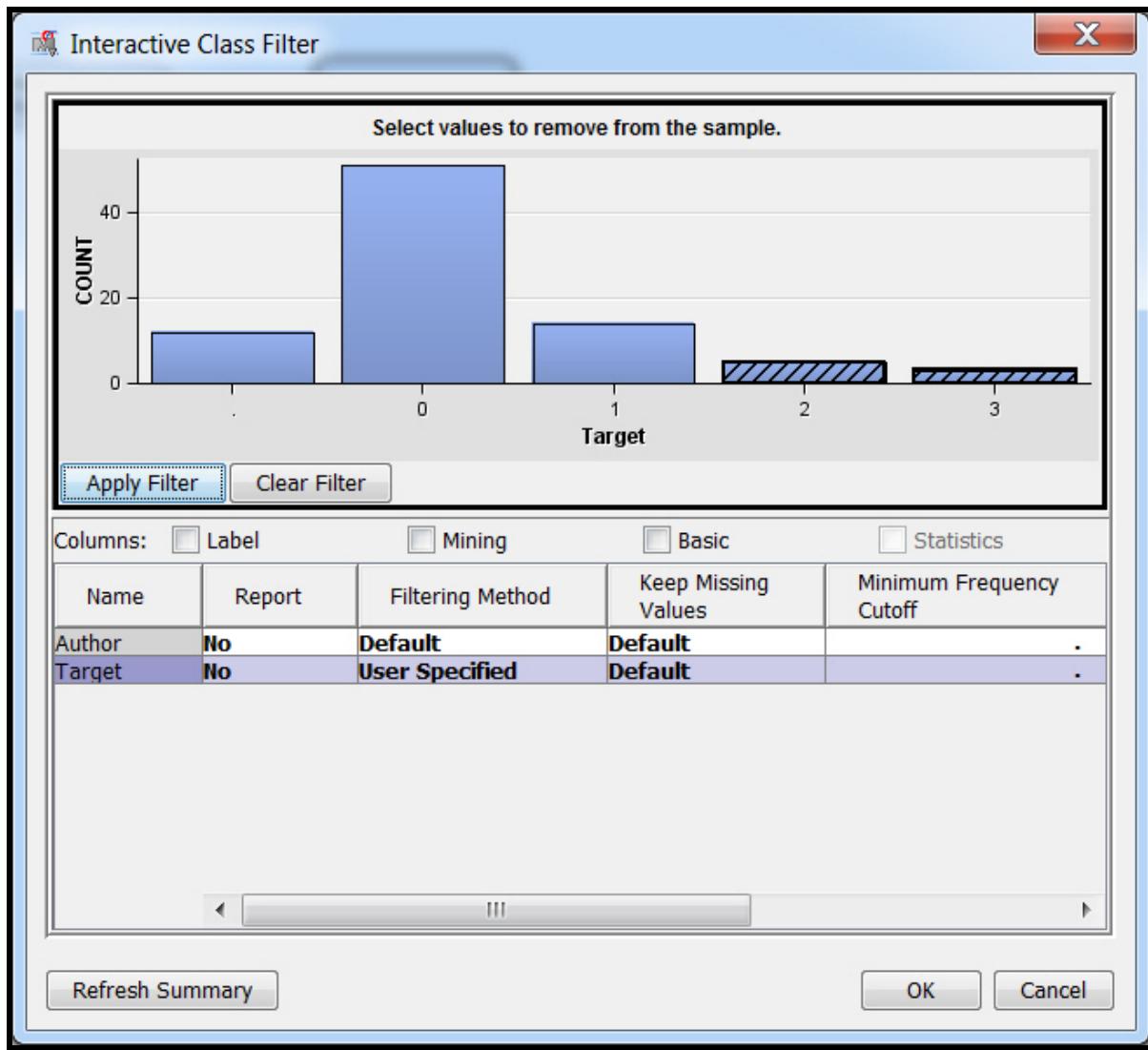
19. Right-click the data source node and select **Run**. When the Run Completed window appears, select **OK**. The green circle with the white check mark indicates that the node ran successfully at least one time. It does not mean that the node is necessarily up-to-date.
20. Click the **Sample** tab, and drag a Filter node into the diagram. Attach the input data source node to the **Filter** node.



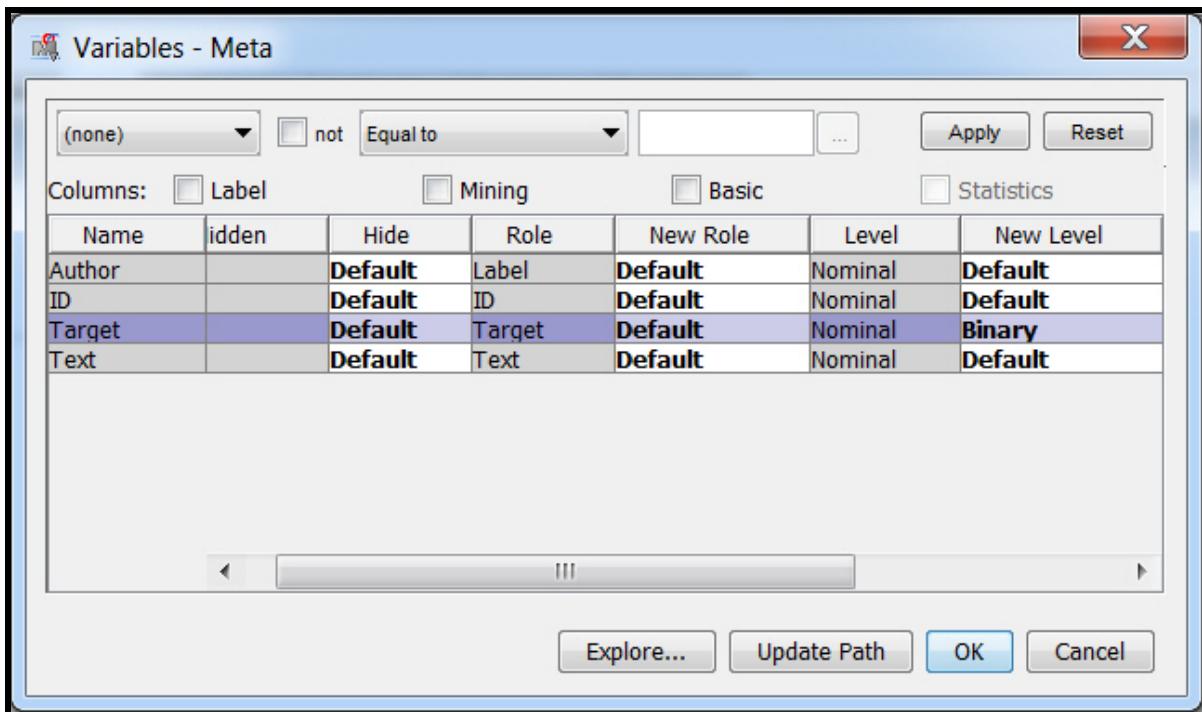
The Filter node Properties panel follows:

.. Property	Value
General	
Node ID	Filter
Imported Data	[...]
Exported Data	[...]
Notes	[...]
Train	
Export Table	Filtered
Tables to Filter	Training Data
Distribution Data Sets	Yes
Class Variables	
Class Variables	[...]
Default Filtering Method	Rare Values (Percentage)
Keep Missing Values	Yes
Normalized Values	Yes
Minimum Frequency Cutoff	1
Minimum Cutoff for Percentage	0.01
Maximum Number of Levels Cut	25
Interval Variables	
Interval Variables	[...]
Default Filtering Method	Standard Deviations from the Me
Keep Missing Values	Yes
Tuning Parameters	[...]
Score	
Create score code	Yes
Update Measurement Level	No

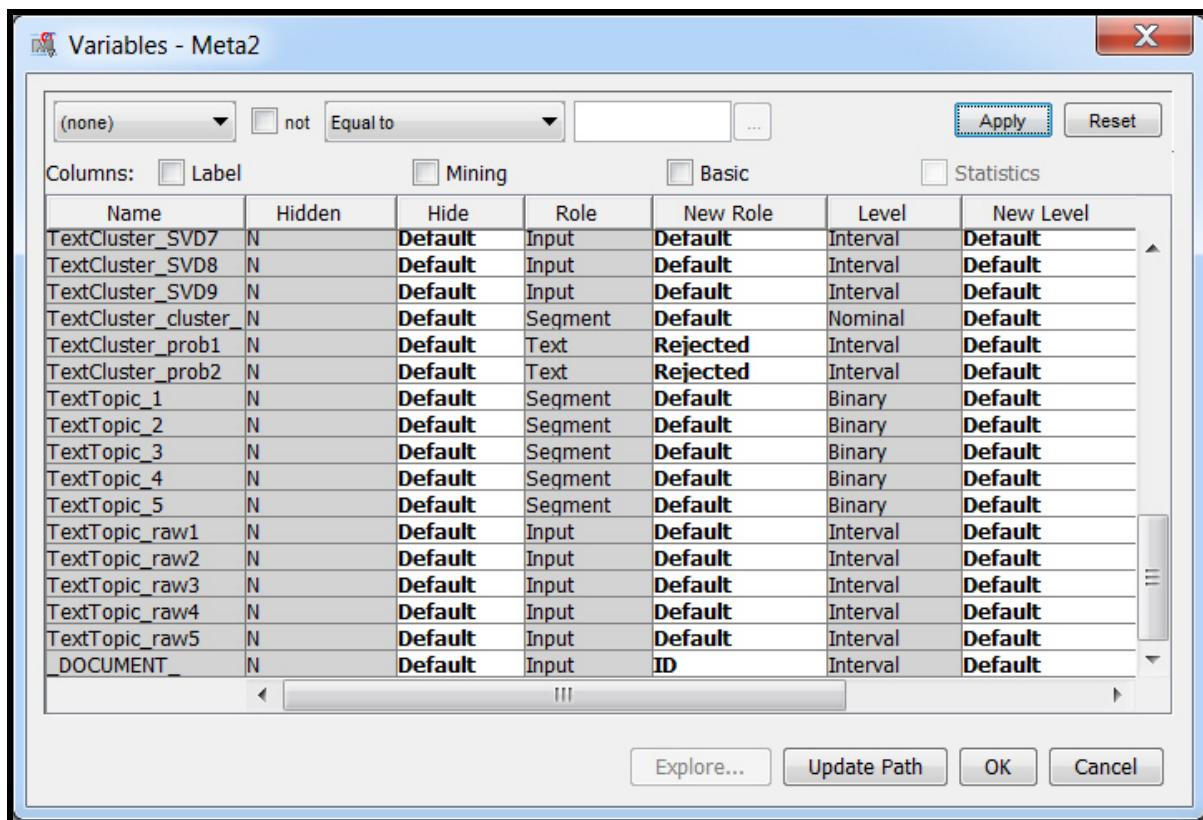
21. Change the Default Filtering Method property to **None** for both class and interval variables.
22. Select the **Class Variables** property. Change the Filtering Method property for the variable **Target** to **User Specified**. Click the **Generate Summary** button. Hold down the control key and click the bars corresponding to values 2 and 3. Then select **Apply Filter**.



23. Click **OK**. Run the filter node. The results window reveals that eight documents have been excluded.
24. On the Utility tab, select the **Metadata** node and drag it into the diagram. Attach it to the Filter node. Select the property **Train**. Change the level of the **Target** variable from **Nominal** to **Binary**.

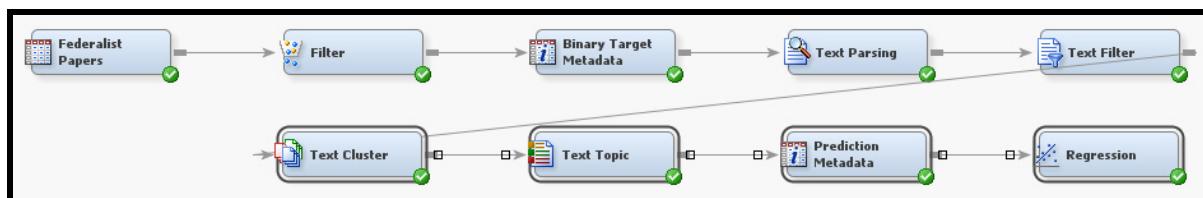


25. Click **OK**. Run the Metadata node.
26. Attach a Text Parsing node to the Metadata node. Change the Synonyms property to **No Data Set to be Specified**. Change the Stop List property to **DMTXT.FederalistStop**. Run the Text Parsing node.
27. Attach a Text Filter node to the Text Parsing node. Change the Term Weight property to **Inverse Document Frequency**. Change the Minimum Number of Documents property to **2**. Run the Text Filter node.
28. Attach a Text Cluster node to the Text Filter node. Set the Exact or Maximum Number property to **Exact**, and set the Number of Clusters property to **2**. Run the Text Cluster node.
29. Attach a Text Topic node to the Text Cluster node. Change the Number of Multi-term Topics property to **5**. Run the Text Topic node.
30. On the Utility tab, drag a Metadata node into the diagram and attach it to the Text Topic node. Change the role of **TextCluster_prob1** and **TextCluster_prob2** to **Rejected**, and change the role of **DOCUMENT** to **ID**. This action fixes an erroneous choice of variable roles that will be corrected in a later release of the software. The erroneous roles of Text can cause a server error for certain downstream actions, and the erroneous role of Input can degrade predictor performance.



31. Run the Metadata node.
32. On the Model tab, drag a Regression node into the diagram and attach it to the Metadata node. Use the default settings, and run the Regression node. Because the target variable is binary, the Regression node will fit a logistic regression model to the data to predict the author of each essay.

The process flow appears below.



33. Select the property **Exported Data** from the Properties panel of the Regression node. Click the **TRAIN** data and click the **Explore** button. Click on the plot wizard icon. In the plot wizard, select **Bar** (chart) and then click **Next**. Specify the role **Category** for variable **Target** and the role **Group** for variable **I_Target**. Click **Finish**. The following plot appears:



The regression model is 100% accurate in predicting the known authors of the 65 attributed essays, and the model predicts that the author of the unattributed essays is Madison. This result agrees with the conclusion of Mosteller and Wallace.

1.03 Multiple Choice Poll

What was the Filter node used for in the Federalist Papers process flow?

- a. to remove unusual or influential documents
- b. to filter out non-informative terms
- c. to remove documents not authored by Madison or Hamilton
- d. to reduce the size of documents that have more than 32,767 characters

64

From Text to Numbers

The Linear Algebra approach uses different names in the literature to quantify documents:

- Vector Space Model (VSM)
- Latent Semantic Indexing (LSI)
- Latent Semantic Analysis (LSA)

Basic calculation per document:

- Boolean counting (0-1) of terms
- Frequency counting of terms
- Information theoretic counting of terms (logarithm of frequency counts)

66

continued...

The VSM is actually distinct from LSA. The VSM does not include a dimensionality reduction component, but LSA employs fast, efficient algorithms to reduce the dimensionality of the problem. Dimensions are determined by the number of documents and the number of terms in the collection.

From Text to Numbers

Adjusting for document size and corpus size \Rightarrow term weights:

- Entropy weights (Shannon information theory)
- Inverse document frequency weights
- Target-based weights
- Others

To illustrate how the use of weights provides a mechanism for enumerating text, consider the use of Boolean weights in a Boolean search. The term *Boolean* just means that exactly two outcome are allowed, which we code as zero (0) or one (1). The concept of a Boolean operation is important in logic, where the result of a logical operation is either TRUE (1) or FALSE (0).

From Text to Numbers: Boolean Search

Document	Term_1	Term_2	Term_3	Term_4	Term_5	Term_6
Doc_01	0	1	0	0	1	0
Doc_02	0	1	0	0	0	1
Doc_03	0	0	1	0	0	0
Doc_04	0	0	1	1	1	1
Doc_05	1	0	0	0	1	0
Doc_06	0	0	0	1	1	1
Doc_07	1	0	1	0	0	0
Doc_08	0	1	0	1	0	1
Doc_09	0	1	1	0	0	1
Doc_10	0	0	1	0	0	1
Doc_11	1	0	0	0	0	1
Doc_12	1	1	0	0	1	0

+

Document	Term_1	Term_2	Term_3	Term_4	Term_5	Term_6
Que_01	1	1	1	0	0	0
Que_02	0	0	0	1	1	1
Que_03	1	0	2	0	3	0

Document/Term Matrix

Query Matrix

continued...

If a term appears in a document, it receives a weight of 1, no matter how often it appears. Otherwise, the term receives a weight of 0. A query can be specified that assigns a weight to each term to represent how important it is in the query. The query could be Boolean as well, with a value of 1 for terms that are sought, and a value of 0 otherwise.

To evaluate how well a document satisfies the query, the weight for each term in the document is multiplied by the corresponding weight for the term in the query. The products are added to give a total score for the document. This score represents how well a document satisfies the query.

From Text to Numbers: Boolean Search

Document	Term_1	Term_2	Term_3	Term_4	Term_5	Term_6	Q1	Q2	Q3
Doc_01	0	1	0	0	1	0	1	1	3
Doc_02	0	1	0	0	0	1	1	1	0
Doc_03	0	0	1	0	0	0	1	0	2
Doc_04	0	0	1	1	1	1	1	3	5
Doc_05	1	0	0	0	1	0	1	1	4
Doc_06	0	0	0	1	1	1	0	3	3
Doc_07	1	0	1	0	0	0	2	0	3
Doc_08	0	1	0	1	0	1	1	2	0
Doc_09	0	1	1	0	0	1	2	1	2
Doc_10	0	0	1	0	0	1	1	1	2
Doc_11	1	0	0	0	0	1	1	1	1
Doc_12	1	1	0	0	1	0	2	1	4

The largest value of the query occurs for the document (or documents) that most closely matches the query.

This illustrates a Boolean search from information retrieval.

69

The operation of scoring a document for a given query is equivalent to taking a dot product between two vectors. If you represent the terms by (t_1, t_2, \dots, t_n) and the query vector by (q_1, q_2, \dots, q_n) , then the dot product is as follows:

$$\sum_{i=1}^n t_i q_i = t_1 q_1 + t_2 q_2 + \cdots + t_n q_n$$

Thus, the third query applied to the first document yields the following score:

$$(0 \times 1) + (1 \times 0) + (0 \times 2) + (0 \times 0) + (1 \times 3) + (0 \times 0) = 3$$

In a pure Boolean search, if the sum of the query values is k , then the search will only retrieve documents that produce an exact score of k . In the Boolean search described above, first a cutoff value is established. Then only documents that produce a score above the cutoff are returned by the query. Search engines pioneered the use of a modified Boolean search with a score cutoff to retrieve Web pages for a given query. Early searches were just keyword searches, which are equivalent to pure Boolean queries. A document either has one or more occurrence of each word from the query in the document (TRUE) or it does not (FALSE). The advantage of the more sophisticated methodology described above is that documents can be sorted by descending score and returned in this order to the user. Thus, the first document returned is more likely to be relevant than the second document, and so on, assuming no ties.

The Text Filter node uses a variant of this methodology. However, a more exciting aspect arises when you consider the data mining side of text mining. Can an algorithm be created to sift through the document collection and derive query weights that best represent the concepts or topics in the collection? The answer is, “Yes!” LSA provides the functionality to **derive** queries, rather than just respond to queries. The first derived query will score each document with respect to the most dominant concept in the collection. The second derived query will score each document with respect to the second most dominant concept in the collection, and so on.

Text Mining

- Algorithms process documents (parsing/filtering).
- A derived vector is associated with each document.
- The vector is typically too large and has too many zeros to work with directly, so transformation methods and dimensionality reduction techniques are applied to produce a more useful final vector representation for each document.
- Converting a document to a well-defined, structured vector permits application of any valid analytic technique to facilitate problem solving.

70

continued...

You can think of the vector associated with each document as the score produced by each derived query. For example, if the dimensionality is set to 50, then 50 sets of query weights will be derived, and each document will produce 50 scores, 1 score for each derived query. The maximum dimensionality is the number of terms in the dictionary or vocabulary (start list) used for the analysis. The number of terms is usually in the thousands or hundreds of thousands. LSA provides a methodology to reduce this maximum dimensionality down to a reasonable dimensionality that will still permit successful mining of the document collection.

Text Mining

“Due to repetitive motion...”

distance



“Bilateral carpal tunnel...”

71

continued...

Continuing the example using a dimensionality of 50, two vectors of length 50 corresponding to two documents can be thought of as coordinates in a 50 dimensional Euclidean space. Any mathematical operation permissible in Euclidean space can be applied to the two vectors. Using three dimensions is easier to visualize, but any number of dimensions is acceptable subject to hardware and software limitations.

Text Mining		SVD Dimensions		
d o c		_SVD_1	_SVD_2	_SVD_3
“Due to repetitive motion...”	<i>i</i>	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$
“Bilateral carpal tunnel...”	<i>j</i>	$x_{j,1}$	$x_{j,2}$	$x_{j,3}$

Euclidean Distance: $D_{i,j} = \sqrt{\sum_{k=1}^3 (x_{i,k} - x_{j,k})^2}$

Cosine Distance: $\cos(i, j) = \frac{\sum_{k=1}^3 (x_{i,k}x_{j,k})}{\sqrt{\sum_{k=1}^3 x_{i,k}^2 \sum_{k=1}^3 x_{j,k}^2}}$

72

You might have a better understanding of Euclidean distance because it is taught in high school algebra. However, measures like cosine distance have turned out to be more useful in text mining practice. The idea is that if the angle between two concept vectors is small, the vectors probably represent the same concept, whereas if the angle is large, then two different concepts are probably being represented. If the angle between two document vectors is small, then the documents probably contain very similar information.



Exercises

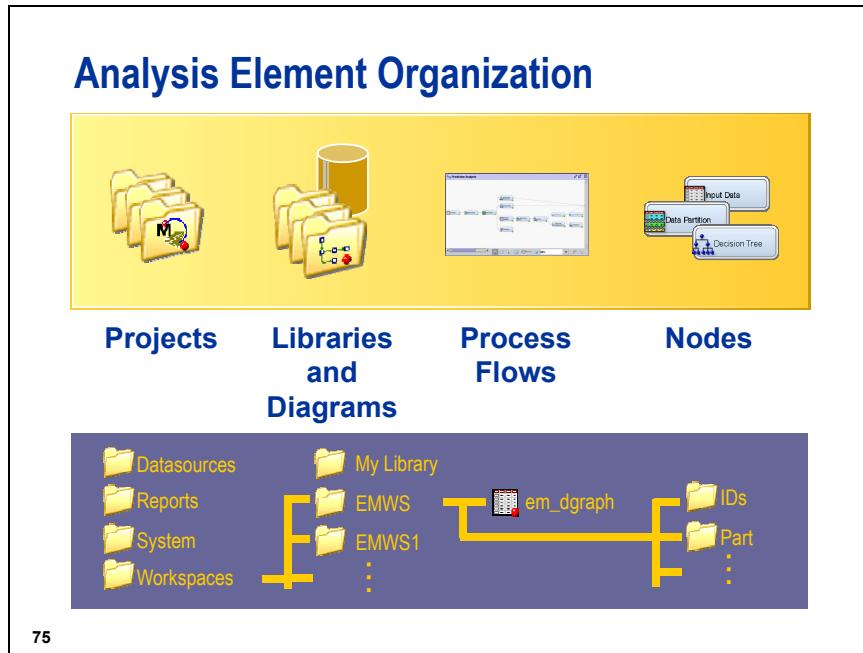
1. Using the Text Miner Nodes to Predict the Author of the Essays

Reproduce the stylometry demonstration with the Federalist Papers.

1.2 Working with Data Sources

Objectives

- Describe SAS Enterprise Miner metadata and detail the types of roles and measurement levels that are supported.
- Explain how to create data sources that can be used by SAS Enterprise Miner projects.
- Provide examples of data sources that are relevant for text mining.

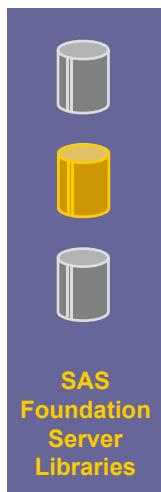


SAS Enterprise Miner organizes projects by placing components of the project in separate folders or directories. The Datasources folder contains metadata for each data source. The Workspaces folder holds all of the details about each diagram, including property settings of nodes used in each process flow.

SAS Enterprise Miner can import data from many sources, including common PC file formats such as Microsoft Excel and common commercial relational databases (for example, Sybase, Teradata, and Oracle), as well as from SAS data sets. The functionality of SAS Enterprise Miner comes from the assignment of roles and levels to variables in a data set. Initially assigning metadata roles makes the building of process flows much easier. Data properties do not have to be repeated or copied and pasted for each new task.

One of the first tasks in any project is to identify one or more relevant data sources. While you can merge tables inside SAS Enterprise Miner, a best practice is to use the query optimization features of the native database to build the analysis table, and then import this table into SAS Enterprise Miner.

Defining a Data Source



- Select a table.
- Define variable roles.
- Define measurement levels.
- Define table role.

76

Variable roles specify how a variable can be used.

Variable Roles

- Assessment
- Censor
- Classification
- Cost
- Cross ID
- Decision
- Frequency
- ID
- Input
- Label
- Prediction
- Referrer
- Rejected
- Residual
- Segment
- Sequence
- Target
- Text
- Text Location
- Time ID
- Web Address

77

These are the most common roles in text mining:

- Text
- Text Location
- ID
- Input
- Target
- Rejected

The document is stored as a variable with a role of Text. If the document is larger than 32,767 characters, then the full pathname of the document as recognized by the server can be supplied as a variable with a role of Text Location. SAS software limits character variables to 32,767 characters. However, SAS Text Miner has no limit on the size of a document. If the document does not fit completely within a character variable, then SAS Text Miner reads the full document from the specified path. (More details about text data are forthcoming.)

 SAS allows $32,767=2^{15}-1$ characters in a character variable. A document is stored as a SAS character variable in a SAS data set. The SAS Text Miner documentation rounds this down to 32,000, but 32,767 is the correct figure.

Additional variables in the data set usually have roles of ID, Input, Target, or Rejected. An *ID* variable identifies the document uniquely. An *input* variable can be used for segmentation or predictive modeling. Only input variables are used to derive segments or clusters. SAS Text Miner converts each document into a collection of inputs. For predictive modeling, the goal is to predict the value of a *target* variable. Only input variables are used to predict the target.

Any other variable in the data that has no purpose for the analysis has a role of *Rejected*.

Measurement Levels

- Categorical (Class, Qualitative)
 - **Unary**
 - **Binary**
 - **Nominal**
 - **Ordinal**
- Numeric (Quantitative)
 - **Interval**
 - Ratio*

* All methods that accommodate an interval measurement scale in SAS Enterprise Miner also support a ratio scale.

78

Elementary statistics textbooks for social science majors usually describe four measurement levels:

- nominal
- ordinal
- interval
- ratio

Other statistics textbooks might only speak of categorical and numeric data.

A variable with a *nominal* measurement scale is purely categorical in nature. There is no numeric interpretation, and there is no natural ordering. Examples include eye color, political party affiliation, and country of origin. An *ordinal* variable is a categorical variable that has an inherent ordering. Thus, ordinal variables are also called *ordered categorical variables*. Examples include course letter grade, response on a Likert scale, or items on a top-10 ranking list.

 Nominal data can be ranked by frequency of occurrence, price, personal preference, and so on. So if the ranking is meaningful and exploited by the analysis, then the nominal variable becomes an ordinal variable.

A *unary* scale implies a single constant value. A *binary* scale implies a nominal scale with only two distinct values.

A variable with an *interval* measurement scale has a numeric interpretation so that the difference between two numeric values is meaningful. A variable with a *ratio* measurement scale is valid as an interval scaled variable, but in addition, the ratio of two numeric values is meaningful. Temperature in degrees Celsius is on an interval scale, but not a ratio scale, because 20 degrees divided by 10 degrees being equal to 2 does not mean that 20 degrees is twice as hot as 10 degrees.

Most numeric data is on a ratio scale. Most analytic methods only require that the data be on an interval scale. All of the methods employed by SAS Enterprise Miner that work for numeric data also work for interval or ratio-scaled values. Consequently, the ratio scale is not supported.

Different nodes expect specific table roles. The Score node scores raw, training, validation, test, and score data sets. The Association node acts on transaction data sets.

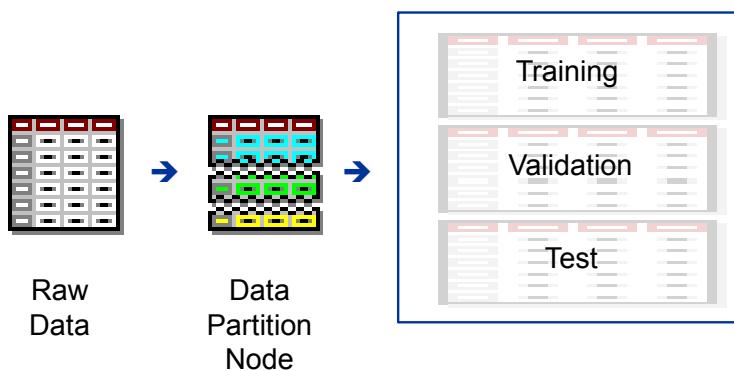
Table Roles

- Raw
- Training
- Validation
- Test
- Score
- Transaction

79

Using the Data Partition node, you can split raw data into training, validation, and test data sets.

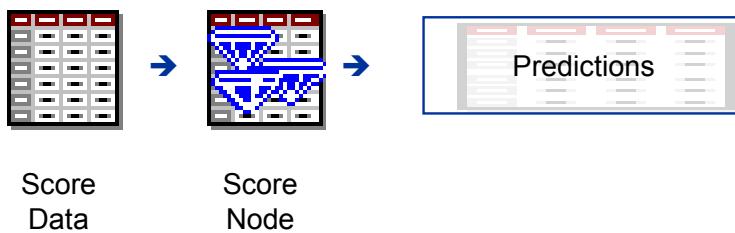
Analysis Data



80

Score data can be scored by the Score node if all of the required data elements are present. The role of the score produced by the Score node is *Prediction* or *Segment*, depending on how the score is produced. Consequently, you need to be familiar with variable roles even if they are not assigned by you.

Scoring (Predicting) New Data



81

SAS Enterprise Miner and SAS Text Miner anticipate the need to create and modify data before an analysis. For text mining, the Text Import node can be deployed in a SAS Enterprise Miner process flow to process a document collection. The Text Import node is discussed in the next chapter.

Working with Text Mining Data Sources

- When documents are stored in separate files, the Text Import node can be used to create an appropriate SAS data set for text mining.
- When documents are stored together (for example, in a Microsoft Excel spreadsheet), then the Import Data Wizard can be used to create a text mining data set.
 - ✍ If the structure does not accommodate storing one document per observation, then more complex SAS programming might be required.

82

The following slide describes how text data is treated by SAS Enterprise Miner.

Working with Text Mining Data Sources

Two supported types of text mining data:

- The data set contains at least one variable with the role **Text**, and documents can be stored completely as a SAS character variable (limited to 32K).
- The data set contains at least one variable with the role **Text Location**, and some documents cannot be stored completely as a SAS character variable.
 - The location must be the full pathname of the document with respect to the Text Miner server.
 - An additional variable with the role **Web Address** can include the path to an unfiltered version of the document to be displayed in an interactive viewer such as the Interactive Filter Viewer.

83

The Text Parsing node accepts train, validate, test, or score data. At least one data source must be train data. Multiple input data source nodes can be attached as predecessors to the Text Parsing node.

The input data source must have at least one variable with a role of Text or Text Location. As stated above, the Text variable can contain an entire document or a truncated piece of an entire document. The Text variable is a character variable, and SAS can only accommodate character variables with lengths up to 32K (32,767 bytes). If a document exceeds 32K in length, then SAS must read the entire document from a location specified in the input data. If no location is specified, then the Text Miner nodes only process the truncated documents.

To process documents that exceed 32K, a variable with the Text Location role must be included in the input data. The text location must be the full pathname of the document folder with respect to the Text Miner server. For example, a document might be visible on your Windows computer at this location:

S:\MyProject\MyDocuments\Doc1.txt

The Text Miner server might recognize the location as follows:

//Sdisk/MyProject/MyDocuments/Doc1.txt

The second form of the document location must be used in the input data.

The Text Filter node can access documents through the Interactive Filter Viewer. The Interactive Filter Viewer only displays the portion of the document stored in the Text variable. If you want to see the entire document in the Interactive Filter Viewer, then you can include a variable with the role Web Address that provides the full pathname of the file that contains unfiltered text. The Topic Viewer in the Text Topic node has the same functionality.

If the input data source contains two or more variables with a role of Text, and the Use status is Yes for these variables, then the Text Parsing node chooses the variable with the largest length. If the lengths are the same, then the variable that appears first in column order is selected. If your data has two or more text variables, you should set the Use status to No for all text variables except the one to be included in the

analysis. You should **not** rely on the default text selection method of the node because the selection default might change between releases of the software (and has changed!).

If you want to include two or more text variables in your text mining project, then you have to connect Text Miner nodes in sequence and change the Use status of the variables as needed.

 This process of *chaining* text mining nodes is illustrated later.

In many cases, you need to preprocess textual data before you can import it into a data source. The Text Import node is designed for this purpose. The Text Import node can be used in file preprocessing to extract text from various document formats or to retrieve text from Web sites by crawling the Web. The node creates a SAS data set that you can use to create a data source to use as input for the Text Parsing node. Depending on which structure (of the two described above) that you use, you have to adjust the roles of the variables accordingly in the Data Source Wizard.

Working with Text Mining Data Sources

Additional data sources:

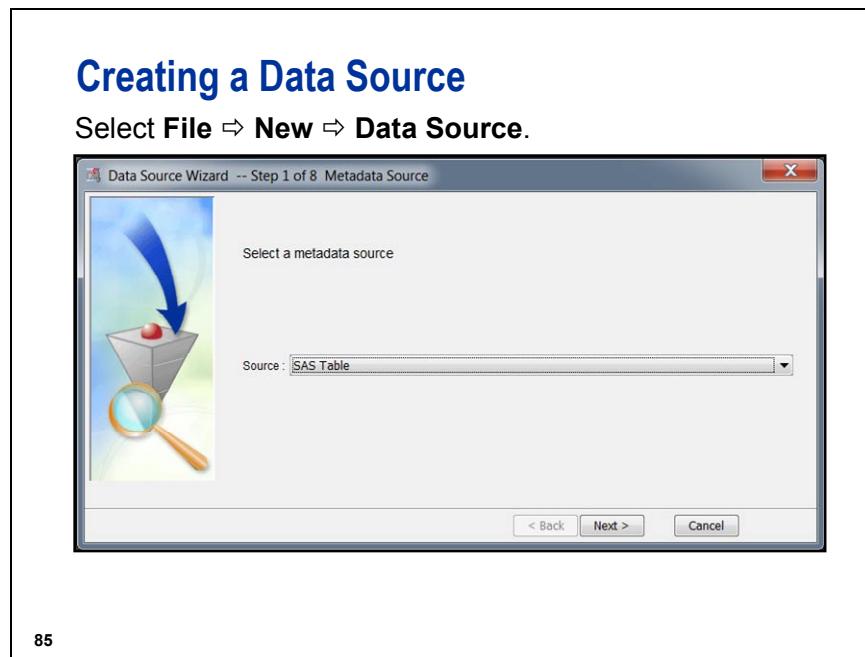
- Dictionaries
 - start lists
 - stop lists
- Synonym tables
- Multi-word term tables
- Topic tables

The software distribution of SAS Enterprise Miner includes the following sample data sets in the Sashelp library:

Data Set	Description	Used In
<i>language_multi</i>	Multi-term lists for various languages	Text Parsing node
<i>languagestop</i>	Stop lists for various languages	Text Parsing node
<i>Engstop</i>	Stop list for the English language	Text Parsing node
<i>Engsynms</i>	Synonym list for the English language	Text Parsing node

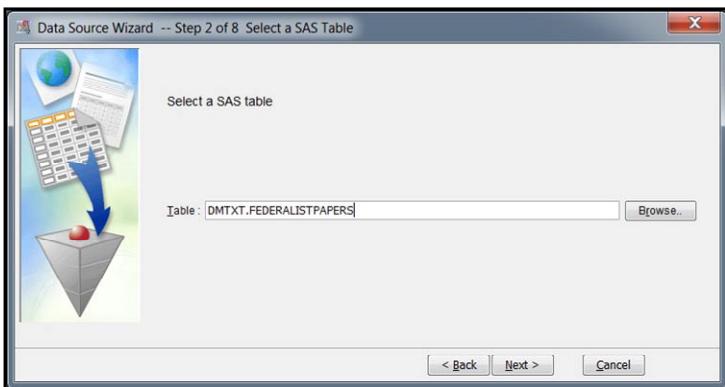
The keyword *language* can be one of Eng, Frnch, Germ, Ital, Port, or Span for the corresponding English, French, German, Italian, Portuguese, or Spanish languages. Other languages are also supported.

The SAS data set **SAMPSIO.TM_abstract_topic** contains a sample topic data set that can be processed by the Text Topic node.



The source table can be a SAS table or any table recognized by your licensed version of SAS, which could include an access engine for a specific commercial database.

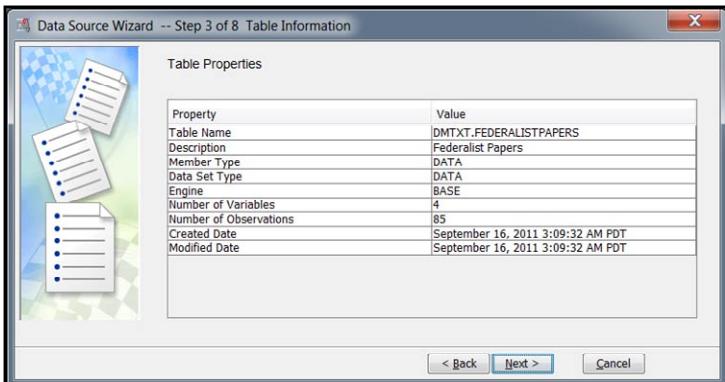
Creating a Data Source



86

The table specified above is named **AAEM61.PVA97NK**. The **AAEM61** library either is identified using the New Library wizard, or is specified in SAS Enterprise Miner start-up code. The **PVA97NK** table refers to **PVA97NK.SAS7BDAT**, which is a file located in the folder defined by the **AAEM61** library reference.

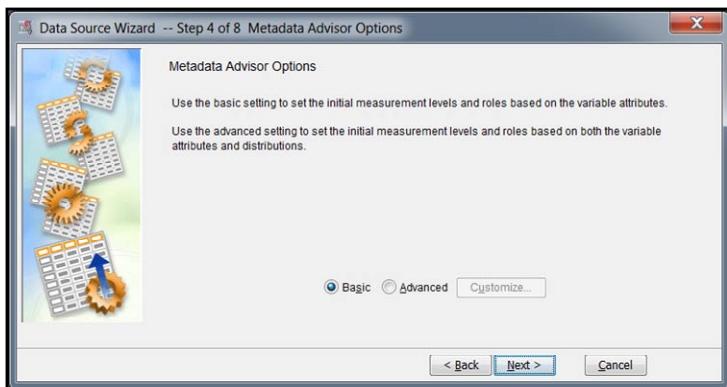
Creating a Data Source



87

The table has 28 columns and 9,686 rows.

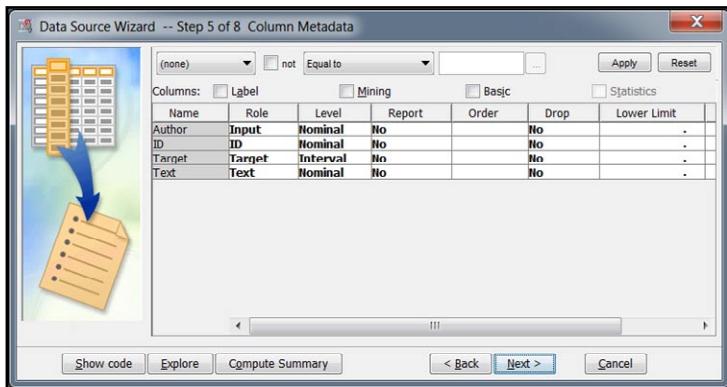
Creating a Data Source



88

The Advanced Advisor is useful for data with many variables. It applies a set of rules to provide preliminary values for the roles and levels of each variable.

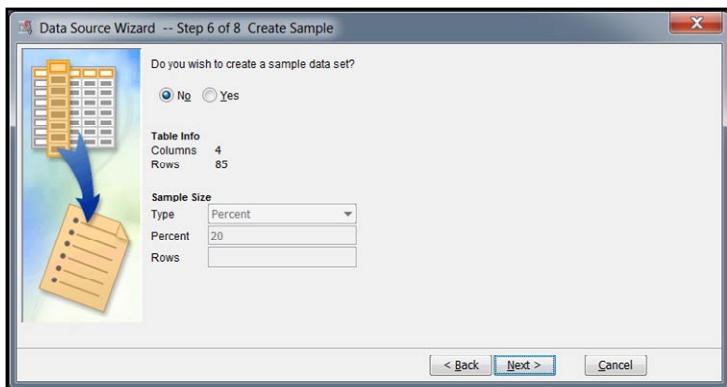
Creating a Data Source



89

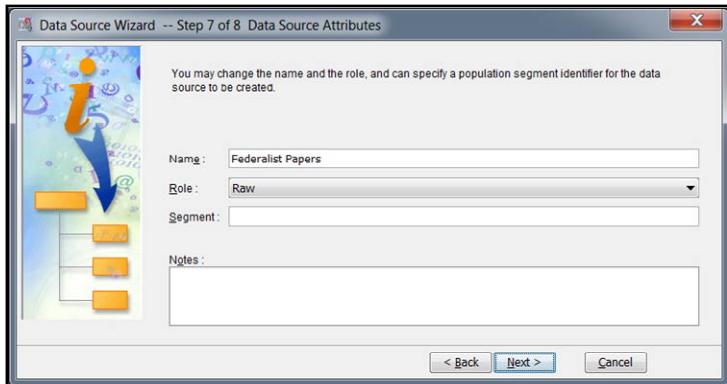
You usually have to manually modify some of the role and level values.

Creating a Data Source



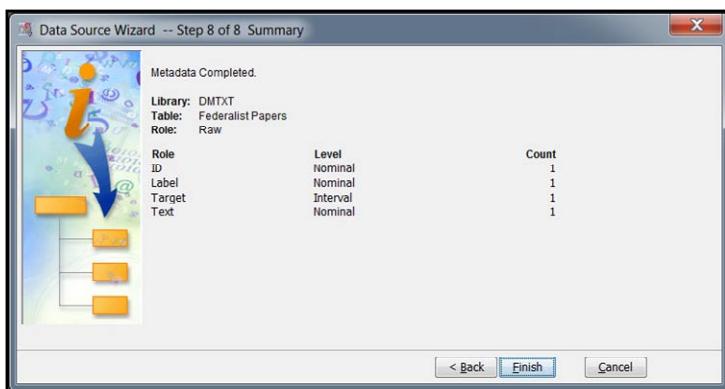
90

Creating a Data Source



91

Creating a Data Source



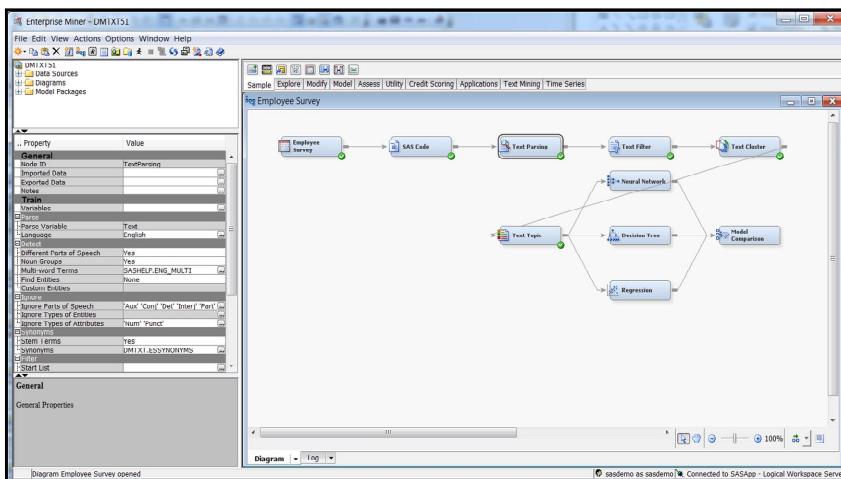
1.3 Using SAS Enterprise Miner and SAS Text Miner

Objectives

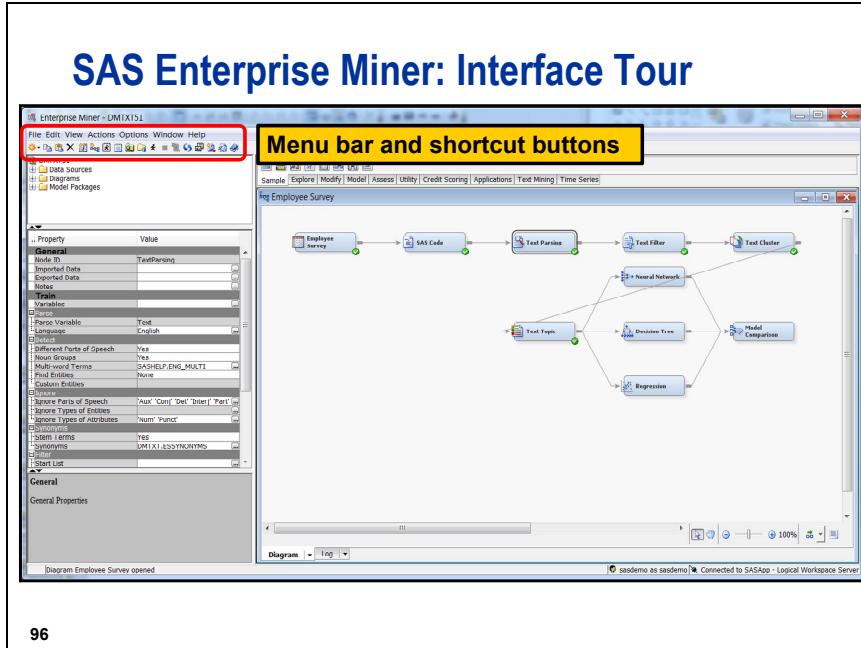
- Tour the SAS Enterprise Miner user interface.
- Describe SEMMA data mining methodology.
- Describe SAS Text Miner.
- Explain the nodes in SAS Text Miner.
- Use the SAS Text Miner nodes to explore a document collection.

94

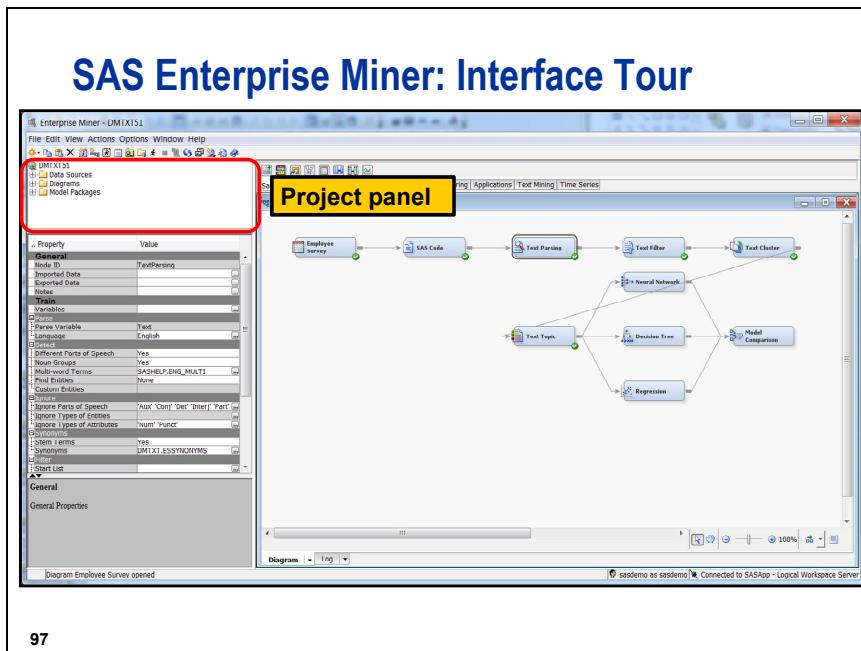
SAS Enterprise Miner



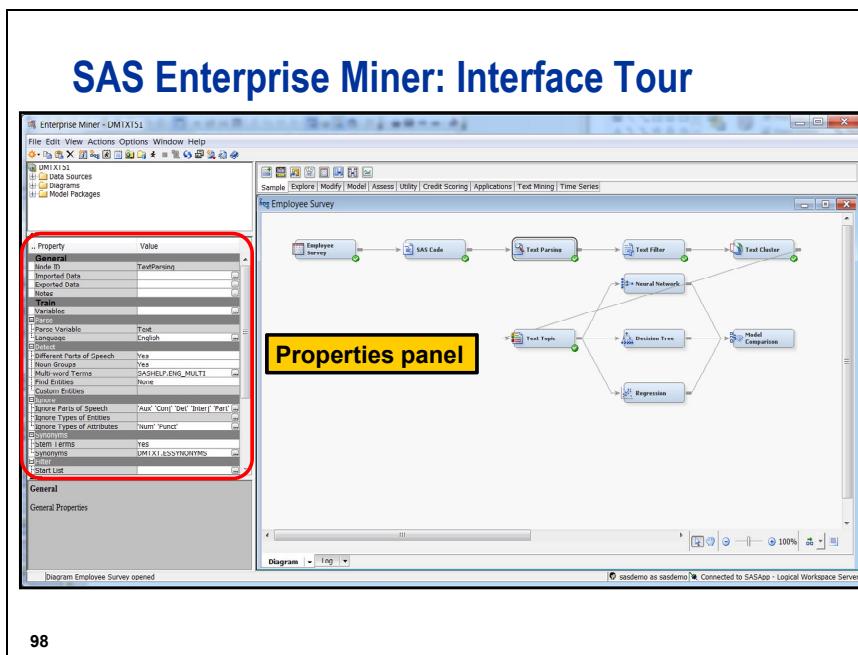
95



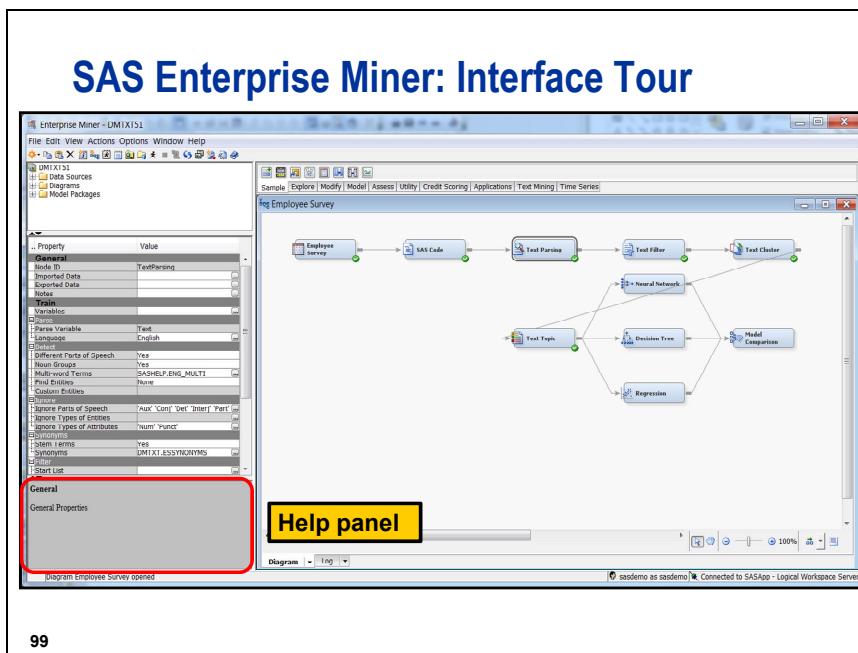
96



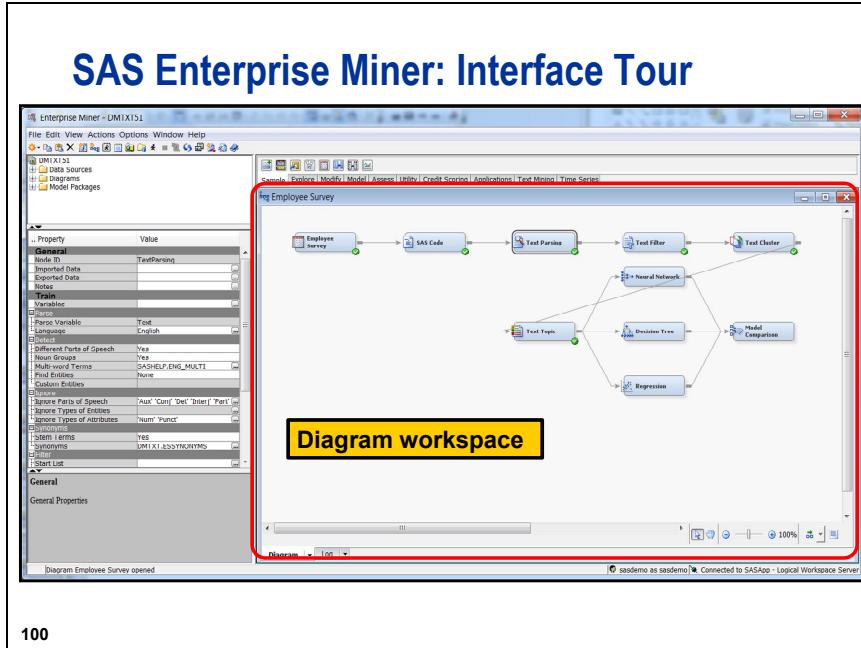
97



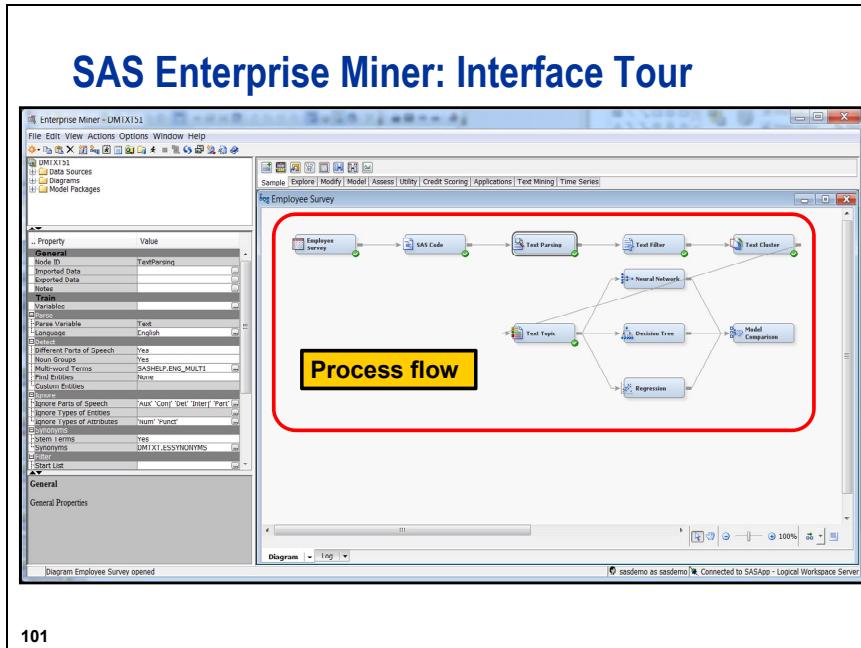
98



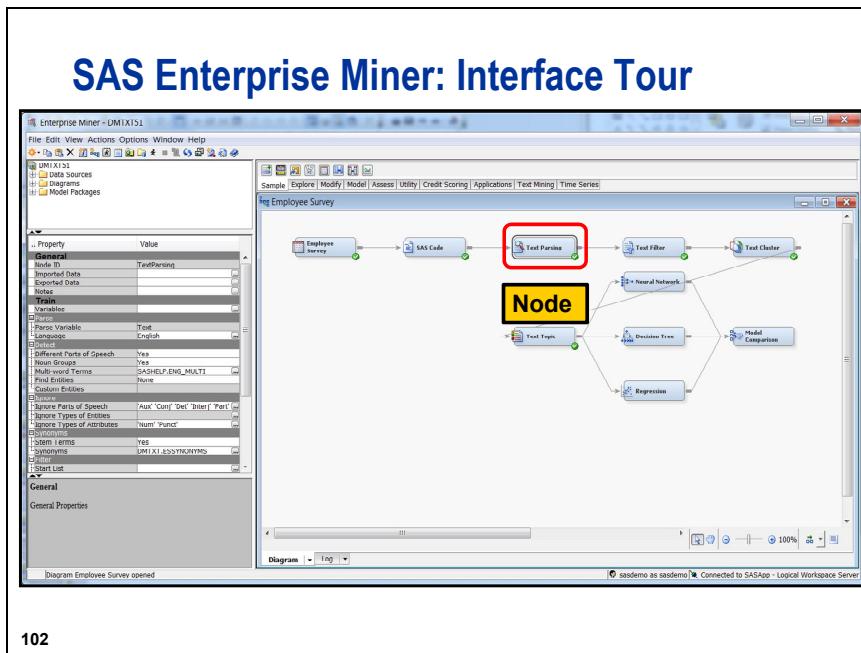
99



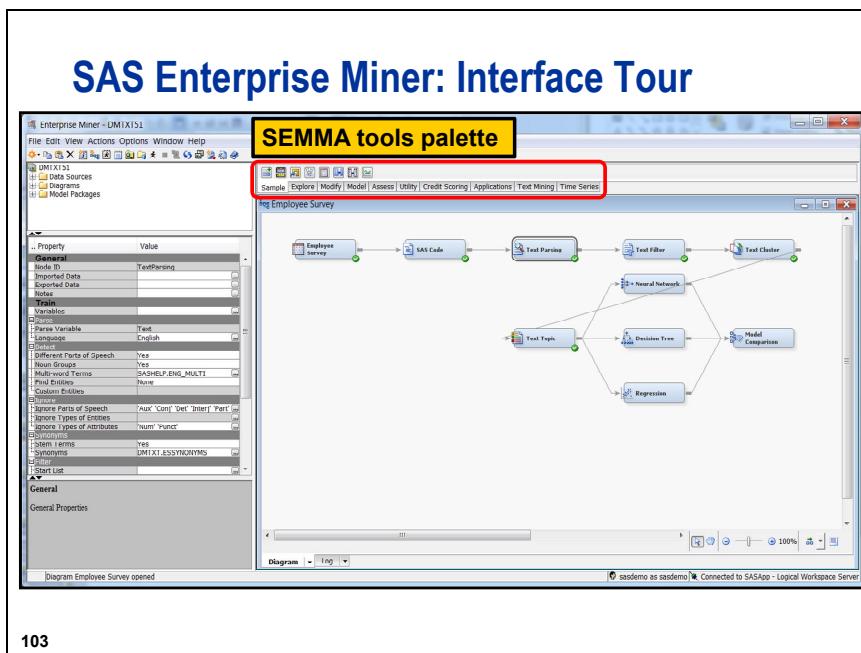
100



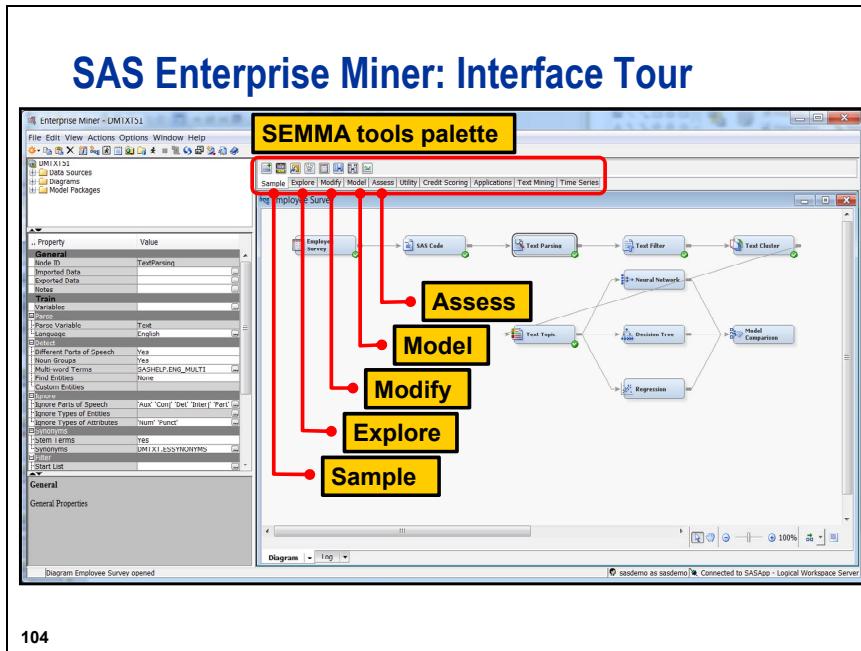
101



102



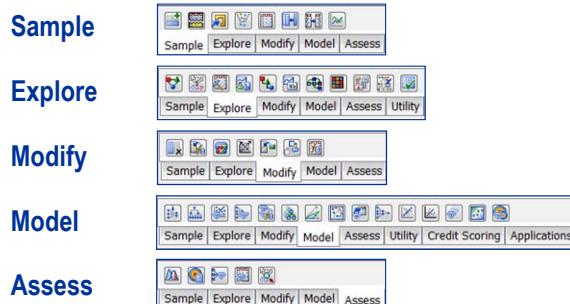
103



104

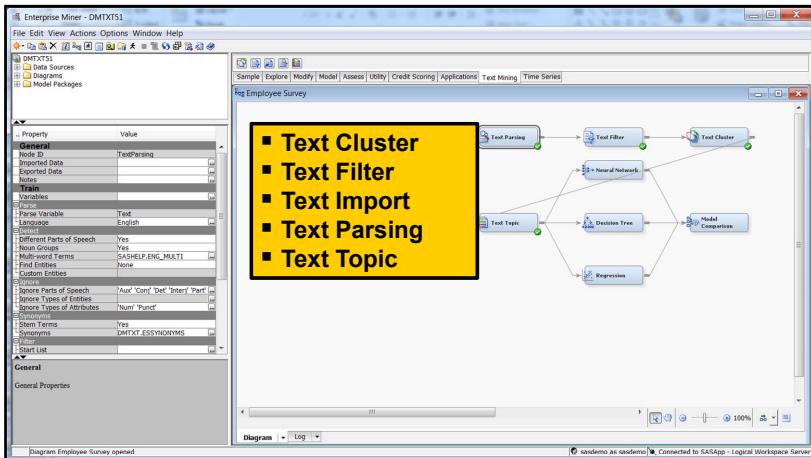
SAS Enterprise Miner: Interface Tour

SEMMA tools palette



105

Using SAS Text Miner



106

Using the Text Parsing Node

.. Property	Value
General	
Node ID	TextParsing
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Parse	
Parse Variable	
Language	English
Detect	
Different Parts of Speech	Yes
Noun Groups	Yes
Multi-word Terms	SASHHELP.ENG_MULTI
Find Entities	None
Custom Entities	
Ignore	
Ignore Parts of Speech	'Aux' 'Conj' 'Det' 'Interj' 'Part'
Ignore Types of Entities	
Ignore Types of Attributes	'Num' 'Punct'
Synonyms	
Stem Terms	Yes
Synonyms	SASHHELP.ENGSYNIMS
Filter	
Start List	
Stop List	SASHHELP.ENGSTOP
Report	
Number of Terms to Display	20000

Text Parsing Properties

107

Using the Text Parsing Node

The Text Parsing node

- performs parsing to build the corpus dictionary
- associates terms with parts of speech and controls which parts of speech to recognize
- performs stemming to equate terms that are different verb tenses of the same verb, or to equate terms that are either singular or plural versions of the same noun
- identifies up to 16 entities such as address, company name, currency, and person's name
- imports custom entities created by a product such as SAS Content Categorization
- controls recognition of numbers or punctuation as separate terms.

108

The Reference Help for SAS Enterprise Miner 6.1 has the following description of the Custom Entities property.

Custom Entities specifies the path (relative to the SAS Text Miner server) to a file that has been output from SAS Content Categorization with Teragram Contextual Extraction and contains compiled custom entities. Valid files have the extension .li. No custom entity should have the same name as a standard entity.



Custom entities are not discussed in this course because you can license SAS Text Miner without licensing SAS Content Categorization. (Other SAS courses provide instruction about SAS Content Categorization.)

Using the Text Parsing Node

The Text Parsing node special tables:

- Synonyms
- Multi-word term dictionary
- Start/stop list - table of terms to include/exclude from the analysis

109

Using the Text Parsing Node

Verb stemming example:

- type
- typed
- typing
- types

Noun stemming examples:

- house, houses
- matrix, matrices
- criteria, criterion

110

Using the Text Parsing Node

Dictionaries when a **stop list** is specified:

- **Corpus dictionary:** the union of all terms in the corpus (derived, not specified)
- **Stop list:** a dictionary of terms to be ignored in the analysis (specified by the user)
- **Start list:** terms in the corpus dictionary that are not in the stop list (derived)

The stop list contains **low information** terms that only add **noise** to the analysis. **Noisy data** has no descriptive or predictive value.

111

Using the Text Parsing Node

Dictionaries when a **start list** is specified:

- **Corpus dictionary:** the union of all terms in the corpus (derived, not specified)
- **Start list:** a dictionary of terms to be used in the analysis (specified by the user)
- **Stop list:** terms in the corpus dictionary that are not in the start list (derived)

The start list can be a technical/business dictionary developed by the analyst or obtained from other sources.

112

Using the Text Filter Node

.. Property	Value
General	
Node ID	TextFilter
Imported Data	[...]
Exported Data	[...]
Notes	[...]
Train	
Variables	[...]
Spelling	
Check Spelling	No
Dictionary	[...]
Weightings	
Frequency Weighting	Default
Term Weight	Default
Term Filters	
Minimum Number of Documents	4
Maximum Number of Terms	.
Import Synonyms	[...]
Document Filters	
Search Expression	[...]
Subset Documents	[...]
Results	
Filter Viewer	[...]
Spell-Checking Results	[...]
Exported Synonyms	[...]
Report	
Terms to View	All
Number of Terms to Display	20000

Text Filter Properties

113

Using the Text Filter Node

Text Filter Properties

- Frequency weights and term weights.
- The Check Spelling property uses a spelling dictionary and word-similarity algorithms to find and correct misspellings.
- The Minimum Number of Documents property performs frequency filtering for rare terms. This property can be used rather than searching for rare terms and adding them to the stop list.
- The Filter Viewer enables you to interactively control terms to drop or keep, interactively create synonyms, and perform queries and view concept links.

114

Using the Text Filter Node

Analysis Features

- Frequency weights
 - Log (default)
 - Binary
 - None (count or frequency)
- Term weights
 - Entropy (default)
 - Inverse Document Frequency
 - Mutual Information
 - None

115

Using the Text Filter Node

Query filters have the following characteristics:

- can be used in the Properties panel and in the Interactive Filter Viewer
- return documents satisfying the query
- can be used to subset the document collection for the continuing downstream analysis of the collection

116

Using the Text Filter Node

Query Operators

- $+term$ returns all documents having at least one occurrence of *term*.
- $-term$ returns all documents having zero occurrences of *term*.
- “*text string*” returns all documents having at least one occurrence of the quoted text string.
- $string1*string2$ returns all documents that have a term that begins with *string1*, ends with *string2*, and has text in between.
- $>\#term$ returns all documents that have *term* or any of the synonyms that are associated with *term*.

Using the Text Cluster Node

.. Property	Value
General	
Node ID	TextCluster
Imported Data	[...]
Exported Data	[...]
Notes	[...]
Train	
Variables	[...]
Transform	
SVD Resolution	Low
Max SVD Dimensions	100
Cluster	
Exact or Maximum Number	Maximum
Number of Clusters	40
Cluster Algorithm	EXPECTATION-MAXIMIZATION
Descriptive Terms	8

118

The Text Cluster node divides a document collection into mutually exclusive clusters. By default, eight descriptive terms are used to describe the clusters.

From **Help** \Rightarrow **Contents**:

“The Text Cluster node uses a descriptive terms algorithm to describe the contents of both EM clusters and hierarchical clusters. If you specify to display m descriptive terms for each cluster, then the top $2*m$ most frequently occurring terms in each cluster are used to compute the descriptive terms.”

“For each of the $2*m$ terms, a binomial probability for each cluster is computed. The probability of assigning a term to cluster j is $\text{prob} = F(k|N, p)$. Here, F is the binomial cumulative distribution function, k is the number of times that the term appears in cluster j, N is the number of documents in cluster j, p is equal to $(\text{sum}-k)/(\text{total}-N)$, sum is the total number of times that the term appears in all the clusters, and total is the total number of documents. The m descriptive terms are those that have the highest binomial probabilities.”

“Descriptive terms must have a keep status of Y and must occur at least twice (by default) in a cluster.”

Using the Text Cluster Node

Analysis Features

- Latent Semantic Analysis (LSA) using the Singular Value Decomposition (SVD)
 - SVD Resolution: Low, Medium, or High
 - Max SVD Dimensions: Up to 250
- Cluster derivation
 - Exact or Maximum Number
 - Maximum number of clusters
 - Cluster Algorithm
 - Expectation-Maximization (EM)
 - Hierarchical

119

The use of *singular value decomposition* (SVD) is called the linear algebra approach to text mining, information retrieval, Web analytics, and so on. As mentioned in a previous section, this algebraic operation is the foundation of an approach that goes by many different names:

- Latent Semantic Indexing (LSI)
- Latent Semantic Analysis (LSA)
- Vector Space Model (VSM)

Using the Text Topic Node

.. Property	Value
General	
Node ID	TextTopic
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
User Topics	...
Term Topics	
Number of Single-term Topics	0
Learned Topics	
Number of Multi-term Topics	25
Correlated Topics	No
Results	
Topic Viewer	...

120

Using the Text Topic Node

Text Topic Properties

- A custom topic table can be supplied by the user. The table can be imported, or the table can be manually created with a table editor.
- The user can request up to 1,000 single-term topics if more than 1,000 terms are in the imported term list. By default, no single-term topics are derived.
- A user can specify up to 1,000 multi-term topics if more than 1,006 terms are imported. By default, 25 multi-term topics are derived.

121

Using the Text Topic Node

Topics

- Single-term topics are not the same as filtering on a single term. For example, a topic can be derived based on the single-term *price*, but documents might be labeled as not having the topic even if the term *price* is present in the document.
- The node might return fewer topics than requested. After the designated number of topics are derived, the node can, for example, decide that topics 24 and 25 are not sufficiently distinct to warrant including both topics. If so, topic 25 (based on order of importance) is dropped.

122

Using the Text Topic Node

Custom Topics

- A topic consists of a label, and one or more terms, with each term having a role and a weight.
- The weight associated with a term-role pair indicates the relative importance of the term-role pair to the topic.
- A weight of 1 is the highest importance, and a weight of 0 is the lowest. Term-role pairs with a weight of 0 are excluded, because, by default, any term-role pair not in the custom topic table will be assigned a weight of 0 and will be ignored when determining the strength of the topic in a given document.

123

Using the Text Topic Node

A Custom Topic Table

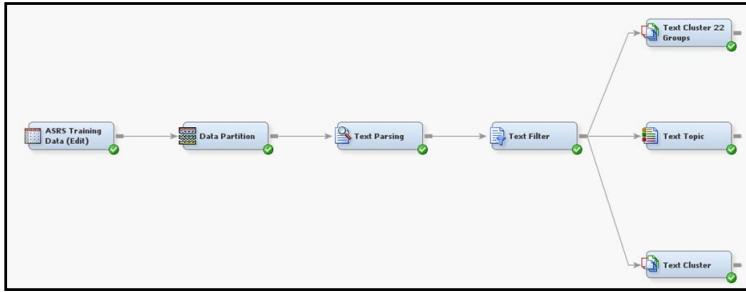
Topic	Term	Role	Weight
analytics	analytics	noun	1.0
analytics	analyze	verb	0.9
analytics	logistic regression	NOUN_GROUP	0.5
data	data	noun	1.0
data	data warehouse	NOUN_GROUP	0.7
data	analyze	verb	0.2

Columns in the custom topic table have names: **_topic_**, **_term_**, **_role_**, and **_weight_**.

124

Weights can be any numeric value, positive or negative. Negative weights imply that the term supports the negative, or opposite, of the concept. A 0-1 system is the easiest to use until you gain familiarity with the creation of custom topics.

Using the Text Miner Nodes





Text Mining SAS Course Descriptions

This demonstration illustrates how to perform text mining using the Text Miner node applied to the SAS course descriptions data.

SAS Education supports over 300 courses. Prospects often have questions about curriculum and specific course content. For example, a prospect might ask for information about courses that discuss neural networks. Text mining provides a solution for automating queries based on keywords.

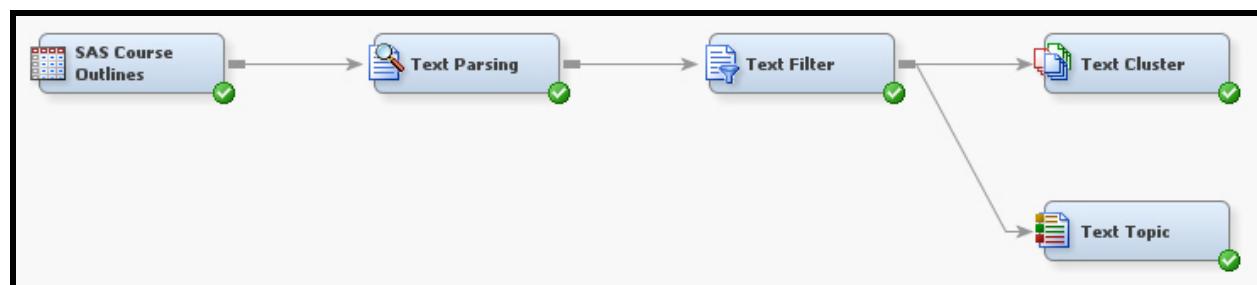
Descriptions of the SAS courses can be found at support.sas.com/training.

The SAS course descriptions data set **DMTXT.SASCOURSES** contains descriptions of courses supported in 2011. The data set has 735 rows (documents) and four columns. Some courses have multiple versions that are associated with different releases of the software.

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper L
CourseCode	ID	Nominal	No		No	.	.
CourseOutline	Text	Nominal	No		No	.	.
CourseTitle	Label	Nominal	No		No	.	.
Date	Time ID	Interval	No		No	.	.

The variable **CourseOutline** contains the course outline text.

The following flow diagram implements the text mining analysis:



Frequency filtering is a methodology to create or add to a stop list. You can run the Text Parsing node with the default stop list and then use frequency filtering to add terms to this list. Frequency filtering specifies a cutoff frequency. Terms with frequency below the cutoff are added to the stop list. You can also specify a cutoff frequency at the high end so that terms with frequency above the cutoff are added to the stop list. For creating a start list, just keep terms with frequencies between the high and low cutoff values. The start list **DMTXT.SASCOURSESTART** contains a start list that was obtained using domain knowledge and frequency filtering. Chapter 3 discusses Zipf's Law as it pertains to identifying low information terms. Zipf's Law provides the justification for employing frequency filtering.

1. Create a diagram named SAS Course Outlines. If you need to, create a data source for the **DMTXT.SASCOURSES** data set. Drag this data source into the diagram.
2. Attach a Text Parsing node to the Input Data Source node. Change the Synonyms property to **No Data Set to be Specified**. Select as a start list **DMTXT.SASCOURSESTART**. Run the Text Parsing node.
3. Attach a Text Filter node to the Text Parsing node. Change Frequency Weighting from **Default** to **Log**. Change Term Weighting from **Default** to **Inverse Document Frequency**. Log is the default frequency weight, but entropy is the default term weight. Inverse document frequency is recommended for documents larger than a paragraph. Run the Text Filter node.
4. Open the Filter Viewer, which is also called the Interactive Filter Viewer.

The screenshot shows the 'Interactive Filter Viewer' application window. It contains two main panes. The top pane, titled 'Documents', displays a table of course outlines with columns for COURSEOUTLINE, COURS..., COURS..., and DATE. The bottom pane, titled 'Terms', displays a table of terms with columns for TERM, FREQ, # DOCS, KEEP ▼, WEIGHT, ROLE, and ATTRIBUTE. Both tables have scroll bars.

COURSEOUTLINE		COURS...	COURS...	DATE
Define Tools importing data into JMP understanding the JMP data table		lw6jov	JMP Ove...	2009-02-...
Introduction touring SAS Enterprise Miner 5.3 placing SAS Enterprise Miner in		aaem53	Applied ...	2008-10-...
Introduction introduction to SAS Enterprise Miner Accessing and Assaying		aaem61	Applied ...	2011-06-...
What Is ETL? Setting up SAS for ETL Importing Model Data Schema		abctl	ETL for ...	2009-09-...
Introduction to Greenhouse Gas Modeling and Terminology references What is		abghgm	Greenho...	2009-06-...
Introduction Overview of SAS Activity-Based Management 7.1 Architecture,		abmicmbc	SAS Acti...	2010-09-...
Introduction What's New in SAS Activity-Based Management 7.1 - Product		abmimpbc	SAS Acti...	2010-09-...
Introducing Activity-Based Management activity-based management: why and		abmo6	ABC Mod...	2009-05-...
Introducing Activity-Based Management activity-based management: why and		abmo64	ABC Mod...	2010-09-...
Introducing Activity-Based Management activity-based management: why and		abmo6h	ABC Mod...	2009-07-...
Activity-Based Management learning why and how to use activity-based		abmo71	ABC Mod...	2010-12-...
Project Implementation project startup project timeline data collection		abmod	Advance...	2010-05-...
Topics Covered in Webinar and Supporting Documentation as Prerequisites		abmts	SAS Acti...	2010-09-...
Reporting for Business Analysis and Model Validation defining business		abrc	SAS Acti...	2010-03-...
Standard ABM Reports and Report Dependencies identify report templates and		abrc64	SAS Acti...	2010-12-...
Standard SAS Activity-Based Management Reports and Report Dependencies		abrc71	SAS Acti...	2011-05-...
What is ETL? describe staging tables describe ETL view staging tables in SAS		abtl64	Data Inte...	2010-12-...
What is ETL? Setting Up SAS For ETL SAS Enterprise Guide and SAS Data		abtl71	Data Inte...	2011-04-...
The one-on-one workshop will be geared toward your specific needs. Areas of		acbacw	Business...	2011-03-...
The one-on-one workshop will be geared toward your specific needs. Areas of		acbacw2	Business...	2011-03-...

Terms							
	TERM	FREQ	# DOCS	KEEP ▼	WEIGHT	ROLE	ATTRIBUTE
	addressed	492	492	<input checked="" type="checkbox"/>	1.579	Prop	Alpha
	data	1815	452	<input checked="" type="checkbox"/>	1.701	Noun	Alpha
[+]	introduction	794	377	<input checked="" type="checkbox"/>	1.963	Noun	Alpha
[+]	create	907	293	<input checked="" type="checkbox"/>	2.327	Verb	Alpha
	overview	506	248	<input checked="" type="checkbox"/>	2.567	Noun	Alpha
	business	494	244	<input checked="" type="checkbox"/>	2.591	Noun	Alpha
[+]	analysis	584	226	<input checked="" type="checkbox"/>	2.701	Noun	Alpha
[+]	define	424	186	<input checked="" type="checkbox"/>	2.982	Verb	Alpha
	management	444	179	<input checked="" type="checkbox"/>	3.038	Noun	Alpha
	analytics	327	179	<input checked="" type="checkbox"/>	3.038	Prop	Alpha
[+]	process	308	169	<input checked="" type="checkbox"/>	3.121	Noun	Alpha
[+]	platform	347	169	<input checked="" type="checkbox"/>	3.121	Noun	Alpha
[+]	review	246	161	<input checked="" type="checkbox"/>	3.191	Verb	Alpha
[+]	model	405	150	<input checked="" type="checkbox"/>	3.293	Verb	Alpha
[+]	report	252	149	<input checked="" type="checkbox"/>	3.302	Verb	Alpha
	web	295	148	<input checked="" type="checkbox"/>	3.312	Noun	Alpha
[+]	include	170	148	<input checked="" type="checkbox"/>	3.312	Verb	Alpha
	studio	349	140	<input checked="" type="checkbox"/>	3.392	Noun	Alpha

If you sort the TERM column in the Terms table by clicking on the header cell containing the word TERM, then you can use a quick-find feature. Select any term in the TERM column. Then type in the first letter of the term that you want to find. The window is scrolled to the first term starting with that letter. You can also use the **Edit ⇔ Find** feature to go directly to a desired term.

- In the Filter Viewer, select **Edit ⇔ Find**, and type in the two word phrase **neural network**. After you click **OK**, you are taken to the first cell in the TERM column that contains the phrase “neural network.” Right-click in the cell containing the phrase “neural network,” and select **Add Term to Search Expression**. The Search window will contain **>#"neural network"**. The quotes are required if the search expression has more than one word. If you look at the filter rules, you will see that this expression will search for documents containing “neural network” and any of the synonyms of “neural network.” Click **Apply**, and the following results appear

The screenshot shows the 'Interactive Filter Viewer' application window. At the top, there is a menu bar with File, Edit, View, Window, and a search bar containing the query ># "neural network". Below the search bar are Apply and Clear buttons. The main area is divided into two sections: 'Documents' and 'Terms'.

Documents Section:

COURSEOUTLINE	TEXTFILTER_SNIPPET	TEXTFILTER_RELEVANCE	COURS...	COURS...	DATE
Introduction touring SAS Enterprise Miner 5.3 placing SAS Enterprise Miner in	... Predictive Modeling with	0.3	aaem53	Applied ...	2008-10...
Introduction introduction to SAS Enterprise Miner Accessing and Assaying	... Predictive Modeling with	0.2	aaem61	Applied ...	2011-06...
Review of Basel I and Basel II the Basel I and Basel II regulation standard	... for scorecard development	0.5	basel52	Credit Ri...	2008-06...
Review of Basel I and Basel II the Basel I and Basel II regulation standard	... for scorecard development	0.5	basel53	Credit Ri...	2008-06...
A Review of Basel II and PD Modeling Basel I and Basel II a brief review of PD	... for scorecard development	0.5	basela52	Advance...	2008-06...
A Review of Basel II and PD Modeling Basel I and Basel II a brief review of PD	... for scorecard development	0.5	basel53	Advance...	2008-06...
A Review of Basel II and PD Modeling Basel I and Basel II a brief review of PD	... for scorecard development	0.5	basellia	Advance...	2008-05...
Review of Basel I and Basel II application scoring, behavioral scoring, and	... for scorecard development	0.5	bb4c	Credit Ri...	2010-03...
Review of Basel I and Basel II application scoring, behavioral scoring, and	... for scorecard development	0.5	bb4c61	Credit Ri...	2010-03...
Predictive Modeling for Customer Intelligence: The KDD Process Model A	... leave-one-out)	0.7	bdcml	Advance...	2009-08...
Refresher: the Customer Analytics Process Model basic nomenclature (e.g.	... leave-one-out)	1.0	bdcml61	Advance...	2011-06...
Refresher: the Customer Analytics Process Model basic nomenclature (e.g.	... leave-one-out)	1.0	bdcml71	Advance...	2011-07...
Introduction to Data Mining what is data mining? directed and undirected data	... build decision trees Neural	0.7	bdmt53	Data Min...	2008-08...
Introduction to Data Mining what is data mining? directed and undirected data	... build decision trees Neural	0.7	bdmt61	Data Min...	2009-11...
Predictive Analytics and Exploratory Data Mining the relationship between	... regression decision trees	0.4	beap	Explorat...	2009-09...
Predictive Analytics and Exploratory Data Mining the relationship between	... regression decision trees	0.4	beap61	Explorat...	2011-02...
What Is Electric Load Forecasting? an overview of the electric power industry	... demand response Artificial	0.2	belf	SAS for ...	2011-03...
Survival Data time-dependent outcomes derived from customer event histories	... regression spline and neural	0.3	bmce	Survival ...	2009-02...
Introduction definition, examples, and brief history review of some basic	... decision trees , neural	0.6	bwawi	Web Ana...	2010-02...
The one-on-one workshop will be geared toward your specific needs. Specific	... decision trees , neural	0.5	dme2e	Expert-to...	2008-05...
Solving Business Problems Using Analytics approaches to solving business	... decision trees , neural	0.5	dmepw	Expert P...	2009-07...
Introduction to Neural Networks using the NLIN procedure for nonlinear	... Introduction to Neural	0.7	dmnn53	Neural N...	2009-08...
Introduction to Neural Networks using the NLIN procedure for nonlinear	... Introduction to Neural	0.7	dmnn61	Neural N...	2010-06...

Terms Section:

TERM	FREQ	# DOCS	KEEP ▼	WEIGHT	ROLE	ATTRIBUTE
model	144	29	<input checked="" type="checkbox"/>	1.096	Verb	Alpha
model	96	28	<input checked="" type="checkbox"/>	1.147	Noun	Alpha
enterprise	68	27	<input checked="" type="checkbox"/>	1.199	Noun	Alpha
miner	71	27	<input checked="" type="checkbox"/>	1.199	Noun	Alpha
neural	49	27	<input checked="" type="checkbox"/>	1.199	Adj	Alpha
regression	60	25	<input checked="" type="checkbox"/>	1.31	Noun	Alpha
analysis	92	24	<input checked="" type="checkbox"/>	1.369	Noun	Alpha
network	69	24	<input checked="" type="checkbox"/>	1.369	Noun	Alpha
data	136	23	<input checked="" type="checkbox"/>	1.431	Noun	Alpha
neural network	48	23	<input checked="" type="checkbox"/>	1.431	Noun Group	Alpha
model	50	23	<input checked="" type="checkbox"/>	1.431	Adj	Alpha
tree	31	20	<input checked="" type="checkbox"/>	1.632	Verb	Alpha
score	34	19	<input checked="" type="checkbox"/>	1.706	Verb	Alpha

The TEXTFILTER_RELEVANCE column contains the result of the inner product calculation described in the previous discussion of a Boolean query. The cutoff is not displayed, but all documents that produced a result above the cutoff are returned. You can see that courses with codes BDMCI61 and BDMCI71 have the highest relevance, whereas AAEM61 and BELF have the lowest relevance of the documents that can be viewed in the window above. If you scroll down, you will see other courses with a relevance value of 0.2. For the most part, the courses with lower relevance scores include neural network material, but the courses include other material as well, and the neural network part takes up a small portion of the course. A total of 31 courses are returned. You can see that at most 23 of these courses contain the noun group “neural network.” This often occurs because other terms tend to co-occur with the term of interest, and co-occurrence can lead to a high relevance score. Recall the discussion of “cats and dogs” being a phrase related to meteorology.

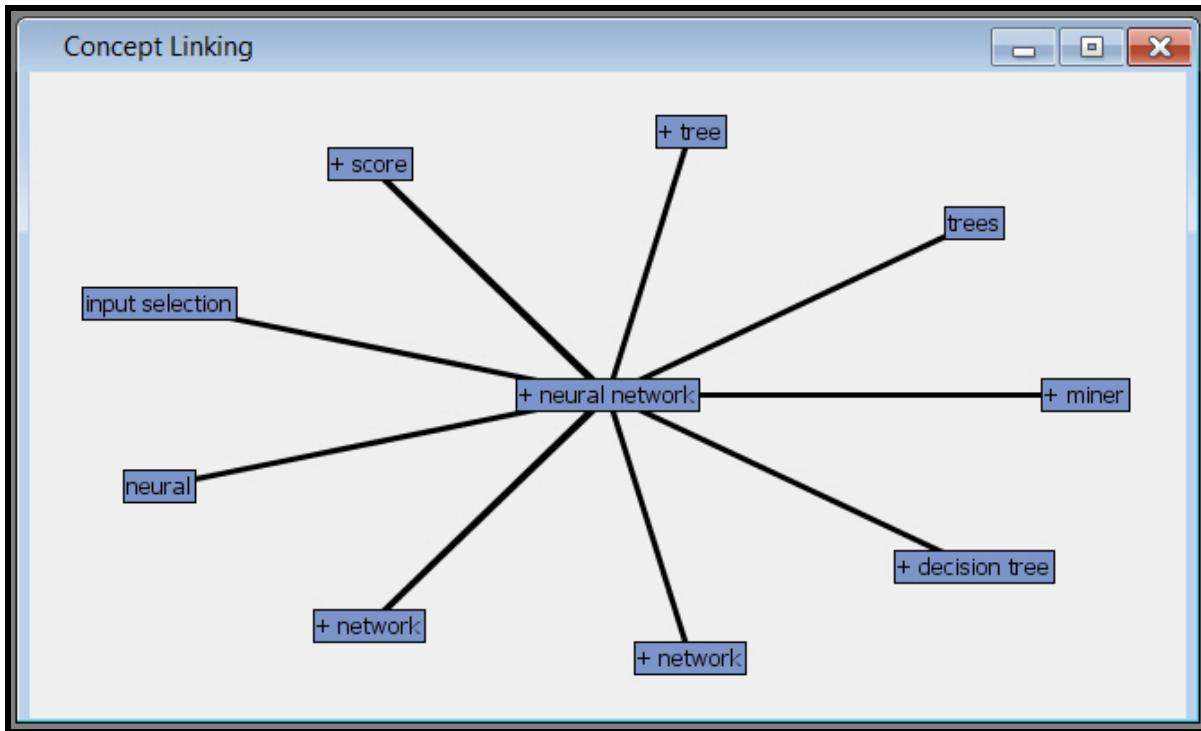
6. Select the document corresponding to the course with code BASEL52. Select **Edit** ⇒ **Toggle Show Full Text**. You can read the course outline for BASEL52.

The screenshot shows the 'Interactive Filter Viewer' application window. At the top, there's a menu bar with File, Edit, View, Window. Below the menu is a toolbar with a magnifying glass icon, a search input field containing '>#"neural network"', and buttons for Apply and Clear. The main area is titled 'Documents' and contains a table with several columns: COURSEOUTLINE, TEXTFILTER_SNIPPET, TEXTFILTER_RELEVANCE, COURS..., COURSETITLE, and D/. The first row of the table is expanded to show the full document content in the COURSEOUTLINE column. This content discusses various topics related to neural networks, including regression, logistic regression, decision trees, linear programming, k-nearest neighbor, cumulative logistic regression, input selection, segmentation modeling, and backtesting. The relevance score for this document is 0.5, and it corresponds to course code baseL52, title 'Credit Risk Modeling for Basel II Using SAS', and ID 200. Below the table is a small grid labeled 'Decision Tree' with two rows: 'modeling' and 'technique'. The 'modeling' row has values 34, 17, 1.867, Prop, Alpha. The 'technique' row has values 33, 17, 1.867, Noun, Alpha.

The smaller relevance score is suggested in that most of the outline discusses topics not related to neural networks, but neural networks and terms that tend to appear in relation to neural networks (for example, KS-statistic and ROC curve) do appear in the document. You can select **Edit** ⇒ **Toggle Show Full Text** again to return to one row per document.

Of the 31 documents returned, most appear to be legitimately related to neural networks, so precision is around 100%. However, there is only one set of courses that address neural networks exclusively. Those are the courses that begin with DMNN (Data Mining with Neural Networks), which include DMNN, DMNN53, and DMNN61. DMNN is obsolete and is no longer offered, so it does not appear in the data set. Surprisingly, DMNN53 and DMNN61 produced a relevance score of 0.7, a value below that of other courses. Consequently, the scoring irregularity suggests some noise in attempting to automatically retrieve only the courses relevant to a customer's inquiry. Using this methodology as-is poses a risk of suggesting irrelevant courses to customers. Furthermore, extra effort will be required to calculate a recall value to see how many false negatives resulted from the neural network query. Despite these setbacks, the results actually look quite promising, at least for this one query.

7. Select **Clear** \Rightarrow **Apply** to return all of the documents in the collection. Navigate back to the neural network row in the Terms table. Right-click on the neural network **TERM** cell and select **View Concept Links**. The concept link plot appears.

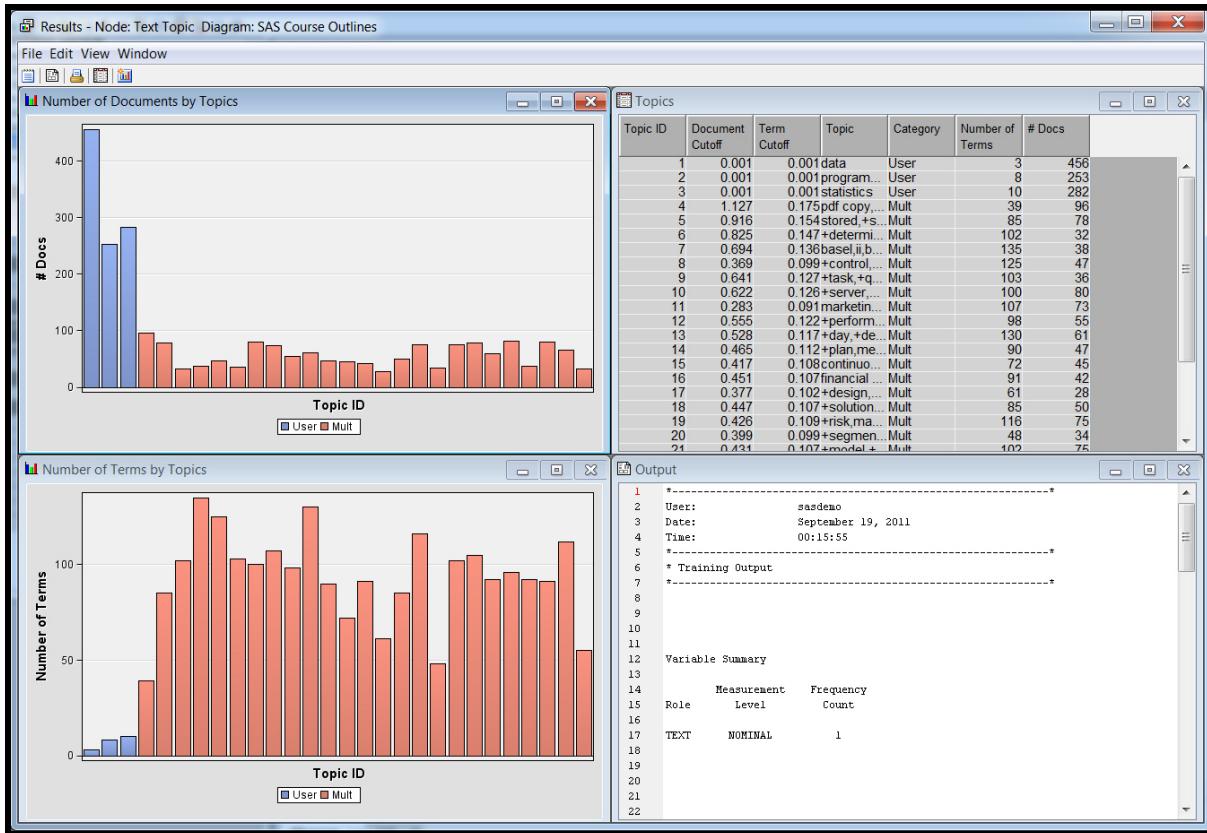


“Decision tree” shows up as having a strong association with “neural network” because of the 20 course outlines that mention decision trees, 13 also mention neural networks.

8. Close the Filter Viewer.

9. Attach a Text Topic node to the Text Filter node. For User Topics, select the data set **DMTXT.SASTOPICS**. Keep all of the other default settings. Run the Text Topic node.

10. Access the Results window.



The three bars to the right in both bar charts relate to the three custom topics: data, programming, and statistics. The remaining bars relate to derived topics. The Number of Terms by Topics bar chart reveals that only a few terms are used to define the custom topics. Perhaps more terms should be used. The Number of Documents by Topics bar chart reveals that the custom topics are much more prevalent than the derived topics. The smallest custom topic, programming, appears in 253 documents. The most popular derived topic appears in 96 documents. You can get the topic frequencies by positioning the cursor over the bar related to a topic. The plots are dynamic.

11. Close the Results window.

12. Open the Topic Viewer. The following window appears:

The screenshot shows the "Interactive Topic Viewer" application window. It has three main sections: "Topics", "Terms", and "Documents".

Topics section:

Topic	Category	Term Cutoff	Document Cutoff
data	User	0.001	0.001
programming	User	0.001	0.001
statistics	User	0.001	0.001
pdf copy, +class, +material, web, +course ma	Mult	0.175	1.127
stored, +store, +platform, studio, business	Mult	0.154	0.916
+determine, +control, +solution, +project, +	Mult	0.147	0.825

Terms section:

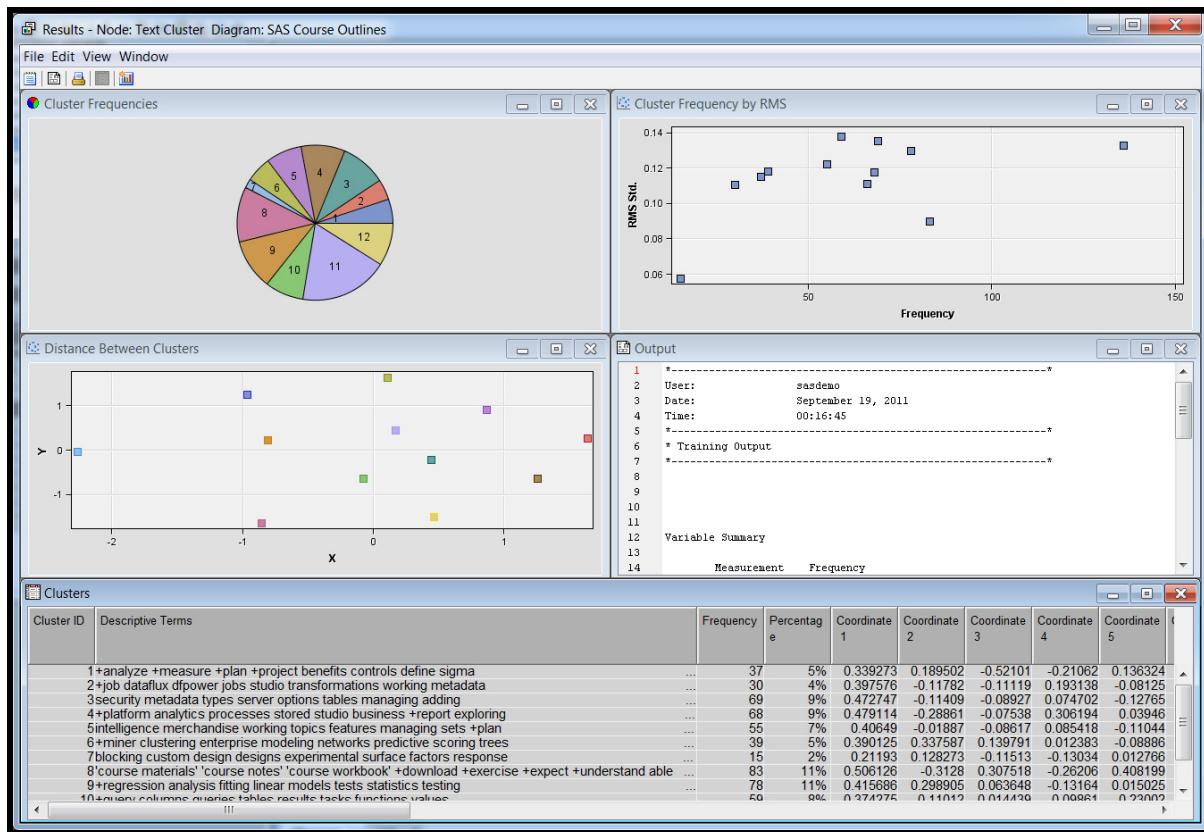
Topic Weight	+	Term	Role	# Docs	Freq
1		data integration	Noun Group	13	17
0.7		data	Noun	452	1815
0.7	+	data set	Noun Group	53	120
0		addressed	Prop	492	492
0	+	introduction	Noun	377	794
0	+	create	Verb	293	907

Documents section:

Topic Weight	CourseOutline	CourseCode	CourseTitle	
0.369	Introduction introducing data	dipcd	Profiling and Cleansing Data	2008-09-25
0.316	Working with Existing SAS	hecpes2	SAS Programming	2008-01-30
0.278	Introduction examine the	hecpes1	SAS Programming	2008-01-30
0.267	Introduction to SAS Data	difast	SAS 9 Data Integration Fast	2009-05-19
0.266	Introduction review of SAS	ecprg2	SAS Programming 2: Data	2009-02-19
0.262	Overview of the SAS	smitech	Technical Overview of SAS	2008-05-20
0.254	Overview of SAS providing	gahsp1	SAS Programming 1 for	2010-05-19

13. A custom topic is like a predefined query. The topic weight shown in the documents window determines whether the topic is present (the query is satisfied). If the topic weight exceeds the document cutoff, then the document is classified as having the topic. In later chapters, you learn how to manipulate topic definitions. For now, close the Topic Viewer.

14. Attach a Text Cluster node to the Text Filter node. You could attach it to the Text Topic node, but the Text Cluster node does not use anything produced by the Text Topic node. If you want to combine the exported data coming from the Text Topic node with the exported data coming from the Text Cluster node, you should connect the nodes in series. In the Federalist Papers demonstration, the nodes were connected in series so that columns produced by both nodes could be used in training a logistic regression model. Use the default setting and run the Text Cluster node. Open the Results window.



A total of 12 clusters are derived based on the cubic clustering criterion. The descriptive terms help identify the courses that appear in each cluster. Careful scrutiny reveals that clusters 1 and 9 are analytic clusters. Cluster 6 appears to be a relatively pure SAS Enterprise Miner cluster. While the Text Filter and Text Topic nodes appear to be the most useful for this project, the Text Cluster node presents a “divide-and-conquer” methodology that could be useful for identifying mutually exclusive categories of courses. Because course topics rarely fall into mutually exclusive sets of documents, the Text Cluster node is of limited value to this study.

Technical Details

The following material is extracted from the Reference Help for SAS Enterprise Miner 6.1.

Parts of Speech in SAS Text Miner

SAS Text Miner can identify the part of speech for each term in a document based on the context of that term. Terms are identified as one of the following parts of speech:

- Abbr (abbreviation)
- Adj (adjective)
- Adv (adverb)
- Aux (auxiliary or modal)
- Conj (conjunction)
- Det (determiner)
- Interj (interjection)
- Noun (noun)
- Num (number or numeric expression)
- Part (infinitive marker, negative participle, or possessive marker)
- Pref (prefix)
- Prep (preposition)
- Pron (pronoun)
- Prop (proper noun)
- Punct (punctuation)
- Verb (verb)
- VerbAdj (verb adjective)

Noun Groups in SAS Text Miner

SAS Text Miner can identify noun groups, such as *clinical trial* and *data set*, in a document collection. Noun groups are identified based on linguistic relationships that exist within sentences. Syntactically, these noun groups act as single units and you can, therefore, choose to parse them as single terms.

- If stemming is on, noun groups are stemmed. For example, the text *amount of defects* is parsed as *amount of defect*.
- Frequently, shorter noun groups are contained within larger noun groups; both the shorter and larger noun groups appear in parsing results.

Entities in SAS Text Miner

An *entity* is any of several types of information that SAS Text Miner can distinguish from general text. If you enable SAS Text Miner to identify them, entities are analyzed as a unit, and they are sometimes normalized. When SAS Text Miner extracts entities that consist of two or more words, the individual words of the entity are also used in the analysis.

Out of the box, SAS Text Miner identifies the following standard entities:

- ADDRESS (postal address or number and street name)
- COMPANY (company name)
- CURRENCY (currency or currency expression)
- DATE (date, day, month, or year)
- INTERNET (e-mail address or URL)
- LOCATION (city, country, state, geographical place/region, political place/region)
- MEASURE (measurement or measurement expression)
- ORGANIZATION (government, legal, or service agency)
- PERCENT (percentage or percentage expression)
- PERSON (person's name)
- PHONE (phone number)
- PROP_MISC (proper noun with an ambiguous classification)
- SSN (Social Security number)
- TIME (time or time expression)
- TIME_PERIOD (measure of time expressions)
- TITLE (person's title or position)
- VEHICLE (motor vehicle including color, year, make, and model)

You can also use SAS Content Categorization with Teragram Contextual Extraction to define custom entities and import these for use in a Text Parsing node. When you create compiled custom entity files, ensure that you specify September 14, 2009 as the compatibility date. (Valid files have the extension .li.) Otherwise, the files cannot be used in SAS Text Miner.

Entities are normalized in these situations:

- SAS Text Miner uses a fixed dictionary of company and organization names in order to identify these entity types, and entities of this type will frequently be associated with a parent. For example, if *IBM* appears in the text, it is returned with the predefined parent *International Business Machines*. Typically, the longest and most precise version of a name is used as the parent form.
- SAS Text Miner normalizes entities that have an ISO (International Standards Organization) standard (dates/years, currencies, and percentages). Rather than return the normalization as a parent of the original term, these normalizations actually replace the original term.
- You can alter any parent forms that are returned by editing the synonym list. Place terms that you want to identify as an entity in the *term* variable, the parent to associate with it in the *parent* variable, and place the entity category in the *category* variable. Then rerun the node.

Attributes in SAS Text Miner

When a document collection is parsed, SAS Text Miner categorizes each term as one of the following attributes, which gives an indication of the characters that compose that term:

- Alpha, if characters are all letters
- Num, if term characters include a number
- Punct, if the term is a punctuation character
- Mixed, if term characters include a mix of letters, punctuation, and whitespace
- Entity, if the term is an entity



Exercises

2. Using the SAS Text Miner Nodes to Investigate SAS Course Outlines

Reproduce the SAS course outlines demonstration.

1.4 Chapter Summary

Data mining incorporates analytic techniques applied to problems of pattern discovery and predictive modeling. SAS Enterprise Miner uses the SEMMA methodology for addressing data mining problems.

Text mining is a specialized area of data mining that brings together algorithms and methods from natural language processing and information retrieval to solve problems involving a collection of documents. SAS Text Miner provides modern techniques for solving text mining problems.

SAS Text Miner includes five nodes: Text Import, Text Parsing, Text Filter, Text Cluster, and Text Topic. The nodes provide complete text mining capabilities.

For Additional Information

- Albright, Russell. 2004. *Taming Text with the SVD*. SAS Institute White Paper.
- Albright, R., J.A. Cox, and K. Daly. 2001. "Skinning the Cat: Comparing Alternative Text Mining Algorithms for Categorization." Proceedings of the 2nd Data Mining Conference of DiaMondSUG, Chicago, IL. DM Paper 113.
- Baldi, Pierre, Paolo Frasconi, and Padraic Smyth. 2003. *Modeling the Internet and the Web: Probabilistic Methods and Algorithms*. West Sussex, England: John Wiley and Sons.
- Berry, Michael W. (ed.). 2001. *Computational Information Retrieval*. Philadelphia: SIAM.
- Berry, Michael W., and Malu Castellanos (eds). 2008. *Survey of Text Mining II*. London: Springer.
- Berry, Michael W., and Murray Browne. 1999. *Understanding Search Engines*. Philadelphia: SIAM.
- Bosch, R.A., and J.A. Smith. 1998. "Separating hyperplanes and the authorship of the disputed Federalist Papers." *American Mathematical Monthly*. Vol. 105, No. 7:601-608.
- Cherniak, Eugene. 1993. *Statistical Language Learning*. Cambridge, Massachusetts: The MIT Press.
- Feldman, Ronen, and James Sanger. 2007. *The Text Mining Handbook*. New York: Cambridge University Press.
- Fung, Glenn. 2003. "Disputed Federalist Papers: SVM and Feature Selection via Concave Minimization." *Proceedings of the 2003 Conference of Diversity in Computing*, pp. 42-46, Atlanta, Georgia.
- Hand, David, Heikki Mannila, and Padhraic Smyth. 2001. *Principles of Data Mining*. Cambridge, Massachusetts: The MIT Press.
- Jurafsky, Daniel, and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, New Jersey: Prentice Hall.
- Konchady, Manu. 2006. *Text Mining Application Programming*. Boston: Charles River Media.
- Love, Harold. 2002. *Attributing Authorship: An Introduction*. Cambridge, Massachusetts: Cambridge University Press.
- Manning, Christopher D., and Hinrich Schutze. 2002. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.
- Mitchell, Tom M. 1997. *Machine Learning*. Boston: WCB McGraw-Hill.
- Mosteller, F., and D.L. Wallace. 1964. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. New York: Springer.
- Sanders, Annette, and Craig DeVault. 2004. "Using SAS® at SAS: The Mining of SAS Technical Support." paper 010-29, SUGI 29, Montréal, Quebec.
- SAS Institute Inc. 2002. *Getting Started with SAS® Text Miner Software, Release 8.2*. Cary, North Carolina: SAS Institute Inc.

Shannon, C.E. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal*, Vol. 27, pp. 379-423 and 623-656.

Song, Min, and Yi-fang Brook Wu. 2009. *Handbook of Research on Text and Web Mining Technologies*. IGI Global. Books 24x7. (accessed June 2, 2010)

Thisted, Ronald A. 1988. *Elements of Statistical Computing*. New York: Chapman and Hall.

Thuraisingham, Bhavani, Latifur Khan, Mamoun Awad, and Lei Wang. 2009. *Design and Implementation of Data Mining Tools*. Auerbach Publications. Books24x7. (accessed June 2, 2010)

Wakefield, Todd. 2004. "A Perfect Storm is Brewing: Better Answers are Possible by Incorporating Unstructured Data Analysis Techniques." *DM Direct*, August 2004.

Wallace, John, and Tracy Cermack. 2004. "Text Mining Warranty and Call Center Data: Early Warning for Product Quality Awareness." paper 003-29, SUGI 29, Montréal, Quebec.

1.5 Solutions

Solutions to Exercises

See the demonstrations in the course notes.

Solutions to Student Activities (Polls/Quizzes)

1.03 Multiple Choice Poll – Correct Answer

What was the Filter node used for in the Federalist Papers process flow?

- a. to remove unusual or influential documents
- b. to filter out non-informative terms
- c. to remove documents not authored by Madison or Hamilton
- d. to reduce the size of documents that have more than 32,767 characters

Chapter 2 Overview of Text Analytics

2.1 Data Preparation for Text Analytics	2-3
Demonstration: Using the Text Import Node	2-10
2.2 Forensic Linguistics.....	2-13
Demonstration: Stylometry for Forensic Linguistics	2-18
2.3 Information Retrieval.....	2-21
Demonstration: Retrieving Medical Information.....	2-25
Exercises	2-33
2.4 Text Categorization	2-34
Demonstration: Categorizing Reports in the ASRS	2-39
Exercises	2-48
2.5 Chapter Summary.....	2-51
2.6 Solutions	2-52
Solutions to Exercises	2-52
Solutions to Student Activities (Polls/Quizzes)	2-60

2.1 Data Preparation for Text Analytics

Objectives

- Explain some of the features of the SAS language and how SAS imports, reads, and modifies data sets.
- Describe the Text Import node for processing document collections.
- Explain the SAS Code node.

3

Often the most challenging part of the data mining process is obtaining and preprocessing the data. SAS provides a rich set of tools for data preparation.

SAS Data Access Features

- SAS provides **access engines** for commercial databases and common file types.
- SAS Enterprise Guide, the SAS windowing environment, and other SAS products or components support a **Data Import wizard**.
- The SAS language supports high-level and low-level **I/O functions** for reading data files.

4

SAS/ACCESS engines provide direct connectivity to popular commercial databases. A SAS/ACCESS engine hooks into the database supervisor to enable direct access to tables in the database. SAS/ACCESS engines also provide connectivity to common file formats, such as Microsoft Excel files.

The Data Import wizard enables access to common file formats, including comma-separated values (CSV) and Microsoft Excel files.

The SAS language provides a complete set of data access functions, including functions for low-level file I/O, so that, theoretically, any file format can be read and processed.

2.01 Multiple Answer Poll

Select all data file formats that you routinely encounter in your work.

- a. PC file formats, for example, Microsoft Excel (ODBC, OLE DB)
- b. Microsoft SQL Server tables
- c. Oracle tables
- d. Sybase tables
- e. Teradata tables
- f. DB2 tables
- g. MySQL tables
- h. Informix tables
- i. other

6

The SAS language also supports Perl regular expressions.

Selected Functionality of the SAS Language

- Support for Perl regular expressions
- Character string functions for searching and modifying text data
- Mathematical and statistical functions for working with numeric data
- Formats and informats for reading and writing data in most recognized data formats
- A macro facility to enable users to program complex operations for enterprise-wide use

7

Perl regular expressions enable you to use terse scripts for complex data operations on text files. For example, to preserve confidentiality, you might want to convert all postal codes to a generic phrase.

A Perl Regular Expression Macro in SAS

```
%macro PrivateUSAPostalCode(TextVar) ;
  &TextVar = prxchange(
    's/\d{5}/_PRIVATE_USA_POSTAL_CODE_/',
    -1,&TextVar);
%mend PrivateUSAPostalCode;
```

This macro uses the PRXCHANGE function and a Perl regular expression to convert the occurrence of a five-digit string into the string _PRIVATE_USA_POSTAL_CODE_.

8

Running SAS Programs

In SAS Enterprise Miner

- Program Editor: View \Rightarrow Program Editor
- SAS Code node: Utility tab

Outside SAS Enterprise Miner

- SAS Enterprise Guide
- SAS windowing environment
- Batch
- Other

9

SAS programs are used sparingly in this course. Preference is given to the use of the SAS Code node for running SAS programs.

2.02 Multiple Choice Poll

Which statement best reflects your situation with respect to SAS programming?

- a. I am comfortable programming using the SAS programming language.
- b. I have experience programming in other languages and am eager to learn how to program in SAS.
- c. I would rather use a point-and-click interface than write programs.
- d. I would rather crawl for a mile through broken glass than write a computer program.

11

Because data preparation can be the most arduous task in text mining, SAS Text Miner includes the Text Import node that facilitates reading all popular commercial document formats.

Previous versions of SAS Text Miner included a SAS macro named %TMFILTER for reading document collections. The %TMFILTER macro is still available in the current release of SAS Text Miner. The Text Import node provides the functionality of the %TMFILTER macro without requiring the user to create and execute a SAS program. A *SAS macro* is a script that can be compiled and executed to perform

complex tasks. At the simplest level, a SAS macro is a script that generates SAS code to be executed by the SAS supervisor. These scripts are often stored in SAS catalogs that can be accessed and viewed by users. Proprietary scripts are stored in compiled form and cannot be read by users. Some of the SAS macros included with SAS Enterprise Miner and SAS Text Miner are compiled. Sample SAS macros are stored in the catalog **SASHELP.EMUTIL**. For example, the SAS source file **SASHELP.EMUTIL.EXTDEMO.SOURCE** provides examples of SAS Enterprise Miner functionality that can be exploited using a SAS Code node. The catalog **SASHELP.EMTEXT** might contain macros related to text mining.

The Text Import Node

The Text Import node converts collections of documents into a single SAS table.

.. Property	Value
General	
Node ID	TextImport
Imported Data	
Exported Data	
Notes	
Train	
Import File Directory	C:\workshop\dmxt51\InsClaimsNotes
Destination Directory	C:\workshop\dmxt51\InsClaimsNotesDest
Language	'English'
Extensions	
Text Size	32000
Web Crawl	
URL	
Depth	1
Domain	Restricted
User Name	
Password	

12

continued...

The Text Import Node

Some of the supported document types:

- Microsoft Word (.doc, .docx)
- Microsoft Excel (.xls, .xlsx)
- Microsoft PowerPoint (.ppt, .pptx)
- Rich Text (.rtf)
- Adobe Acrobat (.pdf)
- ASCII Text (.txt)
- Some others:
 - Lotus
 - Corel
 - Framemaker

Over 100 file formats are supported!

13

The Text Import Node: Output SAS Table

Column Name	Description
TEXT	Text of each document, truncated to the length specified by the Text Size property
URI	Path to the input files that reside in the directory specified by the Import File Directory property
NAME	Name of the input file
FILTERED	Path to the directory that contains the HTML file. (This path corresponds to the value of the Destination Directory property.)
LANGUAGE	Most likely source language of the document, as determined by the LANGUAGE property
TRUNCATED	1 if TEXT contains truncated text, 0 otherwise

14

continued...

The Text Import Node: Output SAS Table

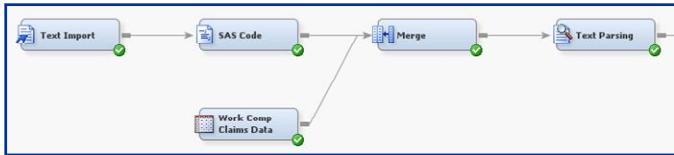
Column Name	Description
OMITTED	1 if the document was skipped (because of an unsupported input file type or some filtering error), 0 otherwise
FILTEREDSIZE	Size of the filtered document, in bytes
EXTENSION	File extension of the input document
CREATED	Date and time that the input document was created
ACCESSED	Date and time that the input document was last accessed
MODIFIED	Date and time that the input document was last modified
SIZE	Size of the input document, in bytes

15

You can use a SAS Code node to modify the SAS table produced by the Text Import node. For example, you can choose to drop variables such as **LANGUAGE**, **TRUNCATED**, **OMITTED**, and **EXTENSION**, because these variables are rarely used beyond the data preparation stage. Many document collections use a naming convention such that the path given in the **URI** field or the filename given in the **NAME** field can be used to derive ID or index variables.

Modifying Imported Data

```
data &EM_EXPORT_TRAIN;
attrib ClaimNo length=$12
      label="Claim Number"
      AdjusterNotes length=$256
      label="Adjuster Notes";
set &EM_IMPORT_DATA;
AdjusterNotes=Text;
ClaimNo=substr(Name,1,12);
keep ClaimNo AdjusterNotes Size;
run;
```



16

In the above example, the SAS Code node modifies the data produced by the Text Import node so that it can be merged with claims data indexed by the variable **ClaimNo**.



Using the Text Import Node

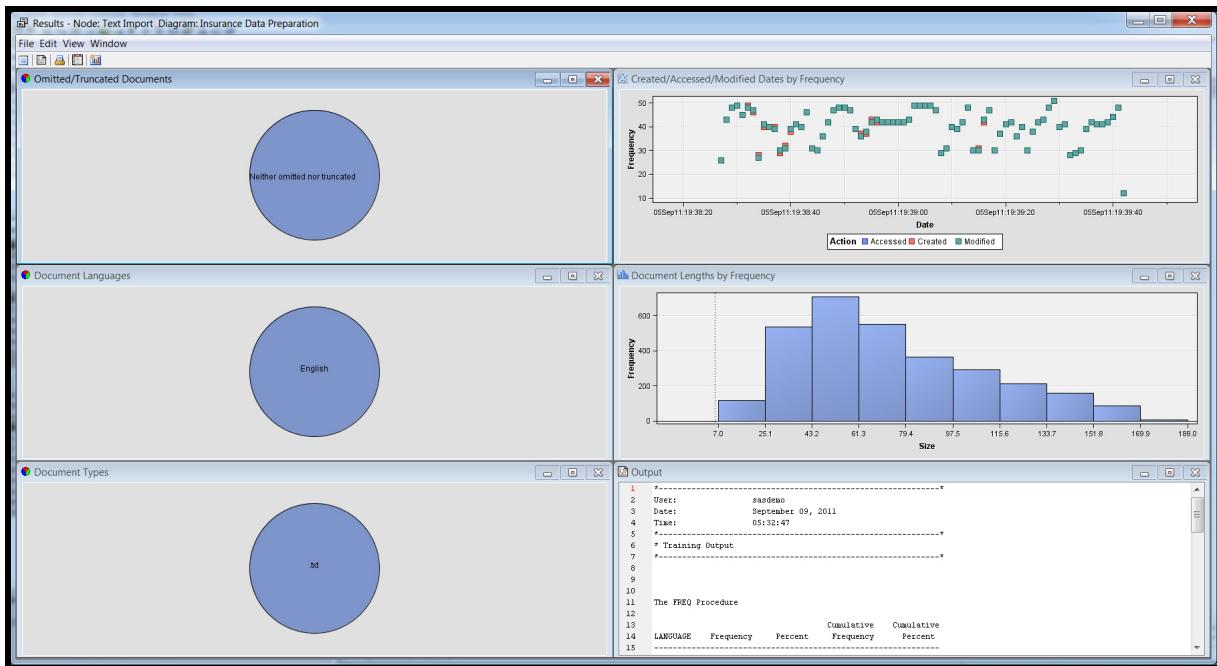
This demonstration illustrates how to use the Text Import node to read the insurance adjuster notes.

The success of this demonstration depends on correctly identifying the location of the adjuster notes source files. The source text (.txt) files are stored in a folder named **InsClaimsNotes**. This folder resides in the course folder. In courses with a virtual lab, the path for the course folder is usually named **D:\workshop\dmtx51**, but variations are possible, such as **S:\workshop\dmtx51**, **C:\workshop\winsas\dmtx51**, **C:\workshop\dmtxt**, **C:\workshop\dmtx51**, or **D:\SAS_Education\dmtx51**. Your instructor will provide you with the correct pathname.

1. Create a diagram and name it Insurance Data Preparation.
2. Drag a Text Import node into the diagram. For the Import File Directory property, enter the pathname for the insurance adjuster notes as determined from the above instructions. The discussion that follows assumes that this path is **C:\workshop\dmtx51\InsClaimsNotes**. Specify a destination directory using the same course pathname, with the destination folder **InsClaimsNotesDest**. For the Text Size property, specify **32000**. An example of the completed Properties panel appears below.

.. Property	Value
General	
Node ID	TextImport
Imported Data	
Exported Data	
Notes	
Train	
Import File Directory	C:\workshop\dmtx51\InsClaimsNotes
Destination Directory	C:\workshop\dmtx51\InsClaimsNotesDest
Language	'English'
Extensions	
Text Size	32000
Web Crawl	
URL	
Depth	1
Domain	Restricted
User Name	
Password	

3. Run the Text Import node. View the results.



There were no omitted or truncated files. The Output window verifies that 3,037 documents were processed.

LANGUAGE	Frequency	Percent	Cumulative	Cumulative
			Frequency	Percent
English	3037	100.00	3037	100.00
<hr/>				
Table of TRUNCATED by OMITTED				
<hr/>				
TRUNCATED OMITTED				
<hr/>				
Frequency				
Percent				
Row Pct				
Col Pct 0 Total				
<hr/>				
0	3037	3037		
100.00		100.00		
100.00				
100.00				
<hr/>	<hr/>	<hr/>	<hr/>	<hr/>
Total 3037 3037				
100.00 100.00				

2.03 Multiple Answer Poll

Which of the following tasks can be performed by the Text Import node?

- a. perform optical character recognition (OCR) of embedded bitmaps in document files
- b. convert Microsoft Word, Excel, and PowerPoint files to ASCII text
- c. process documents having more than 32,000 characters
- d. act as a Web crawler or robot to fetch and convert Internet pages to ASCII text files

2.2 Forensic Linguistics

Objectives

- Define stylometry and explain how it relates to text analytics.
- Illustrate how text mining can be used to support forensic linguistics using stylometry techniques.

23

Stylometry

Stylometry is defined as the use of linguistic style to characterize written language.

Applications:

- Attributing authorship of anonymous or disputed literary works
- Detecting plagiarism
- Forensic linguistics, for example, identifying Theodore Kaczynski as the Unabomber

24

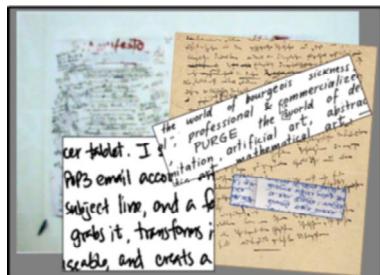
2.04 Poll

Do you anticipate encountering stylometry problems?

- Yes
- No

26

Forensic Linguistics



Special Case:
Stylometry Applied
to Forensics

Problem: Eleven written sources... Who is the author?

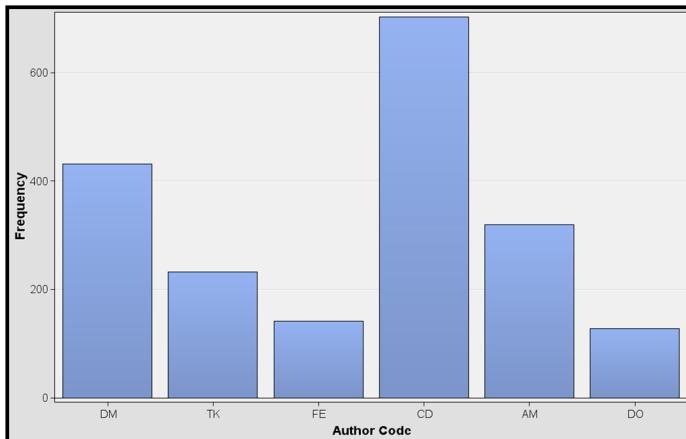
27

continued...

Forensic linguistics typically employs predictive modeling to score a document. The score represents an estimate of the probability that the document was written by a suspect. The value of text mining applied to forensic linguistics is that suspects can be identified for investigation. The text mining results are rarely if ever used as evidence in prosecuting a suspect, although testimony might include a discussion of techniques in describing how the suspect was identified.

While forensic linguistics provides a good example of text mining predictive modeling, the primary topic of a later chapter, it is included in this overview chapter as a popular application of text mining.

Forensic Linguistics



Corpus: 1,958 paragraphs from six authors taken from written works and interviews (Data source: Document Extracts)

28

continued...

The data for this study is real, but the situation is hypothetical. Separation was promoted for educational purposes. In actual forensic linguistic studies, there are rarely such pure results as those achieved here.

The six authors are coded as AM, CD, DM, DO, FE, and TK. The actual author names remain anonymous, except for author TK. TK is Theodore Kaczynski, the so-called Unabomber. The TK documents are paragraphs from the manifesto written by Kaczynski and published in *The New York Times*. The 11 unknown documents are excerpts from interviews with Kaczynski after he was convicted of murder. Thus, while based on real data, this example is artificial.

Forensic Linguistics

Obs #	Pa...	Extracted Text
16001		I read Edward Abbey in mid-eighties and that was one of the things that gave me the idea that,...
26002		Back in the sixties there had been some critiques of technology, but as far as I knew there wer...
36003		The honest truth is that I am not really politically oriented. I would have really rather just be livin...
46004		Unquestionably there is no doubt that the reason I dropped out of the technological system is ...
56005		Many years ago I used to read books like, for example, Ernest Thompson Seton's "Lives of Ga...
66006		I have quite a bit of experience identifying wild edible plants, it's certainly one of the most fulfilli...
76007		One thing I found when living in the woods was that you get so that you don't worry about the fu...
86008		The best place, to me, was the largest remnant of this plateau that dates from the tertiary age. ...
96009		I don't think it can be done. In part because of the human tendency, for most people, there are ...
106010		The big problem is that people don't believe a revolution is possible, and it is not possible pre...
116011		While I was living in the woods I sort of invented some gods for myself. Not that I believed in th...

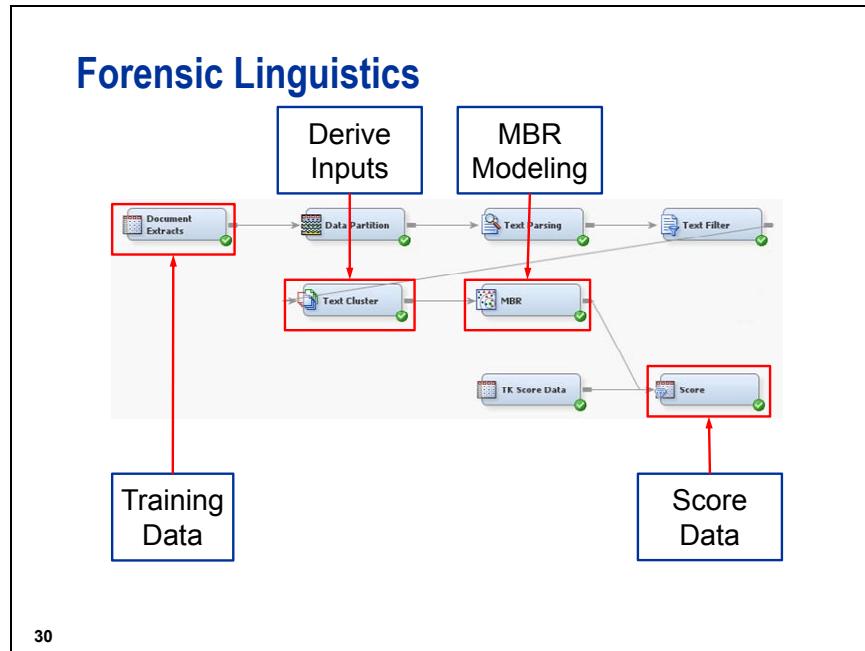
Score Data Set: Eleven documents from the same unknown author

Problem: Find documents in the corpus that are “closest” to the 11 documents and see whether there is evidence of a single author.

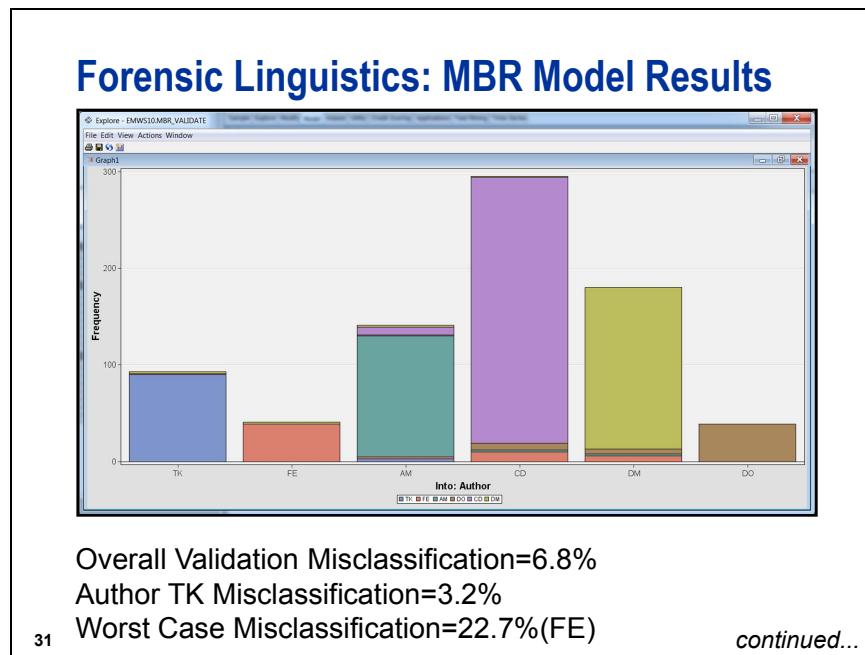
Methodology: Memory-Based Reasoning (MBR)

29

continued...

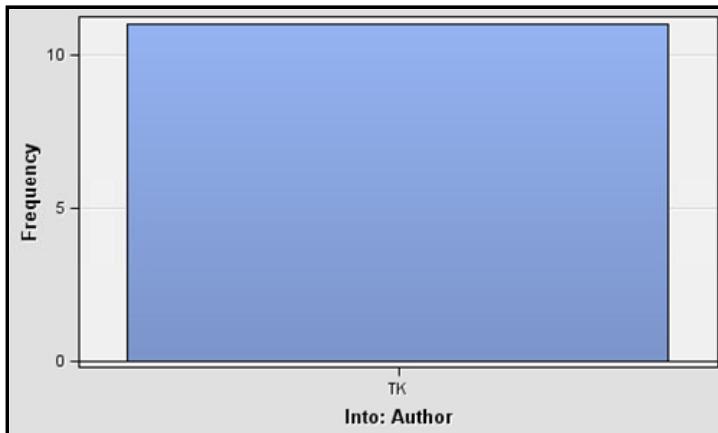


The Text Cluster node is used to derive inputs to be used by the MBR node for modeling. The Text Cluster node derives SVD variables as described in the sections on latent semantic analysis. These SVD variables can be used as inputs for a predictive model.



The misclassification for FE primarily results from crediting AM and CD with writing some of the FE documents. All but three documents that were written by TK are classified as written by TK, yielding the 3.2% misclassification rate. However, three documents that were written by two other authors are classified as written by TK, so there is a small false positive rate.

Forensic Linguistics: MBR Model Results



All 11 unknown cases are classified as being authored by TK. The 11 documents most closely match the documents known to have been authored by TK.

32

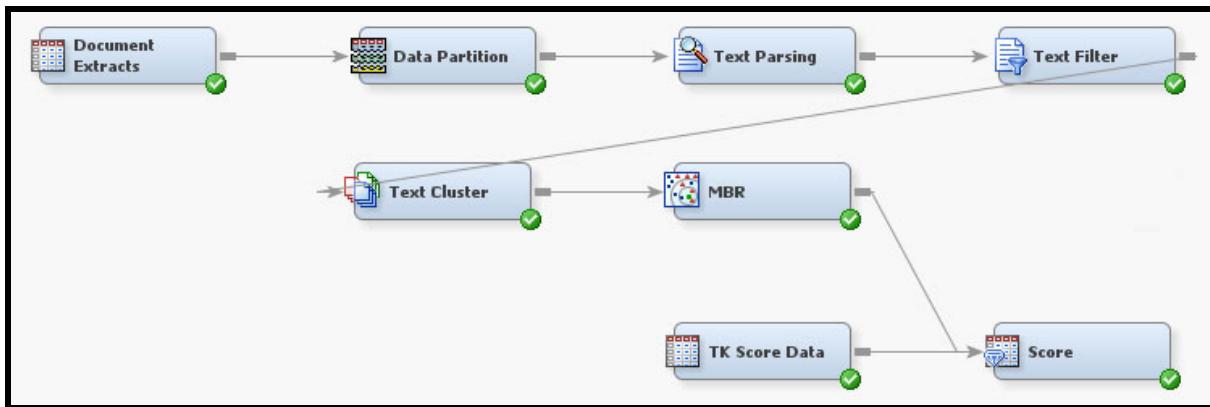
The Memory-Based Reasoning node is configured to find the eight closest documents to a given document. The eight closest documents are polled with respect to the target value (author), and the author receiving the plurality of votes is credited with having authored the given document. Thus, with six authors, the votes could be TK=3, AM=2, CD=2, DM=1, DO=0, FE=0. Consequently, winning the vote does not necessarily translate to high confidence.



Stylometry for Forensic Linguistics

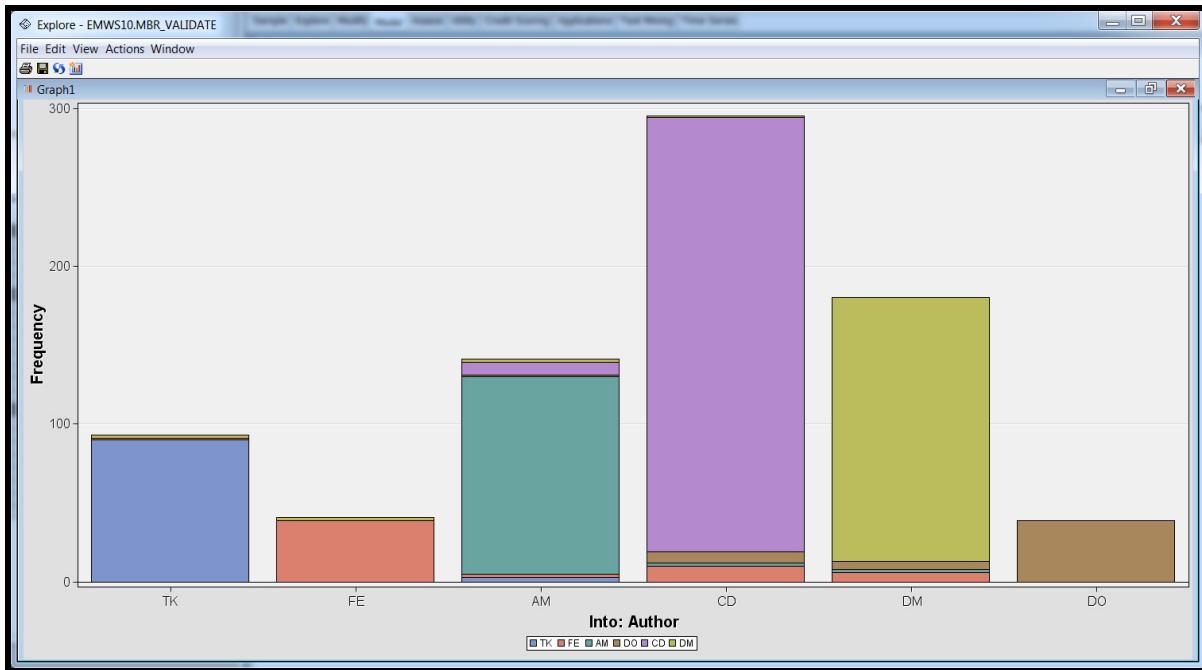
This demonstration illustrates how to use the Text Cluster node and the MBR node in SAS Enterprise Miner to investigate document sources in a forensic setting.

1. If necessary, create data sources for the Document Extracts (**docextracts**) data and for the TK score (**tkscore**) data. Make sure that the TK score data has a role of Score. The variable **Author** in the Document Extracts must have a role of Target.
2. Create a diagram named Forensic Linguistics.
3. Construct the following flow diagram:

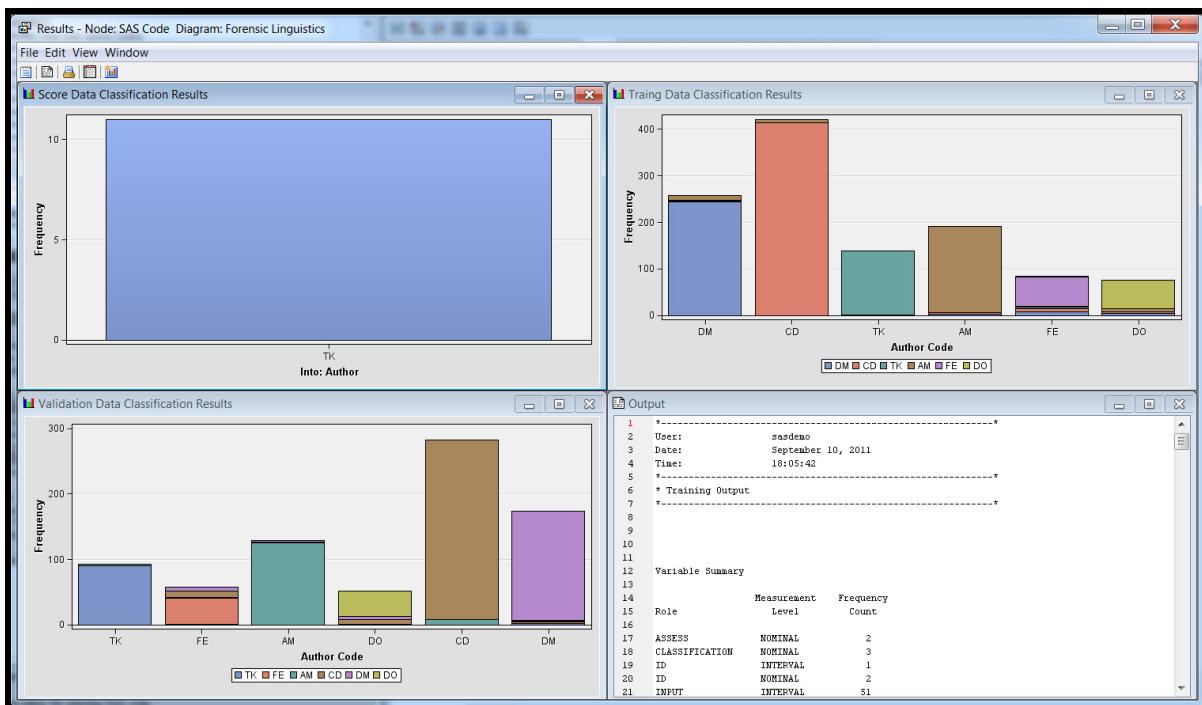


4. For the Data Partition node, specify a Training/Validation/Test split of **60/40/0**.
5. For the Text Parsing node, select **DMTXT.DOCEXTSTOP** for the stop list. Choose no data set to be specified for the synonyms table.
6. For the Text Filter node, choose the Term Weight property **Inverse Document Frequency**.
7. For the Text Cluster node, set the Exact or Maximum Number property to **Exact**. Set the Number of Clusters property to **6**.
8. For the Memory-Based Reasoning node, set the Number of Neighbors property to **8**. Keep all other properties set to the default values.
9. Score node properties remain at the default settings. Run the Score node.
10. Select the **Exported Data** property and select the validation data set. Use the plot wizard to plot the results as a bar chart. Select **I_Author** as the category variable and **Author** as the group variable.

The bar chart appears below.



11. Optionally, to automate the production of plots, you can attach a SAS Code node to the Score node. For the SAS Code node, select **Code Editor**. In the Code pane, right-click and select **Open**. Navigate to the course **sassrc** folder, and select the **SCN_ForensicLinguistics.sas** program. Save the code, and exit the Code Editor. Run the SAS Code node.
12. After the run completes, select **Results**.



The Score Data Classification Results plot shows that all 11 documents are classified as written by TK.

Details

The MBR (Memory-Based Reasoning) node uses a technique named *k-nearest neighbor*. The user supplies a value of *k*, and the node finds the *k* cases that are closest to a given case. The score assigned to the case is the result of polling or averaging as indicated above.

The Memory-Based Reasoning node documentation states: “The Memory-Based Reasoning node assumes that the variables with a model role of ‘input’ are numeric, orthogonal to each other, and standardized.” Because distance measures are employed, the use of standardized inputs prevents any one input variable from dominating the distance calculations. Otherwise, the input with the largest variation would tend to dominate distance calculations. The use of orthogonal inputs prevents a collection of orthogonal inputs from dominating the calculations over a set of correlated inputs.

You can use the Princomp node to derive orthogonal inputs. The SVD values coming from the Text Cluster node satisfy the input requirement for the MBR node.

The MBR node does not enforce the stated requirement for inputs. If the inputs fail to satisfy the requirements, then distance calculations will be biased in ways that might not be intended by the user.

2.05 Multiple Choice Poll

Which of the following statements is true?

- a. The demonstration proves that TK wrote the 11 documents.
- b. MBR works best when all inputs are uncorrelated with each other.
- c. MBR is fast enough to promote real-time scoring of new data.
- d. All of the above statements are true.

2.3 Information Retrieval

Objectives

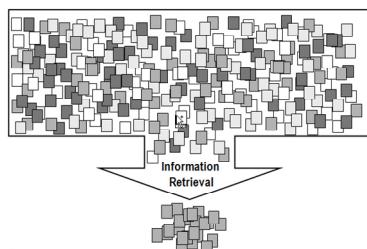
- Describe the science of information retrieval and explain how it relates to SAS Text Analytics.
- Use the Medline medical abstracts data to illustrate an application of information retrieval.

39

Information Retrieval

“Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).”

– Manning, Raghavan, and Schütze (2008)



40

One of the more publicized success stories in information retrieval concerns the discovery by Don Swanson (1988, 1991) that magnesium deficiency could be a source of migraine headaches. Swanson queried medical reports for articles about migraines and nutrition.

Information Retrieval Techniques

- Filtering and querying
- Boolean retrieval
- Indexed search
- Latent Semantic Indexing/Vector Space Model
- Naïve Bayes
- Others

41

For a given corpus of documents, IR groups documents based on the similarity of contents. An IR query can be a Boolean query, a query based on latent semantic indexing, or a query based on some other method of quantifying document content. The Text Filter node uses latent semantic indexing to measure the similarity between a document and the query. Documents that are most similar to the query are returned.

Filtering and Querying

Filtering and Querying Using the Interactive Filter Viewer

- Query operators control how filtering is performed.
- To clear a query, select **Clear** ⇒ **Apply**.
- You can close the Interactive Filter Viewer and save the current query. Rerunning the node exports the results of the query rather than the full data set.

42

continued...

Filtering and Querying

Review of Text Filter Query Operators

- **+term** returns all documents that have at least one occurrence of *term*.
- **-term** returns all documents that have zero occurrences of *term*.
- “*text string*” returns all documents that have at least one occurrence of the quoted text string.
- *string1*string2* returns all documents that have a term that begins with *string1*, ends with *string2*, and has text in between.
- **>#*term*** returns all documents that have *term* or any of the synonyms that were associated with *term*.

43

continued...

Recall that the Interactive Filter Viewer does not recognize **>#** operators mixed with the **+** operator.

Filtering and Querying

The screenshot shows the "Interactive Filter Viewer" application window. At the top, there's a menu bar with File, Edit, View, Window, and a toolbar with icons for Open, Save, Print, and others. Below the toolbar is a search bar with the text "+diabetes" and buttons for Apply and Clear. The main area is divided into two panes: "Documents" and "Terms".

Documents pane: It displays a list of search results. One result is highlighted in yellow with the text "...with type I diabetes and 199...". The columns include ABSTRACT, SNIPPET, RELEVANCE, AUTHOR, INDEX, MEDLINEID, and PES.

Terms pane: This pane shows a table of terms with their frequency (# DOCS), weight, role, and attribute. A yellow box highlights the term "diabetes" with the value 124. A black box contains the query "Query: +diabetes".

TERM	FREQ	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
diabetes	124	63	<input checked="" type="checkbox"/>	0.1	Noun	Alpha
increase	48	30	<input checked="" type="checkbox"/>	0.228	Verb	Alpha
less	78	28	<input checked="" type="checkbox"/>	0.266	Adv	Alpha
control	58	27	<input checked="" type="checkbox"/>	0.276	Noun	Alpha
diabetic	51	27	<input checked="" type="checkbox"/>	0.282	Adj	Alpha
less	51	25	<input checked="" type="checkbox"/>	0.282	Noun	Alpha
isopropyl	27	23	<input checked="" type="checkbox"/>	0.275	Verb	Alpha
group	52	22	<input checked="" type="checkbox"/>	0.373	Noun	Alpha
significantly	30	21	<input checked="" type="checkbox"/>	0.304	Adv	Alpha
level	32	19	<input checked="" type="checkbox"/>	0.334	Noun	Alpha
diabetic	31	19	<input checked="" type="checkbox"/>	0.341	Noun	Alpha
high	21	18	<input checked="" type="checkbox"/>	0.332	Adj	Alpha
mg	50	18	<input checked="" type="checkbox"/>	0.363	Noun	Alpha
compare	22	18	<input checked="" type="checkbox"/>	0.34	Verb	Alpha

44

Boolean Retrieval

- A query can be implemented as a Boolean retrieval.
- A query of the form +word +word –word –word... can be expressed as a binary sequence.
- The binary sequence can be compared to a corresponding sequence for each document.
- When a match occurs, the query is satisfied.
- Exact matches might not be required.
- Indexing of Boolean sequences can speed searches.

(An example of a Boolean search was presented in the last chapter.)

45

Latent Semantic Indexing (LSI)

- LSI is a natural extension of a Boolean retrieval.
- Weighting schemes replace 0-1 Boolean weights.
- LSI is usually implemented using the Singular Value Decomposition (SVD).
- LSI is embedded in the Text Miner nodes through application of the SVD.

46



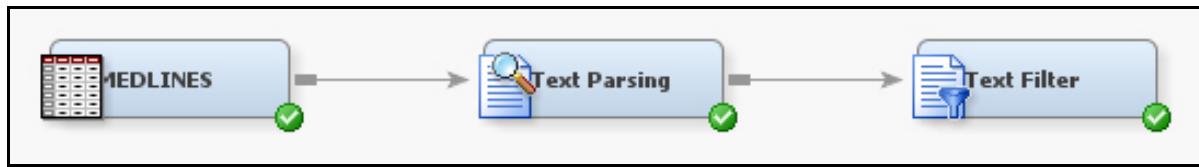
Retrieving Medical Information

This demonstration illustrates the use of the Text Filter node to retrieve medical information from the Medline data.

The **MEDLINES** data source contains a sample of 4,000 abstracts from medical research papers stored in the MEDLINE data repository.

1. Create a new diagram and name it Medline Information Retrieval. Drag the **MEDLINES** data source into the diagram.
2. Attach a Text Parsing node to the input data source node. Set the Synonyms property to **No data set to be specified**. Set the Stop List property to **DMTXT.MEDLINES_STOP**, which is a stop list that was specifically derived for the **MEDLINES** data set. Run the Text Parsing node.
3. Attach a Text Filter node to the Text Parsing node. Set the Maximum Number of Documents property to **5**. Because there is no target variable, the default Term Weight property is **Entropy**. A later chapter suggests that inverse document frequency might be more appropriate for this type of document collection, but for this demonstration, the default term weight is used.

The flow diagram follows.



4. Select the **Filter Viewer** property. This accesses the Interactive Filter Viewer.
5. In the Terms window, click the **TERM** column heading. This sorts the term table by the **TERM** value.
6. Right-click any term in the term table. Select **Find**.

The screenshot shows the 'Interactive Filter Viewer' application window. At the top, there's a menu bar with File, Edit, View, Window. Below the menu is a toolbar with a search icon, a search input field containing 'Search :', and buttons for Apply and Clear. The main area has two panes. The left pane is titled 'Documents' and contains a table with columns: ABSTRACT, AUTHOR, INDEX, MEDLI..., MESHT..., PUBTYPE, SOURCE, and TITLE. The right pane is titled 'Terms' and contains a table with columns: TERM ▲, FREQ, # DOCS, KEEP, WEIGHT, ROLE, and ATTRIBUTE. A context menu is open over the 'Find' entry in the 'Terms' table, listing options: Add Term to Search Expression, Treat as Synonyms, Remove Synonyms, Toggle KEEP, View Concept Links, Find, Repeat Find, Clear Selection, and Print... .

7. Type **diabetes** as the term to find.

The table jumps to the portion of the table that contains the term *diabetes*.

The screenshot shows the Interactive Filter Viewer interface. The top menu bar includes File, Edit, View, and Window. Below the menu is a toolbar with a magnifying glass icon, a search input field containing 'Search :', and buttons for Apply and Clear. The main area has two tabs: 'Documents' and 'Terms'. The 'Documents' tab displays a table with columns: ABSTRACT, AUTHOR, INDEX, MEDLI..., MESHT..., PUBTYPE, SOURCE, and TITLE. The table lists various medical articles. The 'Terms' tab displays a table with columns: TERM ▲, FREQ, # DOCS, KEEP, WEIGHT, ROLE, and ATTRIBUTE. The table lists terms such as 'dh', 'dha', 'dhaka', etc., with 'diabetes' highlighted in blue at the bottom. The 'diabetes' row has a checked 'KEEP' checkbox and a '0.537 Noun' weight.

8. Right-click **diabetes** and select **Add Term to Search Expression**.

This screenshot shows the same Interactive Filter Viewer interface as above, but with a context menu open over the 'diabetes' entry in the 'Terms' table. The menu options are: Add Term to Search Expression (highlighted in blue), Treat as Synonyms, Remove Synonyms, Toggle KEEP, View Concept Links, Find, Repeat Find, Clear Selection, Print..., and Print... (disabled). The 'diabetes' entry in the table is also highlighted in blue. The rest of the table and the 'Documents' pane are visible in the background.

The term *diabetes* is added to the Search window.



- Click **Apply**. The following results appear:

 A screenshot of the "Interactive Filter Viewer" application. At the top, there is a search bar with "diabetes", an "Apply" button, and a "Clear" button. Below the search bar is a "Documents" section containing a table with columns: ABSTRACT, TEXTFILTER_SNIPPET, TEXTFILTER_RELEVANCE, AUTHOR, INDEX, MEDLI..., and MESHT.. The table lists several medical abstracts. Below this is a "Terms" section containing a table with columns: TERM ▲, FREQ, # DOCS, KEEP, WEIGHT, ROLE, and ATTRIBUTE. This table shows the frequency of terms like "diabetes" (124), its role as a noun, and its weight (0.537).

TERM ▲	FREQ	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
di	1	1	<input type="checkbox"/>	0.0	Noun	Alpha
diabetes	124	63	<input checked="" type="checkbox"/>	0.537	Noun	Alpha
diabetes	2	2	<input type="checkbox"/>	0.0	Prop	Alpha
diabetes clinic	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
diabetes durati...	3	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
diabetes mana...	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
diabetes self-c...	1	1	<input type="checkbox"/>	0.0	Noun Group	Mixed
diabetes-40	1	1	<input type="checkbox"/>	0.0	Noun	Mixed
diabetes-induce	2	2	<input type="checkbox"/>	0.0	Verb	Mixed
diabetes-prone	1	1	<input type="checkbox"/>	0.0	Adj	Mixed
diabetic	88	48	<input checked="" type="checkbox"/>	0.554	Adj	Alpha
diabetic	57	34	<input checked="" type="checkbox"/>	0.598	Noun	Alpha
diabetic animal	8	5	<input checked="" type="checkbox"/>	0.812	Noun Group	Alpha

A total of 63 documents have the term *diabetes*, yet a total of 71 documents are returned by the filter. If you close the viewer and save the results, you can examine the 71 documents. Select the **Exported Data** property, and select the **Train** data set. Click the **Explore** button.

The screenshot shows three windows from the SAS Enterprise Miner interface:

- Sample Properties**: A table showing sample properties like Rows (11), Library (EMWS11), and Type (VIEW).
- Sample Statistics**: A table showing descriptive statistics for variables like ABSTRACT, AUTHOR, and TITLE.
- EMWS11.TextFilter_TRAIN**: A list of 63 documents, each with Obs #, INDEX, MEDLINEID, and SOURCE columns. The SOURCE column contains detailed medical abstracts.

More than 63 documents were selected because a latent semantic indexing search was employed. A relevance score is assigned to each document representing the similarity of the document to the query. Documents with relevance scores above an appropriate cutoff value are returned. This means that a document that does not contain the term in the search query can still be selected. The eight documents that are returned but do not have the term *diabetes* likely contain terms that co-occur with the term *diabetes* in other documents.

10. Close the Explore window and return to the Properties panel for the Text Filter node. In the Search Expression property input window, type **+glucose**. You do not need to re-run the Text Filter node; just select the Filter Viewer property.

The following window appears:

The screenshot shows two windows of the Interactive Filter Viewer. The top window is titled 'Interactive Filter Viewer' and has a search bar containing '+glucose'. Below the search bar is a table titled 'Documents' with columns: ABSTRACT, TEXTFILTER_SNIPPET, TEXTFILTER_RELEVANCE, AUTHOR, INDEX, MEDLI.., and MESHT.. The table lists 93 documents. The bottom window is titled 'Terms' and shows a table with columns: TERM, FREQ, # DOCS, KEEP, WEIGHT, ROLE, and ATTRIBUTE. The term 'glucose' is highlighted with a checked 'KEEP' checkbox and a weight of 0.097.

TERM	FREQ	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
gluconeogenesis	10	3	<input type="checkbox"/>	0.0	Noun	Alpha
gluconeogenic	1	1	<input type="checkbox"/>	0.0	Noun	Alpha
gluconeogenic ...	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
+ glucose	176	77	<input checked="" type="checkbox"/>	0.097	Noun	Alpha
glucose antag...	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
glucose antimi...	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
glucose conce...	2	2	<input type="checkbox"/>	0.0	Noun	Alpha
+ glucose conce...	3	3	<input type="checkbox"/>	0.0	Noun Group	Alpha
glucose conce...	2	2	<input type="checkbox"/>	0.0	Noun	Alpha
glucose depriv...	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
glucose disposal	5	3	<input type="checkbox"/>	0.0	Noun Group	Alpha
glucose disposal	5	2	<input type="checkbox"/>	0.0	Noun	Alpha
+ glucose dispos...	4	1	<input type="checkbox"/>	0.0	Noun Group	Alpha

A total of 93 documents are returned, with only 77 actually having the term *glucose* in the document.

11. Close the Interactive Filter Viewer. Type the search expression **+glucose -diabetes**. Be sure to leave a space between **+glucose** and **-diabetes**. Select the **Filter Viewer** property.

The screenshot shows two windows of the Interactive Filter Viewer. The top window is titled 'Interactive Filter Viewer' and displays a search bar with the query '+glucose -diabetes'. Below the search bar is a table titled 'Documents' with columns: ABSTRACT, TEXTFILTER_SNIPPET, TEXTFILTER_RELEVANCE, AUTHOR, INDEX, MEDLI.., and MESHT.. The table lists 72 documents. The second window is titled 'Terms' and displays a table with columns: TERM▲, FREQ, # DOCS, KEEP, WEIGHT, ROLE, and ATTRIBUTE. The term 'glucose' is highlighted with a blue background and has a checked 'KEEP' checkbox.

TERM▲	FREQ	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
gluconeogenic	1	1	<input type="checkbox"/>	0.0	Noun	Alpha
gluconeogenic ...	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
glucose	126	60	<input checked="" type="checkbox"/>	0.085	Noun	Alpha
glucose antag...	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
glucose antim...	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
glucose conce...	2	2	<input type="checkbox"/>	0.0	Noun	Alpha
glucose conce...	2	2	<input type="checkbox"/>	0.0	Noun Group	Alpha
glucose depriv...	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
glucose disposal	3	2	<input type="checkbox"/>	0.0	Noun Group	Alpha
glucose disposa...	1	1	<input type="checkbox"/>	0.0	Noun	Alpha
glucose dispos...	0	0	<input type="checkbox"/>	0.0	Noun Group	Alpha
glucose effect	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha

A total of 72 documents are returned, with 60 having the term *glucose*. These documents are more likely to be related to studies about glucose that do not include the disease diabetes.

12. Close the Interactive Filter Viewer. Type the property **+magnesium** in the Search Expression input window. Select the **Filter Viewer** property. Sort the relevance column to show decreasing relevance.

The following results appear:

The screenshot shows the Interactive Filter Viewer interface. At the top, there's a toolbar with File, Edit, View, Window, and a search bar containing '+magnesium' with Apply and Clear buttons. Below the toolbar is a 'Documents' section with a table. The table has columns: ABSTRACT, TEXTFILTER_SNIPPET, TEXTFILTER_RELEVANCE ▼, AUTHOR, INDEX, MEDLI..., and MESHT... . The abstract column contains snippets from various documents related to magnesium. The relevance column shows values like 1.0, 0.667, etc. The author and index columns provide metadata for each document. Below this table is a 'Terms' section with a table. The terms table has columns: TERM ▲, FREQ, # DOCS, KEEP, WEIGHT, ROLE, and ATTRIBUTE. It lists various terms found in the documents, such as 'magnesium' (FREQ 14, # DOCS 10, WEIGHT 0.14, Role Noun, Attribute Alpha), 'magnet' (FREQ 0, # DOCS 0, WEIGHT 0.0, Role Noun, Attribute Alpha), and 'magnetic field' (FREQ 0, # DOCS 0, WEIGHT 0.0, Role Noun Group, Attribute Alpha). The 'KEEP' column contains checkboxes, many of which are checked.

TERM ▲	FREQ	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
mag	0	0	<input type="checkbox"/>	0.0	Noun	Alpha
magendie	0	0	<input type="checkbox"/>	0.0	Prop	Alpha
maggot	0	0	<input type="checkbox"/>	0.0	Noun	Alpha
magnesium	14	10	<input checked="" type="checkbox"/>	0.14	Noun	Alpha
magnesium ad...	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
magnesium a...	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
magnesium ratio	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
magnesium su...	4	2	<input type="checkbox"/>	0.0	Noun Group	Alpha
magnesium su...	2	2	<input type="checkbox"/>	0.0	Noun	Alpha
magnet	0	0	<input type="checkbox"/>	0.0	Noun	Alpha
magnetic	0	0	<input type="checkbox"/>	0.0	Adj	Alpha
[+] magnetic field	0	0	<input type="checkbox"/>	0.0	Noun Group	Alpha

According to the snippet column, the second and third most relevant documents (ignoring ties) do not actually contain the term *magnesium*.



Exercises

1. Finding a Text String in a Data Set

The data table stored as **DMTXT.MOVIEDATA** (label=Random Movie Synopses) contains 1,527 movie synopses randomly selected from movie descriptions, reviews, and summaries found on the Internet. Some movies might have multiple entries. A start list has been compiled using frequency filtering. The list is stored in **DMTXT.MOVIESTART**.

- a. Identify the name of an actor or actress of interest.
- b. Find all of the movies in the data set that have a synopsis that mentions the selected name.

Bonus: Has Brad Pitt ever portrayed a vampire in a movie?

2.4 Text Categorization

Objectives

- Describe the text categorization problem and explain how it relates to SAS Text Analytics.
- Use the ASRS data to illustrate an application of text categorization.

50

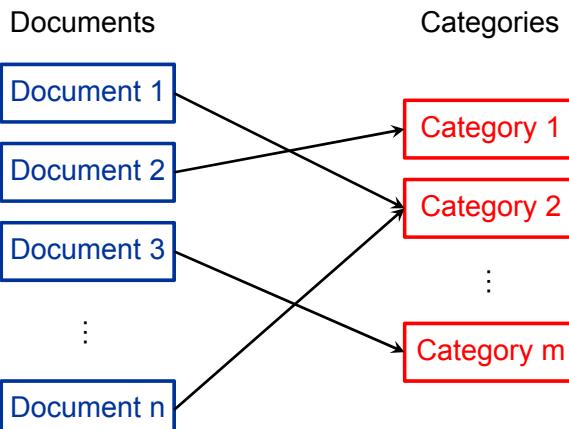
Text Categorization

Applications

- e-Mail filtering
- Call center routing
- News article classification
- Web page classification
- Classifying a technical library based on similar content

51

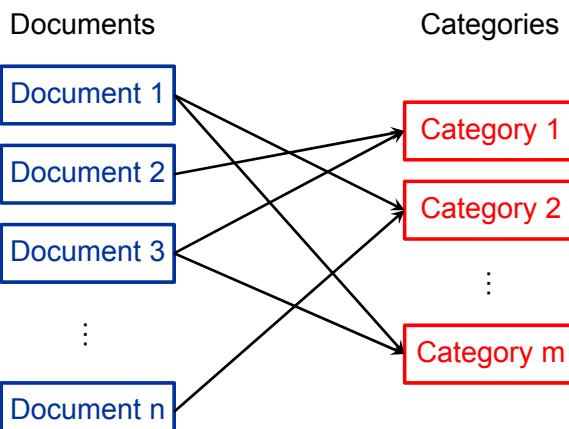
Text Categorization: Mutually Exclusive Categories



52

Text categorization can be used as an automated or semi-automated system for classifying documents. A common example describes a system for categorizing classified ads into categories such as Car Sales, Real Estate, and Employment Opportunities. Another common application involves classifying movies into genres such as Drama, Comedy, and Science Fiction. In the classified ads example, the categories must be mutually exclusive. In the movie genre example, a movie can fall into several categories.

Text Categorization: Multiple Categories per Document



53

Restricting topics to be mutually exclusive is often overly restrictive. You should expect larger documents to include multiple topics. The Text Topic node addresses this deficiency by allowing multiple topics per document and by enabling the user to refine the definitions of topics that are derived from the data.

Predictive modeling techniques for mutually exclusive categories would be restricted to those methods that can accommodate a nominal target variable, but most predictive modeling methods accommodate multiple category nominal targets. Plus, those methods that are designed to accommodate only binary targets can be easily extended to accommodate multiple categories. When a document can be assigned to multiple categories, then the predictive modeling problem just becomes a series of binary response problems. For each category, a predictive model produces a yes/no decision about the document belonging to the category.

Text Categorization

- Unsupervised
 - a special case of the pattern discovery classification problem
 - might be a first step for deriving labels that can then be used for supervised categorization
- Supervised
 - class labels assigned by domain experts
 - supervised classification
 - Classes can be mutually exclusive.
 - One document can belong to two or more classes.
 - models that are trained to automatically assign class labels to new documents

54

 By definition, text categorization is strictly a supervised learning problem because the categories are defined in advance to form a categorical target variable. The use of unsupervised learning techniques can be used to construct a categorization system when grouping documents can be useful, for example, in a division of labor where documents must be manually processed. Unsupervised learning techniques are presented as a tool to support the larger supervised learning problem – for example, when documents have not been classified into categories and no target variable exists in the data. Chapter 5 addresses text categorization as a strictly predictive modeling problem.

A fully automated text categorization system provides a yes/no decision about document membership in a category. These systems are called *hard categorization systems*. Semiautomatic systems provide a list of the top scoring categories for each document to facilitate a human judge making the final decision about category membership. A predictive model assigns a posterior probability of membership to each category for a given document, and the system sorts these probabilities in descending order, listing the top scorers.

Text Categorization: Unsupervised

Clustering/Profiling

- Categories are mutually exclusive.
- Interpreting clusters requires a methodology for assigning descriptive labels or keywords to clusters.
- The Text Cluster node uses a binomial probability formulation to assign descriptive terms.
- The Text Topic node provides term weighting capabilities through single-term topics, and the highest weighted terms can be used as descriptive terms if clusters are obtained outside of SAS Text Miner.

55

Unsupervised techniques can be employed even when a target variable is present. However, the usual setting for using techniques like clustering or neural network profiling is a text categorization problem where target values have not been stored with the text data. After clusters have been derived, and after class labels have been assigned, at least to preliminary test cases, then rules can be derived to try to associate cluster membership with category membership.

Evaluating Unsupervised Categorization

ID	Cluster1	Cluster2	Cluster3	Target1
1	1	0	0	1
2	0	0	0	0
3	0	1	0	1
4	0	1	0	1
5	0	0	0	0
6	0	1	0	1
7	0	0	0	0
8	0	1	1	1
9	0	0	1	0
10	1	1	0	1
11	0	1	0	1
12	0	0	0	0
13	0	0	0	0
14	0	0	1	0
15	0	0	0	0

Rule:
IF (Cluster1=1) AND (Cluster2=1)
THEN Target=1;

Rule Accuracy: 100%

56

Evaluating Unsupervised Categorization

ID	Cluster1	Cluster2	Cluster3	Target1
1	1	0	0	1
2	0	0	0	0
3	0	1	0	1
4	0	1	0	1
5	0	0	0	0
6	0	1	0	1
7	0	0	0	0
8	0	1	1	1
9	0	0	1	0
10	1	1	0	1
11	0	1	0	1
12	0	0	0	0
13	0	0	0	0
14	0	0	1	0
15	0	0	0	0

Rule:
IF (Cluster2=1) AND (Cluster3=1)
THEN Target=1;

Rule Accuracy: 86.7%

A decision tree can be used to try to deduce rules for assigning categories based on cluster membership. The concepts of precision and recall introduced in Chapter 1 are more relevant as assessment measures than overall percentage of correct classification, especially if a category has sparse membership.



Categorizing Reports in the ASRS

This demonstration illustrates how to use the Text Topics node to categorize reports from the Aviation Safety Reporting System (ASRS).



The ASRS can be accessed from the following link:

<http://asrs.arc.nasa.gov/>

From the Web site:

“ASRS captures confidential reports, analyzes the resulting aviation safety data, and disseminates vital information to the aviation community.”

“More than 850,000 reports have been submitted (through October, 2009) and no reporter’s identity has ever been breached by the ASRS. ASRS de-identifies reports before entering them into the incident database. All personal and organizational names are removed. Dates, times, and related information, which could be used to infer an identity, are either generalized or eliminated.”

As with other data sets used in this course, data sets derived from ASRS have been modified. The original data for this demonstration was extracted from the ASRS, pre-processed, and provided to competitors in a text mining competition sponsored by SIAM and the NASA Ames Research Center. The competition results were presented at the Seventh SIAM International Conference on Data Mining held in 2007 in Minneapolis, Minnesota. Participants were prohibited from using the R language, SAS software, and most commercial software. A link that provides access to the original data follows.

<https://c3.ndc.nasa.gov/dashlink/resources/138/>

A single report in the ASRS database can be a composite derivation of two or more reports filed for the same incident. For example, one runway incursion incident can result in three reports: one from the pilot, one from the co-pilot, and one from an air traffic controller. An incident involving two or more aircraft can have reports filed from pilots of all aircraft involved, as well as from air traffic controllers. In both examples, there will be only one ASRS report, but that report will be prepared by NASA professionals based on all reports submitted.

Reports can be submitted by aviation professionals, such as pilots, flight attendants, and mechanics. Reports can also be submitted by non-professionals, such as private pilots.

A report in the ASRS database has many fields, with one field representing a primary narrative describing the incident. This primary narrative is stored in the **Text** variable. All of the other fields have been omitted to simplify the text mining component of the analysis. In practice, an automated labeling system would attempt to use all fields.

NASA manually assigns to each report 1 or more of 54 anomalies, 1 or more of 32 results, 1 or more of 16 contributing factors, and 1 or more of 17 primary problems. For example, the report might describe an event that was a “runway ground incursion” anomaly, with a “took evasive action” result, that was a “human factor” contributing factor, and a “human factor” primary problem. These fields are not available in the contest data. Instead, the contest data has 22 labels, with a value of 1 “if document i has label j”; otherwise, the label has a value of -1. Labels correspond to the topics identified by NASA to aid in the analysis of the reports. The labels are not defined in the competition. For the course data, the 22 labels are named **Target01**, **Target02**, ..., and **Target22**, and an original coding of (-1,1) has been changed to (0,1), with a code of 1 indicating the presence of the label in the document. A document can be associated with one or more labels.

There are two data sources for the ASRS: ASRS Training Data and ASRS Test Data. The training data contains columns indicating which of the 22 manually assigned labels relates to a given report. The test data represents new reports that are not classified. However, for evaluation purposes, classification of the 22 labels has been added. In the competition data, the labels were originally masked in the test data. The goal is to develop a system to automatically detect topics to avoid the time, cost, and error associated with manually labeling the reports.

Because 22 target variables representing topic categories reside in the ASRS data table, this problem can be approached as a predictive modeling problem. This approach is presented in Chapter 5. However, it is common when an organization is transitioning to the use of text analytics for document data sets to be missing the class labels that define category membership. The premise for this demonstration includes the following main assumptions:

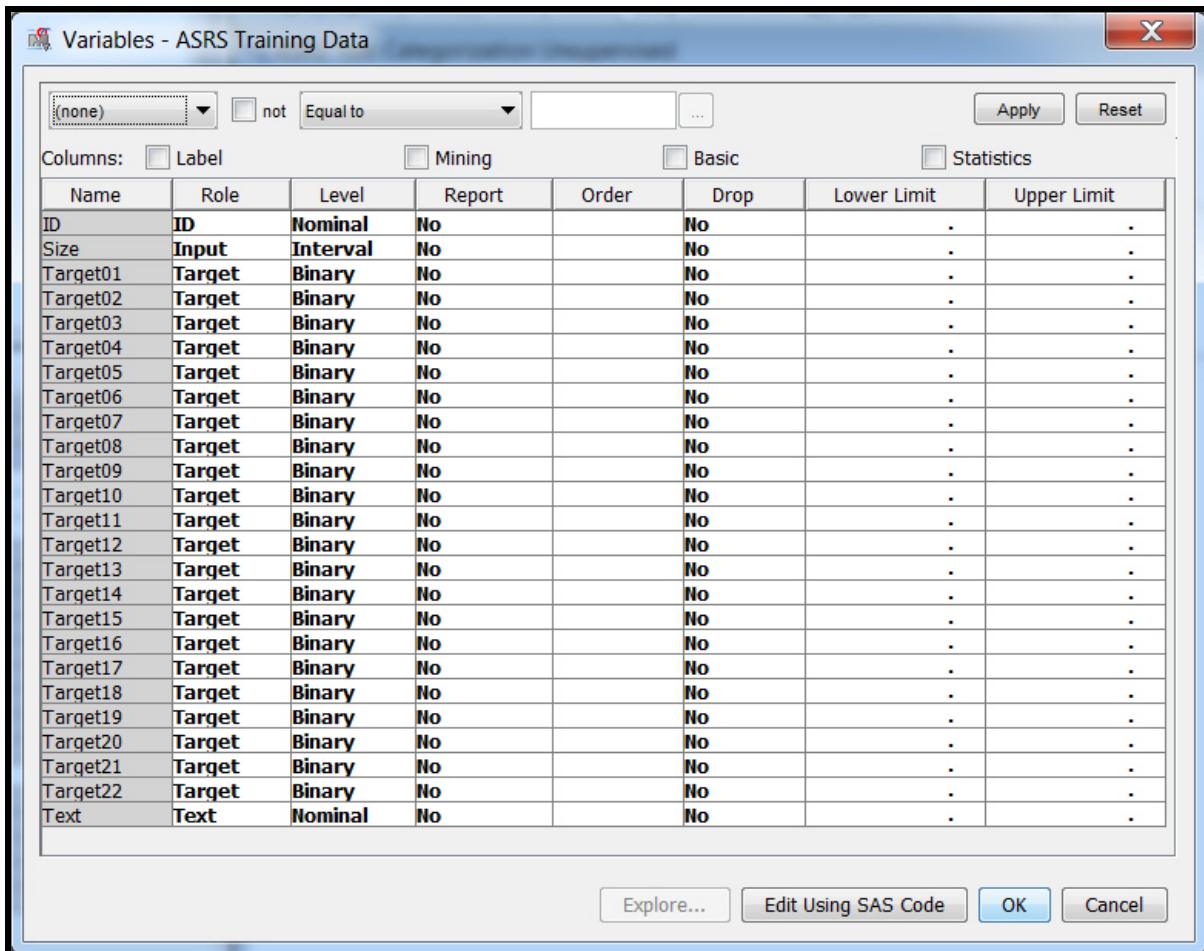
1. In the initial text categorization step, no class labels are available.
2. In the assessment step, experts have assigned class labels to permit evaluation of the text categorization step.

Even when a target variable is available, clustering can be a useful tool for a preliminary analysis to see whether documents naturally separate into predefined categories.

The class labels have been left in the training data for convenience. They will be exported to the last node in the process flow for use in assessment.

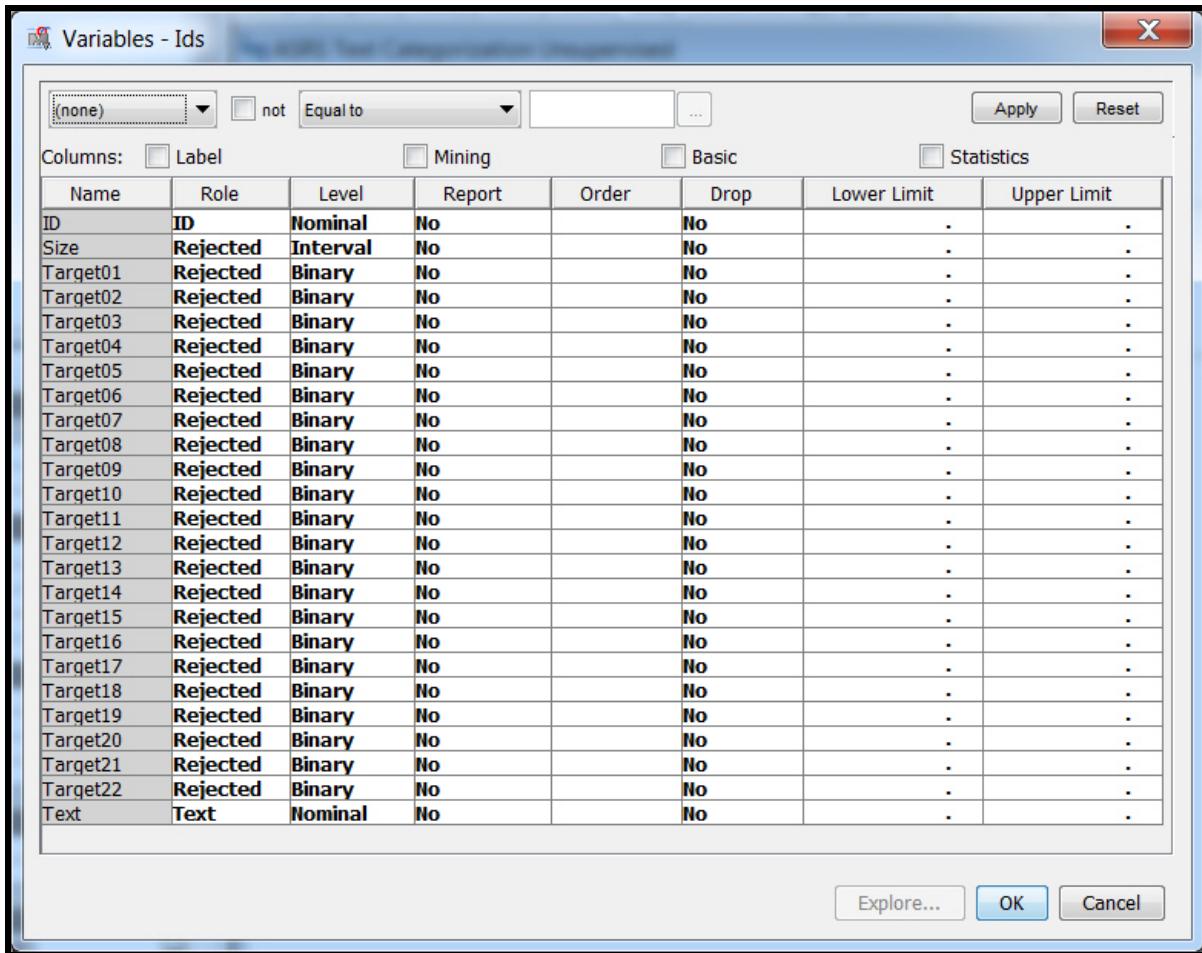
1. Create a new diagram and name it ASRS Text Categorization Unsupervised.

2. Create a data source for the ASRS Training data, **DMTXT.ASRS_TRAINING**. Use the following metadata:



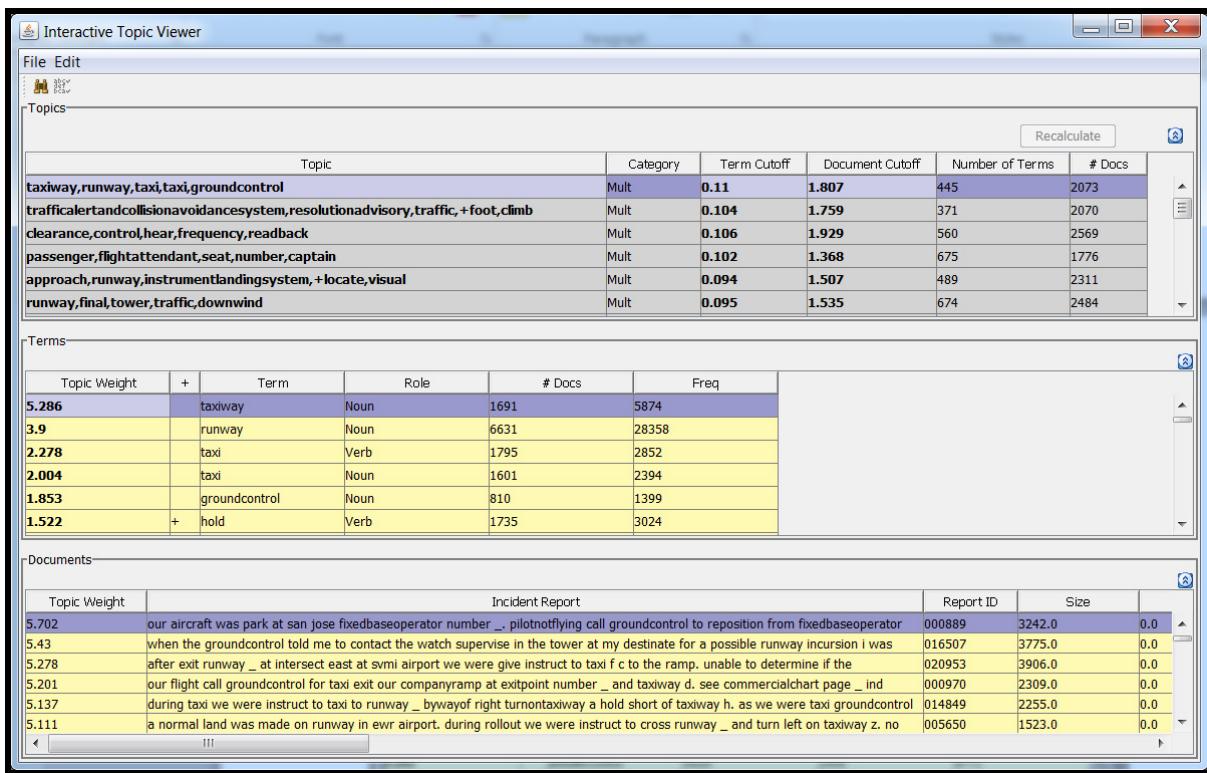
The metadata definitions above will permit use of the data for predictive modeling in Chapter 5.

3. Drag the **ASRS** data source into the diagram. Change the metadata to be consistent with the unsupervised learning scenario established above.



4. To investigate the robustness of the automated assignment, add a Data Partition node to the partition data set **DMTXT.ASRS_TRAINING** into training and validation data sets. Use a 75/25/0 partition.
5. Attach a Text Parsing node to the Data Partition node. A stop list for the ASRS data was derived based on frequency filtering. Terms in fewer than 10 documents as well as the terms with the highest frequencies were added to the default stop list to produce **DMTXT.ASRS_STOP**. This stop list has 12,222 terms. Using such a large stop list can increase processing time. Specify this as the stop list in the Text Parsing node. Specify that no synonym data set will be used. Run the Text Parsing node.
6. Attach a Text Filter node to the Text Parsing node. Open the **Variables** property and verify that only a single text variable will be used – that is, ensure that no target variables have inadvertently been left in the metadata. Set the Term Weight property to **Inverse Document Frequency**. Run the Text Filter node.
7. Attach a Text Topic node to the Text Filter node. Use the default properties for the Text Topic node. Run the Text Topic node.

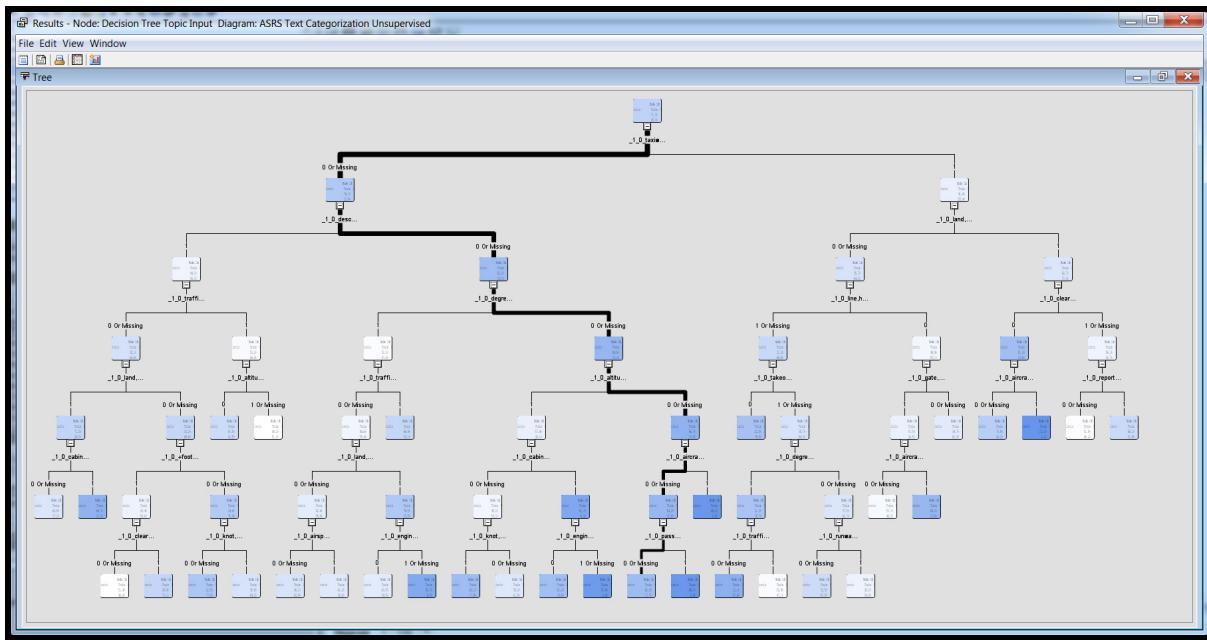
8. Open the Interactive Topic Viewer.



The first topic assigns large weights to the terms *taxiway* and *runway*. Domain expertise helps evaluate the topic categories. There are several topics that assign a large weight to runway, including topic **taxiway, runway, taxi, taxi, groundcontrol**, and topic **takeoff, runway, tower, depart, knot**. The Text Topic node appears to have some success in identifying ground incidents related to runways (for example, runway incursions) and approach to landing incidents, such as instructions to abort a landing on a runway. However, there remains a substantial amount of overlap between topics.

9. Attach a Metadata node to the Text Topic node. Change the role of **TextTarget06** to **Target**, and change the role of the 25 binary TextTopic variables to **Input**. Change the roll of the 25 interval TextTopic raw variables to **Rejected**.
10. Attach a Decision Tree node to the Text Topic node. Change the Assessment Measure property to **Average Square Error**. Run the Decision Tree node.

11. View the results.



The tree has 32 leaves. The Fit Statistics table follows.

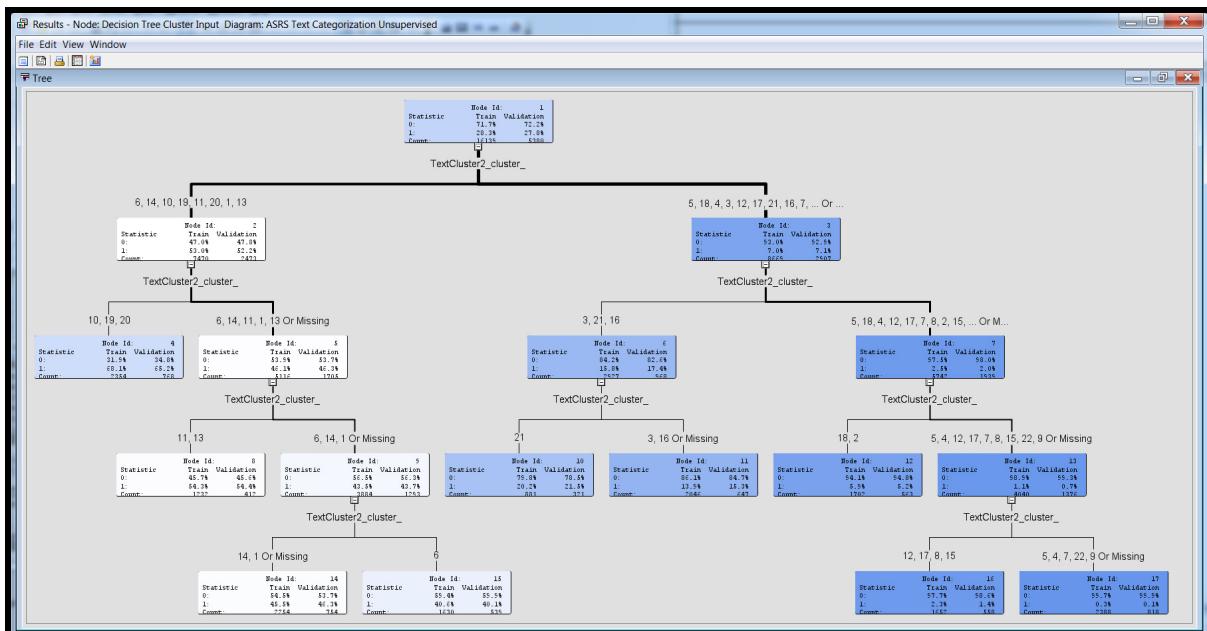
Target	Fit Statistics	Statistics Label	Train	Validation
Target06	<u>NOBS_</u>	Sum of Frequencies	16139	5380
Target06	<u>MISC_</u>	Misclassification Rate	0.212467	0.210037
Target06	<u>MAX_</u>	Maximum Absolute Error	0.992126	0.992126
Target06	<u>SSE_</u>	Sum of Squared Errors	4791.202	1561.685
Target06	<u>ASE_</u>	Average Squared Error	0.148436	0.145138
Target06	<u>RASE_</u>	Root Average Squared Error	0.385273	0.38097
Target06	<u>DIV_</u>	Divisor for ASE	32278	10760
Target06	<u>DFT_</u>	Total Degrees of Freedom	16139	.

The validation misclassification rate is 21.0%, representing an accuracy estimate of 79.0%.

The Variable Importance table indicates which topics contribute to the rules that try to reproduce **TextTarget06**.

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance
TextTopic_15	1_0 descend,+foot,restrict,+cross,arrive	1	1	1
TextTopic_1	1_0 taxiway,runway,taxi,taxi,groundcontrol	1	0.901634	0.8381
TextTopic_10	1_0 degree,head,turn,turn,depart	2	0.624489	0.599873
TextTopic_7	1_0 altitude,+autopilot,+foot,flightlevel,descend	2	0.475637	0.520521
TextTopic_14	1_0 land,gear,flap,extend,land	3	0.454111	0.45227
TextTopic_9	1_0 aircraft,install,inspect,maintain,+find	3	0.359833	0.356703
TextTopic_2	1_0 trafficalertandcollisionavoidancesystem,resolutionadvisory,traffic,+foot,climb	3	0.358987	0.430802
TextTopic_11	1_0 cabin,emergency,smoke,declare,flightattendant	2	0.347351	0.357726
TextTopic_4	1_0 passenger,flightattendant,seat,number,captain	1	0.306389	0.316009
TextTopic_25	1_0 knot,control,knot,flap,speed	2	0.232478	0.244831
TextTopic_3	1_0 clearance,control,hear,frequency,readback	2	0.179383	0.177061
TextTopic_19	1_0 +foot,altitude,approach,airport,descend	1	0.165648	0
TextTopic_16	1_0 line,hold,+hold,+short,hold short line	1	0.158521	0.169703
TextTopic_18	1_0 airspace,classb,classb airspace,visualflightrules,airport	1	0.150751	0.182659
TextTopic_20	1_0 takeoff,runway,tower,depart,knot	1	0.141027	0.181162
TextTopic_23	1_0 gate,aircraft,brake,ramp,captain	1	0.113812	0.109807
TextTopic_13	1_0 engine,number,fuel,emergency,declare	2	0.106978	0.140446
TextTopic_6	1_0 runway,final,tower,traffic,downwind	1	0.086708	0.104502
TextTopic_24	1_0 report,report state,state,+report,airplane	1	0.082536	0.146857

12. Attach a Text Cluster node to the Text Filter node. Set the Exact or Maximum Number property to **Exact**, and set the Number of Clusters property to **22**. Run the Text Cluster node.
13. Attach a Metadata node to the Text Cluster node. Every variable should have a role of Rejected, except **TextTarget06** should have a role of Target, and **TextCluster_cluster_** should have a role of Input.
14. Attach a Decision Tree node to the Text Cluster node. Change the Assessment Measure property to **Average Square Error**. Run the Decision Tree node. Open the Results window. The derived tree has nine leaves. The tree plot follows.



The Fit Statistics table follows.

Target	Fit Statistics	Statistics Label	Train	Validation
Target06	_NOBS_	Sum of Frequencies	16139	5380
Target06	_MISC_	Misclassification Rate	0.22362	0.227881
Target06	_MAX_	Maximum Absolute Error	0.997069	0.997069
Target06	_SSE_	Sum of Squared Errors	4588.89	1537.314
Target06	_ASE_	Average Squared Error	0.142168	0.142873
Target06	_RASE_	Root Average Squared Error	0.377051	0.377985
Target06	_DIV_	Divisor for ASE	32278	10760
Target06	_DFT_	Total Degrees of Freedom	16139	.

The misclassification rate is 22.8%, yielding an accuracy of 77.2%.

The English rules reveal some relationships between clusters and the target.

```
*-----*
  Node = 16
*-----*
if TextCluster2_cluster_ IS ONE OF: 12, 17, 8, 15
then
  Tree Node Identifier = 16
  Number of Observations = 1652
  Predicted: Target06=1 = 0.02
  Predicted: Target06=0 = 0.98

*-----*
  Node = 17
*-----*
if TextCluster2_cluster_ IS ONE OF: 5, 4, 7, 22, 9 or MISSING
then
  Tree Node Identifier = 17
  Number of Observations = 2388
  Predicted: Target06=1 = 0.00
  Predicted: Target06=0 = 1.00
```

A document assigned to any of clusters 4, 5, 7, 8, 9, 12, 15, 17, or 22 is very unlikely to fall into the target category.

```
*-----*
  Node = 4
*-----*
if TextCluster2_cluster_ IS ONE OF: 10, 19, 20
then
  Tree Node Identifier = 4
  Number of Observations = 2354
  Predicted: Target06=1 = 0.68
  Predicted: Target06=0 = 0.32
```

The purest clusters for the presence of **TextTarget06** are identified for leaf 4, but a document in a leaf 4 cluster only has a 68% chance of exhibiting the target category.

Unsupervised learning reveals some separation that appears to be related to **Target06**. However, a supervised approach is appropriate if the category labels are available. This problem is revisited in Chapter 5, using **TextTarget19**. A predictive model for **TextTarget06** that uses all SVD variables will produce superior accuracy. For example, using a neural network model, a misclassification rate of about 15% can be achieved.



Exercises

2. Categorizing Movies into Genres

The SAS table **DMTXT.MOVIESGENRE** has the same columns as **DMTXT.MOVIEDATA**. However, the **GENRE** field has been expanded to variables **GENRE1-GENRE5**, which contain individual genre names, and 10 binary target variables **ACTION, COMEDY, DOCUMENTARY, DRAMA, HORROR, KIDSFAMILY, MYSTERY, ROMANCE, SCIFI, and SUSPENSE**. The target variables indicate the presence or absence of the genre for each movie. For example, the movie *Absolute Power* has **GENRE="Action, Drama, Suspense"**, so **ACTION=1, DRAMA=1**, and **SUSPENSE=1**, while all other target variables have a value of 0 (zero). The program **RefineMovieData.sas** contains the code that created **DMTXT.MOVIESGENRE** from **DMTXT.MOVIEDATA**.

Use the following metadata to create the input data source for the movies data:

Data Source Wizard -- Step 5 of 8 Column Metadata

Name	Role	Level	Report	Order	Drop	Lower Limit
Action	Target	Binary	No		No	-
Comedy	Target	Binary	No		No	-
Documentary	Target	Binary	No		No	-
Drama	Target	Binary	No		No	-
Genre	Rejected	Nominal	No		No	-
Genre1	Rejected	Nominal	No		No	-
Genre2	Rejected	Nominal	No		No	-
Genre3	Rejected	Nominal	No		No	-
Genre4	Rejected	Nominal	No		No	-
Genre5	Rejected	Nominal	No		No	-
Horror	Target	Binary	No		No	-
KidsFamily	Target	Binary	No		No	-
MPAARating	Input	Nominal	No		No	-
Mystery	Target	Binary	No		No	-
NumGenres	Input	Nominal	No		No	-
Romance	Target	Binary	No		No	-
SciFi	Target	Binary	No		No	-
Size	Rejected	Interval	No		No	-
Suspense	Target	Binary	No		No	-
Synopsis	Text	Nominal	No		No	-
Title	ID	Nominal	No		No	-
ViewerRating	Input	Nominal	No		No	-
Year	Input	Interval	No		No	-

Buttons at the bottom include: Show code, Explore, Refresh Summary, < Back, Next >, and Cancel.

The following table shows all of the genres that are used. Only 10 were flagged for further analysis.

Genre	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Action	410	12.53	410	12.53
Animation	43	1.31	453	13.84
Arts	8	0.24	461	14.09
Christmas	8	0.24	469	14.33
Classic	93	2.84	562	17.18
Comedy	602	18.40	1164	35.57
Crime	5	0.15	1169	35.73
Cult	41	1.25	1210	36.98
Documentary	11	0.34	1221	37.32
Drama	718	21.94	1939	59.26
Erotica	15	0.46	1954	59.72
Family	58	1.77	2012	61.49
Finish	1	0.03	2013	61.52
Foreign	37	1.13	2050	62.65
Gay/Lesbian	29	0.89	2079	63.54
German	1	0.03	2080	63.57
Horror	84	2.57	2164	66.14
Independent	9	0.28	2173	66.41
Italian	1	0.03	2174	66.44
Kids	102	3.12	2276	69.56
Martial-Arts	16	0.49	2292	70.05
Music	53	1.62	2345	71.67
Mystery	40	1.22	2385	72.89
Noir	6	0.18	2391	73.07
Religion	4	0.12	2395	73.20
Romance	255	7.79	2650	80.99
Sci-Fi/Fantasy	183	5.59	2833	86.58
Spanish	1	0.03	2834	86.61
Sports	27	0.83	2861	87.44
Suspense	307	9.38	3168	96.82
Thriller	44	1.34	3212	98.17
War	40	1.22	3252	99.39
Western	20	0.61	3272	100.00

Because 33 genres are listed, you might attempt to reproduce all 33. However, notice that many reference a language, and quite a few are sparse. If you restrict attention to genres represented by 50 or more movies, then 11 genres are represented. The following table summarizes these 11 genres:

Obs	Genre	COUNT	PERCENT
2	Action	410	12.5306
6	Classic	93	2.8423
7	Comedy	602	18.3985
11	Drama	718	21.9438
13	Family	58	1.7726
18	Horror	84	2.5672
21	Kids	102	3.1174
23	Music	53	1.6198
27	Romance	255	7.7934
28	Sci-Fi/Fantasy	183	5.5929
31	Suspense	307	9.3826

To obtain the 10 target category variables, the genres **Kids** and **Family** were combined to form **KidsFamily**. Also, the genres **Classic** and **Music** were dropped in favor of adding **Documentary** and **Mystery**. For future analysis, a possible modification of the 10 chosen genres is to combine **Mystery** and **Suspense** into one target variable because both are relatively sparse.

- a. Create a data source for **DMTXT.MOVIESGENRE** using the metadata suggested above.
- b. You can use the same diagram as the previous exercise. Drag the data source into the diagram. Change the metadata by rejecting all of the target variables. (This will affect partitioning and the default choice of term weight.) Attach a **Data Partition** node to the data source node and use a 75/25 split. Attach the **Text Parsing**, **Text Filter**, and **Text Topic** nodes. Use the **DMTXT.MOVIESTART** start list. Select the Term Weight property **Inverse Document Frequency**. In the Text Topic node, change the property Number of Multi-term Topics to **10**. Compare these 10 topics to the 10 genres selected for analysis.
- c. Use a decision tree to help determine whether any of the topics could be used to define a set of rules for assigning a movie to the category **Comedy**. Hint: Attach a **Metadata** node to the **Text Topic** node, and change every variable to **Rejected**, except set the **Comedy** variable to **Target** and set all of the text topic binary variables to **Input**.

2.5 Chapter Summary

SAS provides many tools and products for accessing and processing data. The SAS language features a rich set of character functions for processing text. In addition, Perl regular expressions are supported.

Stylometry traditionally encompasses literary features such as those explained by Love (2002). However, the emergence of analytic techniques for text analysis seems to provide the strongest support for attributing authorship. Applications extend beyond attributing authorship to plagiarism detection and forensic linguistics.

Information retrieval (IR) methods are designed to access relevant information quickly. IR methods can be found in many locations of SAS Text Miner. In particular, the Interactive Filter Viewer in the Text Filter node supports queries to extract relevant documents.

Text categorization can be supervised or unsupervised. The Text Topic node derives topics for the document collection so that each document can have one or more topics. This permits categorization by multiple topics without the usual mutually exclusive categories required by document clustering. The Text Cluster node permits derivation of mutually exclusive categories.

For Additional Information

Love, Harold. 2002. *Attributing Authorship: An Introduction*. Cambridge, United Kingdom: Cambridge University Press.

Mosteller, F., and D.L. Wallace. 1964. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. New York: Springer.

SAS Institute Inc. 2010. Extending *SAS Enterprise Miner with User-Written Nodes*. Cary, North Carolina: SAS Institute Inc.

Swanson, Don R. 1988. "Migraine and Magnesium - 11 Neglected Connections." *Perspectives in Biology and Medicine*. 31 (4), pp. 526-557.

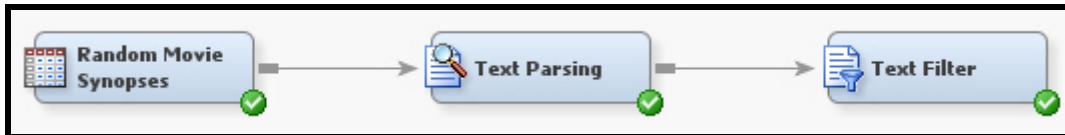
Swanson, Don R. 1991. "Complementary structures in disjoint science literatures." *SIGIR'91. Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*. Pp. 280-289.

2.6 Solutions

Solutions to Exercises

1. Finding a Text String in a Data Set

A possible flow diagram is given below.



Suppose you are interested in finding movies that star Sandra Bullock. There are at least two versions of filter operations to find synopses that mention Sandra Bullock.

Interactive Filter Viewer

File Edit View Window

Search : "Sandra Bullock" Apply Clear

Documents

SYNOPSIS	TEXTFILTER_SNIPPET	TEXTFILTER_RELEVANCE	TITLE	GENRE
The girl-next-door has grown up. Sandra Bullock takes a brave	... Sandra Bullock takes a	0.25	28 Days	Drama P
If you, like me, are a big fan of end-of-the-world science fiction stories	... DAYS , starring Sandra	0.125	28 Days Later	Horror, ... R
If you're psychologically messed up can't get a date, can't	... , Siddha (Sandra Bullock) ,	0.125	Divine Secrets of the Ya-Ya Sisterhood	Comedy,... P
Sandra Bullock shines in this frenetic road comedy/love story	... Sandra Bullock shines in	0.125	Forces of Nature	Comedy,... P
"Hope Floats" is a pretty good movie for a chick flick. Now, before you	... , thanks to Sandra Bullock	1.0	Hope Floats	Comedy,... P
Modern science, contrary to the thousands years of	(played by Sandra Bullock	0.375	Love Potion No. 9	Comedy P
Sandra Bullock, as FBI field agent Gracie Hart, is nothing short of	... Sandra Bullock , as FBI field	0.875	Miss Congeniality	Comedy P
Remember that kid back in high school, the one who was	... , Cassie (Sandra Bullock)	0.375	Murder By Numbers	Suspense R
Dabbling is dangerous. Sally and Gillian Owens, played charmingly by	... played charmingly by	0.125	Practical Magic	Comedy,... P
Set in ancient Egypt, this animated tale follows the Biblical	... his sister (Sandra Bullock)	0.125	Prince of Egypt	Animation... P
A TIME TO KILL is John Grisham's first novel, but the fourth one	... Ellen Roark (Sandra	0.25	Time to Kill, A	Drama, ... R
TWO IF BY SEA is a putative comedy reminiscent of the SMOKEY AND THE	... girlfriend Roz (Sandra	0.25	Two If By Sea	Comedy,... R
"Two Weeks Notice" is not only a conspiracy against the	... it depends on Sandra	0.5	Two Weeks Notice	Comedy,... R
1990s Hollywood gave the viewers plenty of reasons to be	(played by Sandra Bullock	0.25	Vanishing, The	Drama, ... R
WHILE YOU WERE SLEEPING is a gem of a movie. It tells the tale of	... named Lucy (Sandra	0.25	While You Were Sleeping	Comedy,... P

Terms

TERM	FREQ	# DOCS	KEEP ▾	WEIGHT	ROLE	ATTRIBUTE
sandra	24	15	✓	0.035	Prop	Alpha
bullock	30	14	✓	0.085	Noun	Alpha
keep	8	8	✓	0.232	Verb	Alpha
few	8	7	✓	0.296	Adj	Alpha
work	9	7	✓	0.302	Verb	Alpha
romantic	8	7	✓	0.296	Adj	Alpha
hand	8	6	✓	0.384	Noun	Alpha
girl	9	6	✓	0.381	Noun	Alpha
script	12	6	✓	0.387	Noun	Alpha
real	8	6	✓	0.384	Adj	Alpha
woman	12	6	✓	0.415	Noun	Alpha
comedy	11	6	✓	0.365	Noun	Alpha
problem	8	6	✓	0.36	Noun	Alpha
begin	6	6	✓	0.338	Verb	Alpha
small	7	5	✓	0.455	Adj	Alpha

Interactive Filter Viewer

File Edit View Window

Search : +sandra +bullock

Apply Clear

Documents

SYNOPSIS	TEXTFILTER_SNIPPET	TEXTFILTER_RELEVANCE	GENRE	TITLE	M
The girl-next-door has grown up. Sandra Bullock takes a brave	... Sandra Bullock takes a	0.427	Drama	28 Days	P
If you, like me, are a big fan of end-of-the-world science fiction stories	... DAYS , starring Sandra	0.174	Horror, ...	28 Days Later	R
If you're psychologically messed up can't get a date, can't	... , Sidda (Sandra Bullock) ,	0.174	Comedy,...	Divine Secrets of the Ya-Ya Sisterhood	P
Sandra Bullock shines in this frenetic road comedy/love story	... Sandra Bullock shines in	0.174	Comedy,...	Forces of Nature	P
"Hope Floats" is a pretty good movie for a chick flick. Now, before you	... , thanks to Sandra Bullock	1.0	Comedy,...	Hope Floats	P
Modern science, contrary to the thousands years of	... (played by Sandra Bullock	0.444	Comedy	Love Potion No. 9	P
Sandra Bullock, as FBI field agent Gracie Hart, is nothing short of	... Sandra Bullock , as FBI field	0.826	Comedy	Miss Congeniality	P
Remember that kid back in high school, the one who was	... , Cassie (Sandra Bullock)	0.444	Suspense	Murder By Numbers	R
Dabbling is dangerous. Sally and Gillian Owens, played charmingly by	... played charmingly by	0.174	Comedy,...	Practical Magic	P
Set in ancient Egypt, this animated tale follows the Biblical	... his sister (Sandra Bullock)	0.174	Animation	Prince of Egypt	P
A TIME TO KILL is John Grisham's first novel, but the fourth one	... Ellen Roark (Sandra	0.27	Drama, ...	Time to Kill, A	R
TWO IF BY SEA is a putative comedy reminiscent of the SMOKEY AND THE	... girlfriend Roz (Sandra	0.27	Comedy,...	Two If By Sea	R
"Two Weeks Notice" is not only a conspiracy against the	... it depends on Sandra	0.539	Comedy,...	Two Weeks Notice	P
1990s Hollywood gave the viewers plenty of reasons to be	... (played by Sandra Bullock	0.348	Drama, ...	Vanishing, The	R
WHILE YOU WERE SLEEPING is a gem of a movie. It tells the tale of	... named Lucy (Sandra	0.27	Comedy,...	While You Were Sleeping	P

Terms

TERM	FREQ	# DOCS	KEEP ▾	WEIGHT	ROLE	ATTRIBUTE
sandra	24	15	<input checked="" type="checkbox"/>	0.035	Prop	Alpha
bullock	30	14	<input checked="" type="checkbox"/>	0.085	Noun	Alpha
keep	8	8	<input checked="" type="checkbox"/>	0.232	Verb	Alpha
few	8	7	<input checked="" type="checkbox"/>	0.296	Adj	Alpha
work	9	7	<input checked="" type="checkbox"/>	0.302	Verb	Alpha
romantic	8	7	<input checked="" type="checkbox"/>	0.296	Adj	Alpha
hand	8	6	<input checked="" type="checkbox"/>	0.384	Noun	Alpha
girl	9	6	<input checked="" type="checkbox"/>	0.381	Noun	Alpha
script	12	6	<input checked="" type="checkbox"/>	0.387	Noun	Alpha
real	8	6	<input checked="" type="checkbox"/>	0.384	Adj	Alpha
woman	12	6	<input checked="" type="checkbox"/>	0.415	Noun	Alpha
comedy	11	6	<input checked="" type="checkbox"/>	0.365	Noun	Alpha
problem	8	6	<input checked="" type="checkbox"/>	0.36	Noun	Alpha
begin	6	6	<input checked="" type="checkbox"/>	0.338	Verb	Alpha
small	7	5	<input checked="" type="checkbox"/>	0.455	Adj	Alpha

Bonus: Brad Pitt played a vampire in the movie *Interview with the Vampire*, as the following query suggests. (The movie *Seven* is a false positive with relevance score less than half.)

Interactive Filter Viewer

File Edit View Window

Search : +brad +pitt +vampire

Apply Clear

Documents

SYNOPSIS	TEXTFILTER_SNIPPET	TEXTFILTER_RELEVANCE	GENRE	TITLE	V
INTERVIEW WITH THE VAMPIRE is the screen version of Anne Rice's	... INTERVIEW WITH THE	1.0	Horror, ...	Interview with the Vampire	
I used to avoid Brad Pitt movies like the plague, like famine,	... used to avoid Brad Pitt	0.449	Drama, ...	Seven	

Terms

TERM	FREQ	# DOCS	KEEP ▾	WEIGHT	ROLE	ATTRIBUTE
be	27	2	<input type="checkbox"/>	0.0	Verb	Alpha
no	2	2	<input type="checkbox"/>	0.0	Adv	Alpha
ago	2	2	<input type="checkbox"/>	0.0	Adv	Alpha
one	2	2	<input type="checkbox"/>	0.0	Num	Alpha
movie	11	2	<input type="checkbox"/>	0.0	Noun	Alpha
not	8	2	<input type="checkbox"/>	0.0	Adv	Alpha
boy	2	2	<input type="checkbox"/>	0.0	Noun	Alpha
imagine	2	2	<input type="checkbox"/>	0.0	Verb	Alpha
people	3	2	<input type="checkbox"/>	0.0	Noun	Alpha
make	3	2	<input type="checkbox"/>	0.0	Verb	Alpha
being	3	2	<input type="checkbox"/>	0.0	Noun	Alpha
give	3	2	<input type="checkbox"/>	0.0	Verb	Alpha
have	10	2	<input type="checkbox"/>	0.0	Verb	Alpha

2. Categorizing Movies into Genres

- The metadata table was given in the statement of the exercise.
- The following table summarizes the metadata for the **DMTXT.MOVIESGENRE** data source after it has been modified in the diagram.

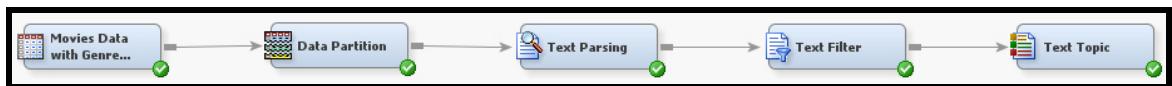
Variables - Ids2

Columns: Label Mining Basic Statistics

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Action	Rejected	Binary	No	No	No	.	.
Comedy	Rejected	Binary	No	No	No	.	.
Documentary	Rejected	Binary	No	No	No	.	.
Drama	Rejected	Binary	No	No	No	.	.
Genre	Rejected	Nominal	No	No	No	.	.
Genre1	Rejected	Nominal	No	No	No	.	.
Genre2	Rejected	Nominal	No	No	No	.	.
Genre3	Rejected	Nominal	No	No	No	.	.
Genre4	Rejected	Nominal	No	No	No	.	.
Genre5	Rejected	Nominal	No	No	No	.	.
Horror	Rejected	Binary	No	No	No	.	.
KidsFamily	Rejected	Binary	No	No	No	.	.
MPAARating	Input	Nominal	No	No	No	.	.
Mystery	Rejected	Binary	No	No	No	.	.
NumGenres	Input	Nominal	No	No	No	.	.
Romance	Rejected	Binary	No	No	No	.	.
SciFi	Rejected	Binary	No	No	No	.	.
Size	Rejected	Interval	No	No	No	.	.
Suspense	Rejected	Binary	No	No	No	.	.
Synopsis	Text	Nominal	No	No	No	.	.
Title	ID	Nominal	No	No	No	.	.
ViewerRating	Input	Nominal	No	No	No	.	.
Year	Input	Interval	No	No	No	.	.

Buttons: Apply, Reset, Explore..., OK, Cancel

The diagram for this analysis appears below.



The Interactive Topic Viewer displays the 10 derived topics.

The screenshot shows the 'Interactive Topic Viewer' application window with three main panels:

- Topics:** A table showing 10 derived topics with their corresponding Category, Term Cutoff, Document Cutoff, Number of Terms, and # Docs. The topics listed include '+show,+rate,+kid,+script,+recommend', '+hollywood,+hand,+protagonist,later,+order', '+viewer,+moment,+relationship,+minute,+love', '+woman,+mother,+school,+girl,+husband', '+comedy,+funny,+joke,humor,+laugh', '+effect,special,+action,+special effect,human', '+war,+battle,+president,+soldier,+history', '+cop,+crime,+thriller,police,+action', '+bond,bond,james,connery,jeffrey', and 'best,granger,gauge,+oscar,+love'. The 'Term Cutoff' column has values like 0.074, 0.064, etc., and the '# Docs' column has values like 166, 107, etc.
- Terms:** A table showing terms with their Topic Weight, Role, # Docs, and Freq. The terms listed include 'show' (Noun), 'rate' (Verb), 'kid' (Noun), 'script' (Noun), 'recommend' (Verb), 'sex' (Noun), 'violence' (Noun), 'teenager' (Noun), 'acting' (Noun), and 'nudity' (Noun). The 'Topic Weight' column has values like 0.792, 0.523, 0.411, etc., and the 'Freq' column has values like 724, 410, 450, etc.
- Documents:** A table showing document details such as Topic Weight, Synopsis, Action, Comedy, Documentary, Drama, Genre, Genre1, and Genre2. The documents listed include 'DEAD MAN', 'Remember being 11', 'THREE WISHES is a', 'A LITTLE PRINCESS', 'From the opening', 'PERSUASION is a', 'THE NEON BIBLE is', 'For someone whi is 0.0', 'THE SOUND OF', 'Two absolute', and 'One of Martin'. The 'Genre' column includes categories like Drama, Independent, Kids, Sci-Fi, Family, Comedy, Romance, Action, Classic, and Music.

There seems to be little relationship between the topics and the genres. For example, the term *kid* in the first topic might suggest Kids/Family, but examination of the Terms table reveals large weights for sex, violence, and nudity. It is possible these terms show up because of negative references, for example, "It is refreshing to see a movie that does not contain violence." The following table suggests possible associations:

Topic	Genre
+viewer,+moment,+relationship,+minute,+love	Romance
+comedy,+funny,+joke,humor,+laugh	Comedy
+effect,special,+action,+special effect,human	Sci-Fi/Fantasy
+war,+battle,+president,+soldier,+history	Action
+cop,+crime,+thriller,police,+action	Action

Part c examines the genre Comedy, so it is useful to examine the **Comedy** topic. This will be exported as the binary variable **TextTopic_5** and the raw SVD variable **TextTopic_raw5**. Select the **Comedy** topic, right-click, and select **Select Current Topic**.

The screenshot shows the "Interactive Topic Viewer" application window with three main tabs:

- Topics**: A table showing topics with their composition, category, and various metrics. The top topic, "+show, +rate, +kid, +script, +recommend", is highlighted in blue.
- Terms**: A table showing terms with their topic weight, role, and document frequency. The term "comedy" has the highest topic weight of 0.512.
- Documents**: A table showing movies with their topic weights across various genres. The top movie, "Scary Movie 3", has a topic weight of 0.671 and is categorized under Comedy and Horror.

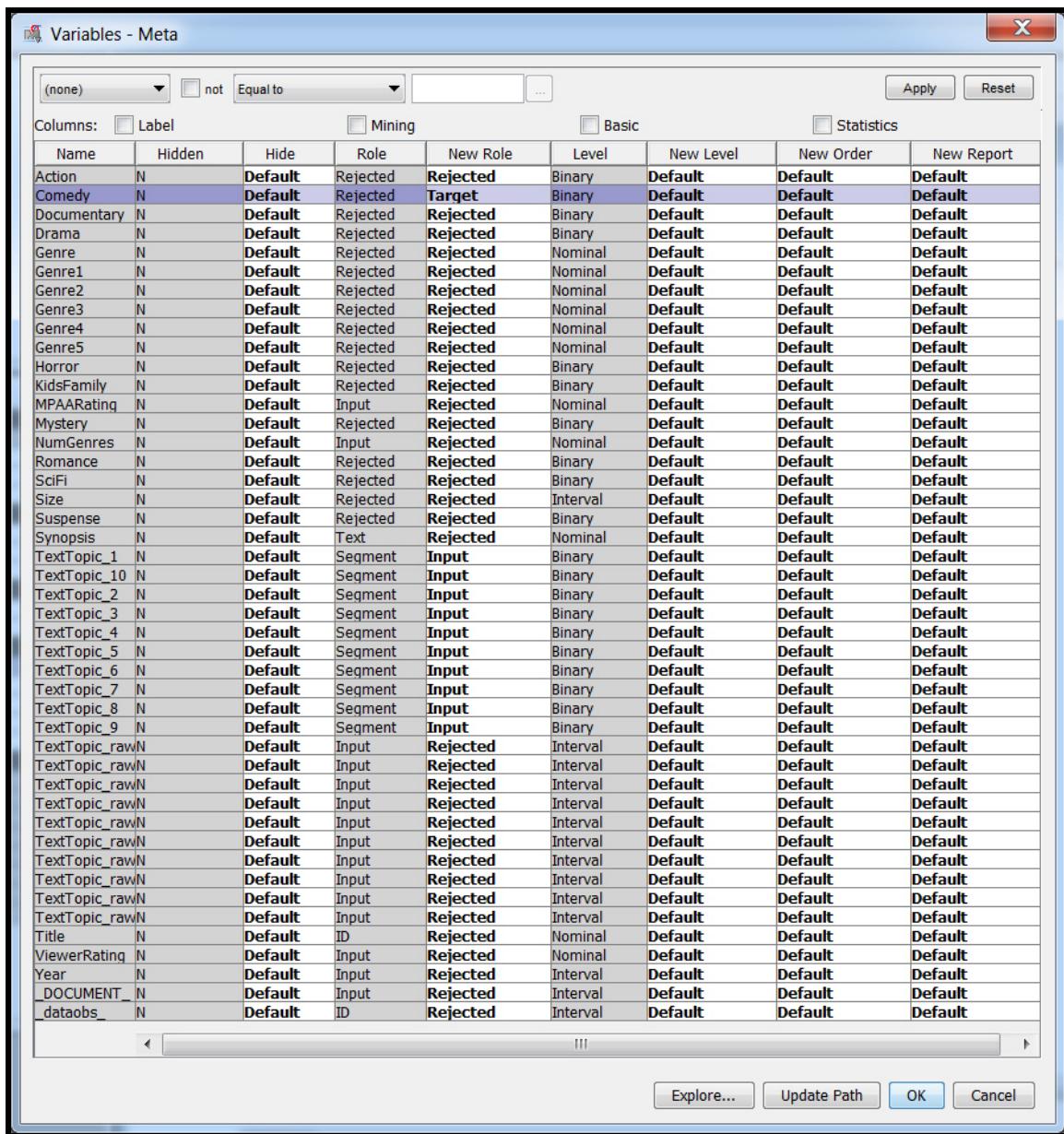
Topic	Category	Term Cutoff	Document Cutoff	Number of Terms	# Docs
+show, +rate, +kid, +script, +recommend	Mult	0.074	0.703	633	166
hollywood, +hand, +protagonist, later, +order	Mult	0.064	0.601	693	107
+viewer, +moment, +relationship, +minute, +love	Mult	0.063	0.569	788	180
+woman, +mother, +school, +girl, +husband	Mult	0.051	0.411	676	170
+comedy, +funny, +joke, humor, +laugh	Mult	0.046	0.342	696	169
+effect, special, +action, +special effect, human	Mult	0.041	0.286	675	144
+war, +battle, +president, +soldier, +history	Mult	0.041	0.298	730	147
+cop, +crime, +thriller, police, +action	Mult	0.041	0.279	725	159
+bond, bond, james, connery, jeffrey	Mult	0.038	0.268	594	124
best, granger, gauge, +oscar, +love	Mult	0.035	0.214	659	162

Topic Weight	+	Term	Role	# Docs	Freq
0.512	+	comedy	Noun	271	431
0.345	+	funny	Adj	181	244
0.312	+	joke	Noun	97	135
0.287		humor	Noun	157	196
0.287	+	laugh	Noun	119	143
0.257		funny	Noun	157	174
0.225		comedic	Adj	95	109
0.199	+	kid	Noun	266	450
0.198		fun	Noun	126	145
0.193		comic	Adj	78	95

Topic Weight	Synopsis	Action	Comedy	Documentary	Drama	Title	Genre
0.671	"Have you ever had	0.0	1.0	0.0	0.0	Scary Movie 3	Comedy, Horror
0.6	With a	0.0	1.0	0.0	0.0	America's Sweethearts	Comedy, Romance
0.583	A kinda-sorta cross	0.0	1.0	0.0	0.0	Head of State	Comedy
0.573	Say what you want	0.0	1.0	0.0	0.0	Mean Girls	Comedy, Romance
0.565	"Harold and Kumar	0.0	1.0	0.0	0.0	Harold and Kumar Go to White Castle	Comedy
0.56	Literally three days	0.0	1.0	0.0	0.0	Scary Movie 2	Comedy, Horror
0.557	In "Christmas with	0.0	1.0	0.0	0.0	Christmas With The Kranks	Christmas, Comedy
0.553	"Joe Dirt" is	0.0	1.0	0.0	0.0	Joe Dirt	Comedy
0.552	After the	0.0	1.0	0.0	0.0	Daddy Day Care	Comedy
0.548	Because THE	0.0	1.0	0.0	0.0	Ladies Man, The	Comedy
0.520	Well, who would	0.0	1.0	0.0	0.0	Election	Comedy

The results look very promising. The top scoring movies in the Documents table are all comedies.

- c. Attach a Metadata node to the Text Topic node. The changes are summarized below.

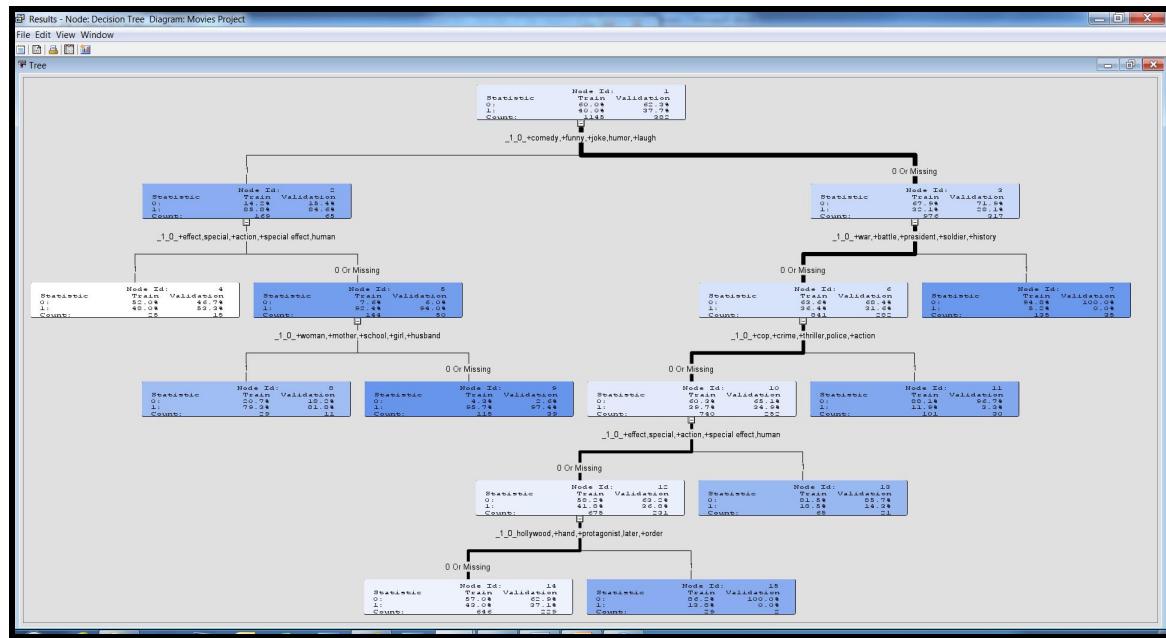


The screenshot shows the 'Variables - Meta' dialog box from a software application. The dialog has a grid of columns for various variables. The columns are labeled: Name, Hidden, Hide, Role, New Role, Level, New Level, New Order, and New Report. There are also buttons for 'Apply', 'Reset', 'Explore...', 'Update Path', 'OK', and 'Cancel'. The 'Mining' column header is selected, indicated by a blue border. The 'Role' column contains values like 'Rejected', 'Target', and 'Text'. The 'Level' column contains values like 'Binary', 'Nominal', and 'Interval'. The 'New Role' column often contains 'Rejected' or 'Default'. The 'New Level' column often contains 'Default'. The 'New Order' and 'New Report' columns contain 'Default'.

Name	Hidden	Hide	Role	New Role	Level	New Level	New Order	New Report
Action	N	Default	Rejected	Rejected	Binary	Default	Default	Default
Comedy	N	Default	Rejected	Target	Binary	Default	Default	Default
Documentary	N	Default	Rejected	Rejected	Binary	Default	Default	Default
Drama	N	Default	Rejected	Rejected	Binary	Default	Default	Default
Genre	N	Default	Rejected	Rejected	Nominal	Default	Default	Default
Genre1	N	Default	Rejected	Rejected	Nominal	Default	Default	Default
Genre2	N	Default	Rejected	Rejected	Nominal	Default	Default	Default
Genre3	N	Default	Rejected	Rejected	Nominal	Default	Default	Default
Genre4	N	Default	Rejected	Rejected	Nominal	Default	Default	Default
Genre5	N	Default	Rejected	Rejected	Nominal	Default	Default	Default
Horror	N	Default	Rejected	Rejected	Binary	Default	Default	Default
KidsFamily	N	Default	Rejected	Rejected	Binary	Default	Default	Default
MPAARating	N	Default	Input	Rejected	Nominal	Default	Default	Default
Mystery	N	Default	Rejected	Rejected	Binary	Default	Default	Default
NumGenres	N	Default	Input	Rejected	Nominal	Default	Default	Default
Romance	N	Default	Rejected	Rejected	Binary	Default	Default	Default
SciFi	N	Default	Rejected	Rejected	Binary	Default	Default	Default
Size	N	Default	Rejected	Rejected	Interval	Default	Default	Default
Suspense	N	Default	Rejected	Rejected	Binary	Default	Default	Default
Synopsis	N	Default	Text	Rejected	Nominal	Default	Default	Default
TextTopic_1	N	Default	Segment	Input	Binary	Default	Default	Default
TextTopic_10	N	Default	Segment	Input	Binary	Default	Default	Default
TextTopic_2	N	Default	Segment	Input	Binary	Default	Default	Default
TextTopic_3	N	Default	Segment	Input	Binary	Default	Default	Default
TextTopic_4	N	Default	Segment	Input	Binary	Default	Default	Default
TextTopic_5	N	Default	Segment	Input	Binary	Default	Default	Default
TextTopic_6	N	Default	Segment	Input	Binary	Default	Default	Default
TextTopic_7	N	Default	Segment	Input	Binary	Default	Default	Default
TextTopic_8	N	Default	Segment	Input	Binary	Default	Default	Default
TextTopic_9	N	Default	Segment	Input	Binary	Default	Default	Default
TextTopic_rawN	N	Default	Input	Rejected	Interval	Default	Default	Default
TextTopic_rawN	N	Default	Input	Rejected	Interval	Default	Default	Default
TextTopic_rawN	N	Default	Input	Rejected	Interval	Default	Default	Default
TextTopic_rawN	N	Default	Input	Rejected	Interval	Default	Default	Default
TextTopic_rawN	N	Default	Input	Rejected	Interval	Default	Default	Default
TextTopic_rawN	N	Default	Input	Rejected	Interval	Default	Default	Default
TextTopic_rawN	N	Default	Input	Rejected	Interval	Default	Default	Default
Title	N	Default	ID	Rejected	Nominal	Default	Default	Default
ViewerRating	N	Default	Input	Rejected	Nominal	Default	Default	Default
Year	N	Default	Input	Rejected	Interval	Default	Default	Default
DOCUMENT_N	N	Default	Input	Rejected	Interval	Default	Default	Default
dataobs_	N	Default	ID	Rejected	Interval	Default	Default	Default

Attach a Decision Tree node. As has been suggested before, changing the Assessment Measure property to **Average Square Error** is recommended. Run the Decision Tree node.

View the derived tree.



The Variable Importance table shows that **TextTopic_5**, the topic identified as a Comedy topic, is the most important variable for classifying a document as a comedy.

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
TextTopic_5	_1_0+_c...	1	1	1	
TextTopic_7	_1_0+_w...	1	0.521819	0.410735	0.78712
TextTopic_6	_1_0+_ef...	2	0.422269	0.407369	0.96471
TextTopic_8	_1_0+_c...	1	0.407104	0.38651	0.94941
TextTopic_2	_1_0_holl...	1	0.23888	0.019993	0.08369
TextTopic_4	_1_0+_w...	1	0.121953	0.1096	0.89871
TextTopic_3	_1_0+_vi...	0	0	0	
TextTopic_1	_1_0+_s...	0	0	0	
TextTopic_10	_1_0_be...	0	0	0	
TextTopic_9	_1_0+_b...	0	0	0	

The Fit Statistics table follows.

Target	Fit Statistics ▲	Statistics Label	Train	Validation
Comedy	<u>ASE</u>	Average Squared Error	0.178682	0.169067
Comedy	<u>DFT</u>	Total Degrees of Freedom	1145	
Comedy	<u>DIV</u>	Divisor for ASE	2290	764
Comedy	<u>MAX</u>	Maximum Absolute Error	0.956522	0.956522
Comedy	<u>MISC</u>	Misclassification Rate	0.29345	0.26178
Comedy	<u>NOBS</u>	Sum of Frequencies	1145	382
Comedy	<u>RASE</u>	Root Average Squared Error	0.422708	0.411177
Comedy	<u>SSE</u>	Sum of Squared Errors	409.1815	129.1668

The validation misclassification rate is lower than the training misclassification rate. Recall that you did not use the **Comedy** target variable in the Data Partition step, so training and validation samples were not stratified on **Comedy**. Furthermore, recall that the decision criterion by default is to classify into the value having the highest probability, which for binary response models implies a cutoff of 50%. If you use a precision/recall break even analysis, the misclassification rate is actually around 33% for precision and recall of about 66%. (You can determine these values by attaching a **SAS Code** node to the **Decision Tree** node and using the program **SCN_PrecisionRecallPlot.sas**.)

Solutions to Student Activities (Polls/Quizzes)

2.03 Multiple Answer Poll – Correct Answers

Which of the following tasks can be performed by the Text Import node?

- a. perform optical character recognition (OCR) of embedded bitmaps in document files
- b. convert Microsoft Word, Excel, and PowerPoint files to ASCII text
- c. process documents having more than 32,000 characters
- d. act as a Web crawler or robot to fetch and convert Internet pages to ASCII text files

20

2.05 Multiple Choice Poll – Correct Answer

Which of the following statements is true?

- a. The demonstration proves that TK wrote the 11 documents.
- b. MBR works best when all inputs are uncorrelated with each other.
- c. MBR is fast enough to promote real-time scoring of new data.
- d. All of the above statements are true.

36

Chapter 3 Algorithmic and Methodological Considerations in Text Mining

3.1 Methods for Parsing and Quantifying Text	3-3
3.2 Quantifying Concepts Using Latent Semantic Analysis	3-24
3.3 Chapter Summary.....	3-34
3.4 Solutions	3-35
Solutions to Student Activities (Polls/Quizzes)	3-35

3.1 Methods for Parsing and Quantifying Text

Objectives

- Explain tokenization and describe the transition from tokens to words in a language.
- Define frequency (local) weights and term (global) weights and describe how weights are used to construct the term with a document frequency matrix.
- Provide guidelines for choosing weights.

3

Text Mining Definitions

Corpus

A collection of documents is called a *corpus*.

Tokens, Separators, and Terms

A document consists of a set of tokens. A *token* is a contiguous string of characters that does not contain a separator. A *separator* is a special character such as a blank or mark of punctuation. A *term* is a token with a specific meaning in a given language.*

* In some languages other than English, a term might contain blanks or other separators.

4

Types of Text Extraction Ordered by Increasing Complexity

1. Token extraction
2. Term extraction (token + language \Rightarrow term)
3. Concept extraction (nouns, noun phrases)
4. Entity extraction (associates nouns with entities – for example, Person: Mr. White, Location: White House)
5. Atomic fact extraction (associates nouns with verbs, that is, subject \Rightarrow action – for example, terrorist \Rightarrow bombed)
6. Complex fact extraction (natural language understanding)

(Wakefield 2004)

5

SAS Sentiment Analysis provides atomic fact extraction capabilities. However, many atomic fact extraction exercises must be customized. For example, you could train a predictive model to mimic categories assigned by professionals. Supervised classification often requires problem-specific tasks related to data preparation and model building. SAS Enterprise Miner and SAS Text Miner provide a framework for developing custom atomic fact extraction solutions.

Characteristics of a Document

A document consists of the following:

- letters
- words
- sentences
- paragraphs
- punctuation
- possible structural items (chapters, sections)

The elements of a document can be

- counted (for example, the number of characters, words, or sentences)
- summarized (for example, relative frequency).

6

Weighting strategies are introduced later. Relative frequency is a useful weighting mechanism that adjusts for document size. The relative frequency of word X is the number of times that X appears in a document divided by the number of words in the document.

Zipf's Law

Let t_1, t_2, \dots, t_n be the terms in a document collection arranged in order from most frequent to least frequent.

Let f_1, f_2, \dots, f_n be the corresponding frequencies of the terms. The frequency f_k for term t_k is proportional to $1/k$.

Zipf's law and its variants help quantify the importance of terms in a document collection. (Konchady 2006)

"The product of the frequency of words (f) and their rank (r) is approximately constant."

7

In practice, Zipf's Law is derived as a Power Law, with free parameters that can be estimated based on the document collection. The general formula is shown here:

$$f_k = C / (\omega + k)^\theta$$

where C is a constant such that, for given ω and θ , $\sum_{k=1}^n f_k = T$, the total number of words in the document

collection. The parameters ω and θ are estimated for a given document collection.

Konchady (2006) relates Zipf's Law to quantifying the importance of a term: "...the number of meanings of a word is inversely proportional to its rank." (Konchady 2006, page 87) Application of Zipf's Law permits identification of important terms for purposes such as describing concepts or topics. You will not encounter Zipf's Law (or similar theoretical laws) directly, but you can see the results of Zipf's Law in text mining applications, for example, in the list of terms used to define a topic. Along with methods such as Hidden Markov Models (HMM), the implementation is often hidden from the user; only the results of the methodology are visible.

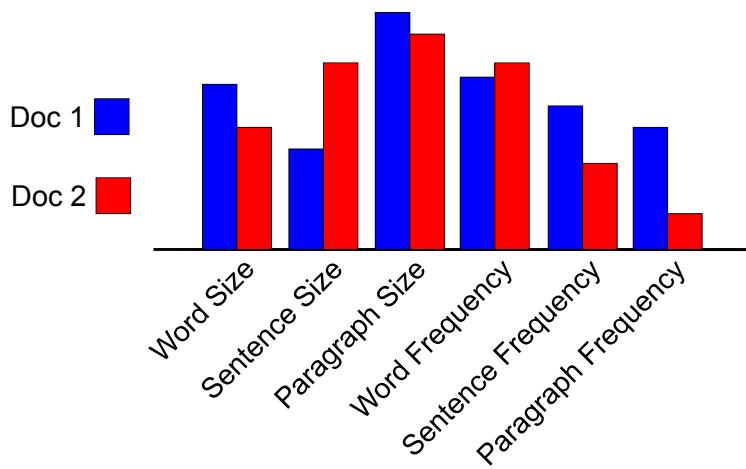
Relevance of Zipf's Law to Text Mining

- Often, a few, very frequent terms are not good discriminators.
 - *stop* words, for example, the, and, an, or, of
 - often words that are described in linguistics as “closed-class” words, which is a grammatical class that does not get new members
- Typically, there is the following in a document collection:
 - a high number of infrequent terms
 - an average number of average frequency terms
 - a low number of high frequency terms
- ✍ Terms that are neither high nor low frequency are the most informative.

8

Frequency filtering is suggested by Zipf's Law.

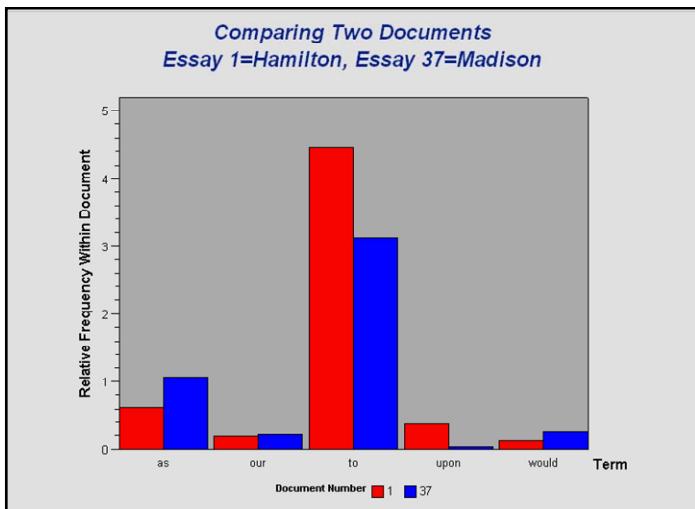
Comparing Two Documents



9

You can compare two documents using a variety of quantities. In practice, comparisons are usually made with respect to specific well-defined objects, such as words, sentences, or paragraphs. Thus, you often see comparisons by word statistics, with sentence and paragraph information used to calculate additional word statistics. Statistics are often relative in nature. For example, for words, you might calculate relative frequency per document, relative frequency per paragraph, and relative frequency per sentence. Basic statistics are limited in usefulness and are primarily used for applications such as stylometry.

Comparing Essay 1 to Essay 37



10

continued...

Word statistics can be good differentiators. However, most problems that can be solved with the help of text mining require a large number of words, typically thousands of words, rather than five words as is used in the illustration.

Comparing Essay 1 to Essay 37

Frequency Percent Row Pct Col Pct	PaperNum	Table of PaperNum by Word					
		Word					Total
	PaperNum	as	our	to	upon	would	
1	1	10 4.52 10.75 25.64	3 1.36 3.23 33.33	72 32.58 77.42 45.86	6 2.71 6.45 85.71	2 0.90 2.15 22.22	93 42.08
	37	29 13.12 22.66 74.36	6 2.71 4.69 66.67	85 38.46 66.41 54.14	1 0.45 0.78 14.29	7 3.17 5.47 77.78	128 57.92
	Total	39 17.65	9 4.07	157 71.04	7 3.17	9 4.07	221 100.00

11

continued...

Two distance measures are popular in text mining: Euclidean distance and cosine distance. A third measure uses the association between two documents given a common set of terms. Two documents quantified by a list of n terms produce a 2-by- n contingency table. Several measures of similarity or distance can be derived based on a contingency table analysis.

Comparing Essay 1 to Essay 37

Statistic	DF	Value	Prob
Chi-Square	4	12.4514	0.0143
Likelihood Ratio Chi-Square	4	13.0976	0.0108
Mantel-Haenszel Chi-Square	1	3.6720	0.0553
Phi Coefficient		0.2374	
Contingency Coefficient		0.2309	
Cramer's V		0.2374	
WARNING: 40% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

12

continued...

The phi coefficient varies between -1 and +1. A value near +1 represents strong positive association. A value near -1 suggests strong negative association. A value near zero suggests little or no association.

Comparing Essay 1 to Essay 37

Distance Measure	Not Similar	Actual Value	Similar
Phi	1.0	0.237	→ 0.0
Euclidean	∞	?	24.27
Cosine	0.0	0.978	→ 1.0

The phi coefficient and cosine distance judge the two essays to be similar. Interpreting Euclidean distance requires knowledge of the distances to other documents in the collection.

13

The phi coefficient judges similarity using the chi-square statistic for a contingency table. A large chi-square implies that knowing the level of one variable (term) helps predict the level of another variable (document). When two documents are very similar, knowing the term frequencies is *not* useful in predicting which document the terms appear in. Thus, a small value for phi implies that two documents are similar.

In Euclidean space, the angle between the two documents is about 12° . An angle of 0° is perfect association, and $\text{cosine}(0)=1$. Cosine is a useful measure because it is a relative measure and is not affected by the size of the documents. Phi has similar advantages. Both phi and cosine distance judge essays 1 and 37 to be similar. On the other hand, judging Euclidean distance requires knowledge of the distances to other documents for comparison.

Text mining software rarely provides the actual distance calculations. Instead, the results of the calculations are displayed. For example, the Text Topic node develops topics that are defined by term weights and topic weights. Therefore, distance measures can be developed to see how close topics are to each other. The Text Topic node considers two topics to be similar if the angle between them is less than three degrees.

3.01 Multiple Choice Poll

Which of the following conditions indicates that two documents are similar?

- a. a value of the phi coefficient near 0
- b. a value of cosine distance near 1
- c. a small value for Euclidean distance
- d. all of the above

Comparing Terms

Methods for comparing documents can also be applied to comparing terms.

Term	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8
diabetes	12	9	2	0	0	4	0	1
insulin	2	0	2	0	3	6	0	0

- 2 by n contingency table \Rightarrow phi coefficient
- 2 vectors of dimension $n \Rightarrow$ Euclidean distance; cosine distance

17

Comparing Term *diabetes* to Term *insulin*

Distance Measure	Not Similar	Actual Value	Similar
Phi	1.0	0.651	0.0
Euclidean	∞	13.96	0.0
Cosine	0.0	0.455	1.0

 The data is hypothetical.

18

Conditional Counts: Concept Linking

Centered term: a term that is chosen to investigate

diabetes (63/63)

+insulin (14/58)

Concept linked term: a term that co-occurs with a centered term

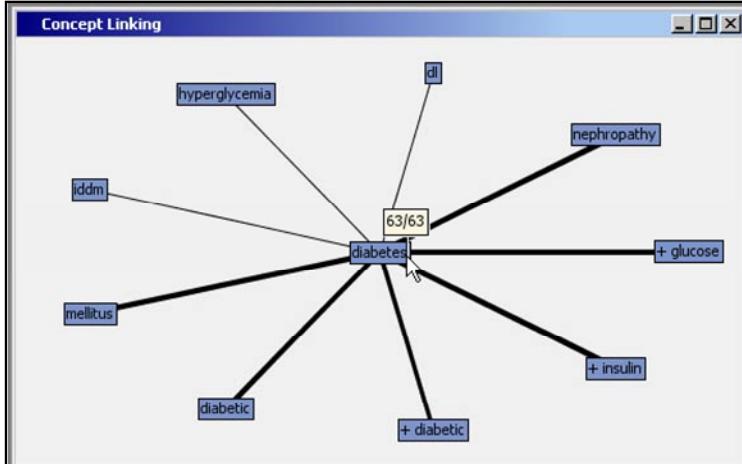
✍ This data is from the Medline abstracts.

19

continued...

Concept linking is available in the Interactive Filter Viewer of the Text Filter node. In the viewer, access the Terms table, select a term, right-click, and select **View Concept Links**.

Conditional Counts: Concept Linking



The term *diabetes* occurs in 63 documents.

20

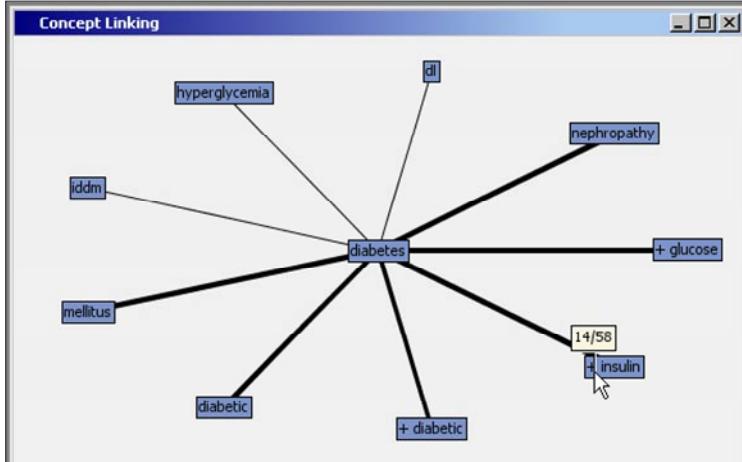
continued...

The Reference Help provides the following description:

“The width of the line between the centered term and a concept link represents how closely the terms are associated. A thicker line indicates a closer association.”

The actual metric employed to judge association strength is not given.

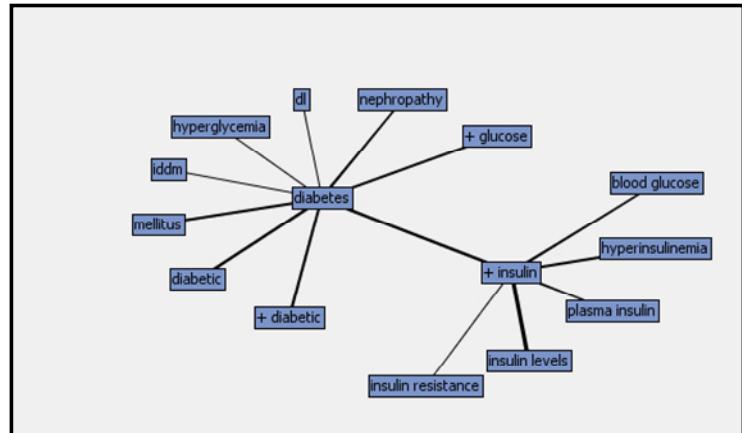
Conditional Counts: Concept Linking



The term *insulin* and its variants occur in 58 documents, and

²¹ 14 of those documents also contain the term *diabetes*. *continued...*

Conditional Counts: Concept Linking



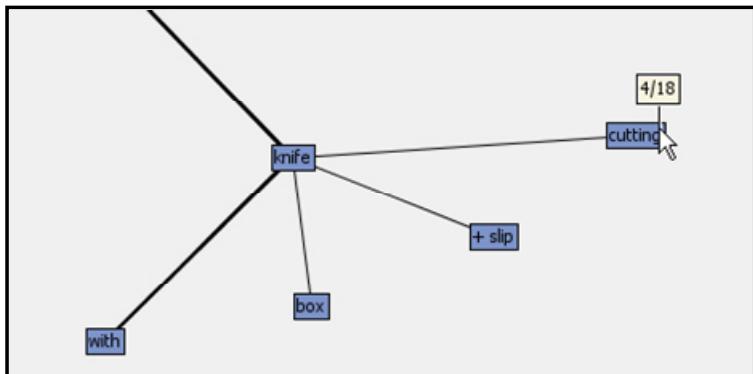
Terms that are primary associates of *insulin* are secondary associates of *diabetes*.

²²

While the concepts related to document and term distances, associations, and similarities are relevant for text mining, the raw frequency counts of terms in documents is typically too primitive to be used for text mining. Weighting strategies and sophisticated linear algebra techniques help move from counting words to extracting concepts.

Setup for the Poll

Consider the following section of a concept linking plot:



24

3.02 Multiple Choice Poll

How many documents contain the term *cutting* as indicated in the setup slide?

- a. 18
- b. 4
- c. 14
- d. 22

25

Quantifying Terms

Example Corpus: 5,280 terms → 5,280 variables

Strategy 1 (existence):

if (Term exists in document) then Variable=1
else Variable=0

Strategy 2 (frequency):

Variable=count(Term in document)

27

When you have 5,280 columns rather than 6 or 8 columns, new problems arise with respect to measuring distance, similarity, and association between documents and terms.

Quantified Text Data

Documents	Term Variables				
	Variable 1	Variable 2	...	Variable 5280	
Document 1					
:					
Document 3000					

Data



Problems: (1) Too big
(2) Sparse ⇒ mostly zeros

28

When the ratio of terms to documents is small, then most documents will not contain most terms. A matrix dominated by zeros can pose problems when you are trying to extract information. Many queries will have a null result. This is especially true of queries investigating the co-occurrence of many terms. One strategy for improving the text data matrix is to transition from simple counts to more complex weighting schemes.

Frequency Weights (Local Weights)

None $L_{ij} = a_{ij}$

Binary $L_{ij} = \begin{cases} 1 & \text{if term } i \text{ is in document } j \\ 0 & \text{otherwise} \end{cases}$

Log $L_{ij} = \log_2(a_{ij} + 1)$

a_{ij} is the number of times that term i appears in document j .

29

The *frequency weights*, often called *local weights* in the text mining and information retrieval literature, represent the first step in quantifying documents. Unfortunately, absolute counts can be influenced by documents that have high variability with respect to size.



The keyword None does not imply that no frequency weights are employed. The untransformed term-document frequency matrix uses simple counts (frequencies). The selection **None** implies that no transformation will be applied to the term-document matrix. The keywords Binary and Log imply that a transformation of the raw counts will be performed. An alternate set of keywords to help avoid ambiguities might be Frequency, Binary, and Log Frequency, but SAS Text Miner uses the untransformed versus transformed view.

Weighted Term-Document Frequency Matrix

Term	ID	Documents		D_n
		D1	D2	
T1	1	$L_{1,1}$	$L_{1,2}$	$L_{1,n}$
T2	2	$L_{2,1}$	$L_{2,2}$	$L_{2,n}$

⋮

⋮

L_{ij} = frequency weight for term i and document j

30

Term weights, often called *global weights* in the literature, modify frequency weights to adjust for document size and term distribution.

Term Weights in SAS Text Miner

- Entropy (default)
- IDF (Inverse Document Frequency)
- Mutual Information (Target-Based)

31

Entropy weights and inverse document frequency (IDF) weights seem to dominate the literature, but researchers are also fond of creating new global weights to address limitations of existing weights for specific document collections. ***You will probably find that entropy weights or IDF weights are adequate for your problem if your problem can in fact benefit from text analytics.***

A brief discussion of the formulas behind the weights begins below. Although you might gain some insight by looking at the mathematics, often experimentation rather than intuition is the best strategy for choosing weights. Experience with similar text analytic problems can help you develop your own guidelines.

Term Weight Formulas

$a_{i,j}$ = frequency that term i appears in document j

g_i = frequency that term i appears in document collection

n = number of documents in the collection

d_i = number of documents in which term i appears

$$p_{i,j} = a_{i,j} / g_i$$

Term Weight Formulas

Entropy

$$G_i = 1 + \sum_{j=1}^{d_i} \frac{p_{ij} \log_2(p_{ij})}{\log_2(n)}$$

$$0 \leq G_i \leq 1$$

Low Information \longrightarrow High Information

33

continued...

Because the logarithm of zero is undefined, the product in the numerator is taken to be zero if the proportion p is zero. If a term appears exactly one time in exactly one document, then the entropy weight for the term is one. If a term appears exactly one time in every document, then the entropy weight for the term is zero.

$$G_i = 1 + \frac{(1/1) \log_2(1/1)}{\log_2(n)} = 1 + \frac{(1)(0)}{\log_2(n)} = 1$$

$$G_i = 1 + \frac{\sum (1/n) \log_2(1/n)}{\log_2(n)} = 1 + \frac{n(1/n)(-\log_2(n))}{\log_2(n)} = 1 + \frac{-\log_2(n)}{\log_2(n)} = 1 - 1 = 0$$

Term Weight Formulas

IDF (Inverse Document Frequency)

$$G_i = 1 + \log_2 \left(\frac{n}{d_i} \right)$$

$$1 \leq G_i < \infty$$

Low Information \longrightarrow High Information

34

continued...

If a term appears in every document, then the IDF weight is 1. The maximum weight for a fixed document collection occurs when the term appears in exactly one document, and the weight becomes $1 + \log_2(n)$. No upper limit exists because no theoretical limit exists for the number of documents in a collection.

Entropy and IDF weights achieve a maximum when exactly one term appears in exactly one document. Both weights are at minimum or near minimum if a term appears exactly one time in every document.

Term Weight Formulas

Mutual Information:

$$G_i = \max_k \left(\log \left(\frac{P(x_i, C_k)}{P(x_i)P(C_k)} \right) \right)$$

$$x_i = \begin{cases} 1 & \text{if term } i \text{ is present} \\ 0 & \text{if term } i \text{ is absent} \end{cases}$$

Target variable T takes categorical values C_1, \dots, C_k .

Term i	Target Category			
	C_1	C_2	\dots	C_k
$x=0$				
$x=1$				

2 by k
Contingency Table

35

Multiplying the local and global weights produces an adjusted count that is often superior to using raw counts alone.

Weighted Term-Document Frequency Matrix

$$\hat{a}_{ij} = G_i L_{ij}$$

G_i = term weight for term i

L_{ij} = frequency weight for term i in document j

36

continued...

The weighted term-document frequency matrix is the foundation of the linear algebra approach to text mining.

Weighted Term-Document Frequency Matrix

		Documents		
		D1	D2	Dn
Terms				
T1		$\hat{a}_{1,1}$	$\hat{a}_{1,2}$	$\hat{a}_{1,n}$
T2		$\hat{a}_{2,1}$	$\hat{a}_{2,2}$	$\hat{a}_{2,n}$
T_m		$\hat{a}_{m,1}$	$\hat{a}_{m,2}$	$\hat{a}_{m,n}$

37

Term Weight Guidelines

- Entropy and IDF weights give high weights to rare or low frequency terms.
- Entropy and IDF weights give moderate to high weights for terms that appear with moderate to high frequency, but in a small number of documents.
- Entropy and IDF weights vary inversely to the number of documents in which a term appears.
- Entropy is often superior for distinguishing between small documents that contain only a few sentences.
- Entropy is the only term weight that depends on the distribution of terms across documents.

38

continued...

Term Weight Guidelines

- In general, if a term tends to occur in documents having the same target category, then that term will receive a high weight for mutual information term weights.

39

A simulation study artificially creates a document collection and distributes terms across the documents using various strategies – for example, creating rare terms and creating terms with frequency counts that follow a certain distribution.

Term	Term Freq	Doc Freq	Entropy	IDF	Mutual Information
armadillo	102	2	0.8495	6.6439	0.4943
bear	105	64	0.1264	1.6439	0.1839
cat	113	59	0.1405	1.7612	0.0421
cow	110	66	0.1107	1.5995	0.2177
dog	107	66	0.1183	1.5995	0.0478
gopher	106	55	0.1580	1.8625	0.2665
hamster	109	65	0.1194	1.6215	0.4308
horse	109	62	0.1315	1.6897	0.1818
kitten	105	62	0.1303	1.6897	0.0307
moose	1934	100	0.0973	1.0000	0.0000
mouse	108	63	0.1296	1.6666	0.0943
otter	1	1	1.0000	7.6439	0.4943
pig	107	58	0.1440	1.7859	0.0592
puppy	115	58	0.1576	1.7859	0.5447
raccoon	967	50	0.2478	2.0000	0.1086
seal	10	10	0.5000	4.3219	0.2712
squirrel	100	100	0.0000	1.0000	0.0000
tiger	117	70	0.1027	1.5146	0.0070
walrus	25	25	0.3010	3.0000	0.0480
zebra	3812	100	0.1008	1.0000	0.0000

40

You can verify the IDF calculations using the Doc Freq column and noting that there are 100 documents in the simulation. For example, for the term *armadillo*, the IDF term weight is as follows:

$$1 + \log_2(100 / 2) = 1 + \log_2(50) = 6.6439$$

The gray scale version is difficult to interpret, but the color version of the table highlights low information terms with a red background, and high information terms with a green background.

The two terms in the mutual information column are the two terms that defined the target variable. If a document contained both the term *hamster* and the term *puppy*, then the target variable was assigned a value of 1. Otherwise, it received a value of 0. Two terms received a higher target-based term weight than *hamster* because the two terms by happenstance appear less frequently than *hamster* when the term *puppy* is not present. In particular, by happenstance, the one document that contains *otter* also contains *hamster* and *puppy*, and the two documents that contain *armadillo* also contain both *hamster* and *puppy*.

The results show that entropy and IDF weights tend to produce similar results. IDF is recommended for larger documents, whereas entropy might be more appropriate for smaller documents. Of course, these results cannot be extrapolated to all document collections. In particular, a typical document in the simulated collection is small, so the results would be more useful for document collections such as the Medline data, but less useful for multi-page reports.

3.03 Multiple Choice Poll

Which term weight is recommended for collections of small documents – for example, a paragraph or smaller?

- a. entropy
- b. IDF
- c. mutual information
- d. none ($G=1$)

3.2 Quantifying Concepts Using Latent Semantic Analysis

Objectives

- Review how linear algebra is used in information retrieval.
- Illustrate the use of linear algebra in the methodology named Latent Semantic Analysis (LSA).

45

3.04 Multiple Choice Poll

Which response best describes your preference?

- a. I am eager to learn the technical details of Latent Semantic Analysis and the Singular Value Decomposition.
- b. I would like to understand the concepts that are important to LSA, but I would prefer to skip the math.
- c. I do not really care how LSA works. I only want to know how to use the software to solve my problem.
- d. I am only here to keep from doing real work.

47

Linear Algebra in Text Mining

- For a document collection, terms in a start list are counted in some fashion for each document.
(Counting \Rightarrow Frequency Weights)
- With documents as rows (observations) and terms as columns (variables), the counts form a **matrix**.
- Linear algebra includes the study of **matrices** and **matrix properties**.
- The **theory** of linear algebra guarantees that every matrix accommodates a **singular value decomposition (SVD)**.
- The SVD of a document-by-term matrix produces a set of **coefficients** that can be used to derive **concepts or topics** from the document collection.

48

The use of linear algebra, and in particular, the use of SVD in text analytics, is named latent semantic analysis (LSA).

In the following slide, the first two SVD vectors are displayed for a document collection having 10 terms.

SVD Concepts

SVD Coefficients		
Term	First Vector	Second Vector
cow	0.01450	0.32506
dolphin	0.36954	-0.02047
goat	0.04167	0.38839
horse	0.01800	0.49270
otter	0.53649	-0.02920
pig	0.01290	0.35068
seal	0.54615	-0.02805
sheep	0.01654	0.61274
walrus	0.35931	-0.01593
whale	0.38147	-0.00040

49

Values near zero occur for *cow*, *goat*, *horse*, *pig*, and *sheep* for the first vector. The remaining values are larger, suggesting that the first SVD dimension might indicate whether a document exhibits a marine mammal concept. Similarly, the marine mammals in the second vector have values near zero, suggesting a land mammal, or perhaps a farm animal concept.

The SVD vectors contain weights that are multiplied by the corresponding term-frequency weight products in the term-document frequency matrix. The following example uses frequency weight option None and term weight option None. Thus, the values for the column Doc A are simply counts of how often a term appears in document A.

SVD Concepts

Term	SVD Weight	Doc A	Concept Strength
cow	0.01450	2	0.02900
dolphin	0.36954	0	0.00000
goat	0.04167	0	0.00000
horse	0.01800	0	0.00000
otter	0.53649	0	0.00000
pig	0.01290	3	0.03870
seal	0.54615	0	0.00000
sheep	0.01654	0	0.00000
walrus	0.35931	1	0.35931
whale	0.38147	0	0.00000

Marine mammal concept is **absent** in Doc A. 0.42701

50

Multiplying the SVD weights by the term-frequency product and summing produces a quantity that represents the strength of the term in the document. The small sum suggests that document A might not exhibit a marine mammal concept.

SVD Concepts

Term	SVD Weight	Doc B	Concept Strength
cow	0.01450	0	0.00000
dolphin	0.36954	1	0.36954
goat	0.04167	0	0.00000
horse	0.01800	0	0.00000
otter	0.53649	0	0.00000
pig	0.01290	1	0.01290
seal	0.54615	3	1.63845
sheep	0.01654	0	0.00000
walrus	0.35931	0	0.00000
whale	0.38147	0	0.00000

Marine mammal concept is **present** in Doc B. 2.02089

51

The marine mammal concept is stronger for document B.

Each document has a value for each SVD dimension.

SVD Concepts

Term	SVD Weight	Doc B	Concept Strength
cow	0.01450	0	0.00000
dolphin	0.36954	1	0.36954
goat	0.04167	0	0.00000
horse	0.01800	0	0.00000
otter	0.53649	0	0.00000
pig	0.01290	1	0.01290
seal	0.54615	3	1.63845
sheep	0.01654	0	0.00000
walrus	0.35931	0	0.00000
whale	0.38147	0	0.00000

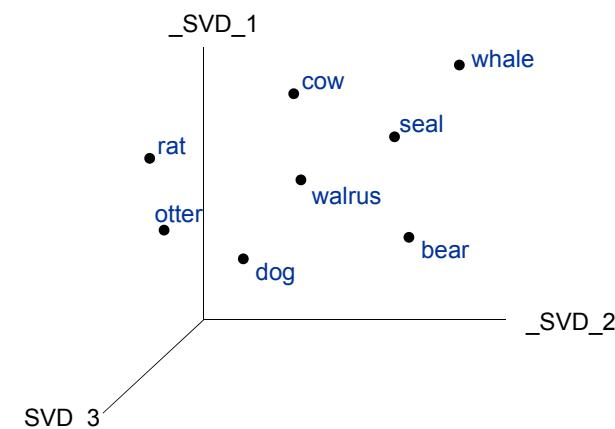
This is the value in the SVD_k column of Doc B. 2.02089

52

SAS Text Miner only reveals the results of the calculation. The Text Topic node starts with the basic SVD concept vectors, and then applies a rotation transformation to try to produce concepts that can be more easily interpreted. The methodology comes from *factor analysis*.

The following slide illustrates documents with concepts indicated by a single keyword – for example, *walrus*.

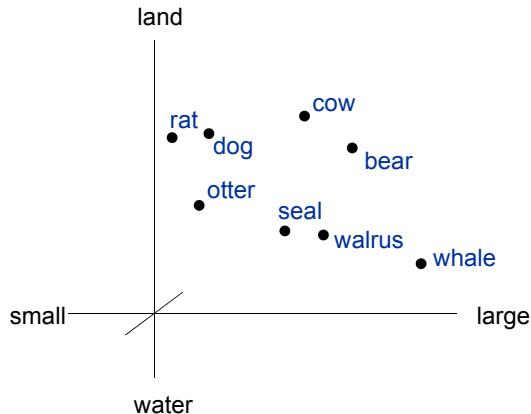
SVD Concepts



53

In the original SVD space, the concepts are not easily identified.

SVD Concepts: Rotated to Form Topics



54

After rotation, the results above appear.

Apparently information is concentrated in two dimensions, and the two concepts are based on size and habitat. The factor analysis approach is promising for revealing the concepts that are extracted by the singular value decomposition. The Text Topic node uses a factor rotation of the original SVD columns.

Documents with SVD Weights: Clusters

Text	TextCluster_SVD1	TextCluster_SVD2	TextCluster_SVD3
If you're a die-hard	0.444762	0.085007	-0.02054
BIG NIGHT is a charmi	0.619079	0.107943	0.274286
Rose are red.	0.640147	0.422202	0.105264
Based on the novel by	0.663075	0.266579	-0.20034
A story of kidnapping.	0.367888	0.082505	0.019349
As with 2001's "The Fa	0.634069	0.215578	-0.24684
Stephen Daldry's BILLY	0.572333	0.431265	0.149547
THE BIRDCAGE is a ...	0.52775	0.208953	0.174304
In a bracing and welco	0.500374	0.036629	-0.11877
Nicole Kidman's versat	0.654763	0.189677	-0.00792
In 1992 French writer	0.512185	-0.2885	0.140257
BLACK BEAUTY is not a	0.502661	0.304246	0.081929
Pow! Boom! Bang! Cr...	0.536203	0.015859	-0.000706
I remember seeing this	0.307794	0.072569	-0.21323
"Ask not what your fie	0.576648	0.129828	0.206674
In the early 1990s it	0.570828	-0.18631	-0.08289

✓ The SVD concept appears to be present in the document.

55

The Text Cluster node produces the original SVD columns.

Documents with SVD Weights: Topics

Text	TextTopic_raw1	TextTopic_raw2	TextTopic_raw3
There are many ways in	0.976	1.693	0.811
The many characters in	0.714	0.649	1.176
By the time this movie	0.649	0.530	0.411
For more than two yea	0.157	0.155	0.148
In today's day and age	0.753	0.613	0.729
The mere idea of _2_Fa	0.653	0.661	0.897
Good-looking Myles Be	0.837	0.527	0.314
Long time ago there we	0.948	1.610	0.794
This MTV-production r	0.194	0.195	0.141
"2001" is probably th	0.462	0.558	0.461
"What am I doing in th	0.461	0.250	0.241
We throw around aphori	0.649	0.684	0.479
The girl-next-door has	0.148	0.191	0.191
If you, like me, are a	0.407	0.357	0.303
Demian Lichtenstein's	0.522	0.226	0.364

- ☛ The SVD concept appears to be present in the document.

56

The Text Topic node produces rotated SVD columns.

The SVD columns above populate the exported training data from the Text Cluster and Text Topic nodes. There are options for extracting the SVD component vectors from the results of the Text Miner nodes. SAS programming using the SAS Code node is required. The Project Start Code must set the SVD macro variable as follows:

```
%let TM_SVDDATA=1;
```

This statement sets a macro variable that causes SVD results to be saved for later reference. The saved SVD tables must be accessed using SAS Enterprise Miner macro variables and special naming conventions. The following sample program displays the U matrix from the singular value decomposition.

```
%global LastParsing LastCluster TermData
      SMatrix UMatrix;
%let LastParsing= ;
%let LastCluster= ;

proc print data=&EM_IMPORT_DATA_EMINFO;
run;

proc sql noprint;
  select data into :LastCluster
  from &EM_IMPORT_DATA_EMINFO
  where key="LastTextCluster";
  select data into :LastParsing
  from &EM_IMPORT_DATA_EMINFO
  where key="LastTextParsing";
quit;

%put NOTE: Last SAS Text Parsing Node: &LastParsing;
%put NOTE: Last Text Cluster Node: &LastCluster;
```

```
%let
TermData=%sysfunc(strip(&EM_LIB)).%sysfunc(strip(&LastParsing))_TERMS;
%let
SMatrix=%sysfunc(strip(&EM_LIB)).%sysfunc(strip(&LastCluster))_svd_s;
%let
UMatrix=%sysfunc(strip(&EM_LIB)).%sysfunc(strip(&LastCluster))_svd_u;
%EM_REGISTER(KEY=TERMS,TYPE=DATA);
%EM_REGISTER(KEY=SVDS,TYPE=DATA);
%EM_REGISTER(KEY=SVDU,TYPE=DATA);

%macro GetUMatrix();
%if %sysfunc(exist(&TermData)) %then %do;
  %EM_REPORT(KEY=TERMS,VIEWTYPE=DATA,AUTODISPLAY=Y);
  data &EM_USER_TERMS;
    set &TermData;
    rename KEY=_TERMINUM_;
  run;
  proc contents data=&EM_USER_TERMS;
  run;
%end;
%if %sysfunc(exist(&SMatrix)) %then %do;
  %EM_REPORT(KEY=SVDS,VIEWTYPE=DATA,AUTODISPLAY=Y);
  data &EM_USER_SVDS;
    set &SMatrix;
  run;
  proc contents data=&EM_USER_SVDS;
  run;
%end;
%if %sysfunc(exist(&UMatrix)) %then %do;
  %EM_REPORT(KEY=SVDU,VIEWTYPE=DATA,AUTODISPLAY=Y);
  data &EM_USER_SVDU;
    set &UMatrix;
    rename INDEX=_TERMINUM_;
  run;
  proc contents data=&EM_USER_SVDU;
  run;
%if %sysfunc(exist(&EM_USER_TERMS)) %then %do;
  proc sql;
    create table &EM_USER_SVDU as
      select a.*, b.TERM
      from &EM_USER_SVDU a, &EM_USER_TERMS b
      where a._TERMINUM_=b._TERMINUM_;
  quit;
%end;
%end;
%mend GetUMatrix;

%GetUMatrix();
```

Dimensionality Reduction

- There are as many SVD dimensions as there are kept terms in the document collection. For example, if there are 5,280 kept terms (Keep Status=Yes), then there are 5,280 SVD columns.
- The SVD algorithm derives SVD weights in order of importance. Thus, the concept represented by _SVD_1 is more important than the concept represented by _SVD_2, and so on.
- There is a cutoff k such that SVD concepts beyond the first k are not important. The value k is usually much smaller than the total number of terms.
- A matrix with k SVD columns will be much smaller than a matrix with the full number of kept terms.

57

continued...

Dimensionality Reduction

- The user specifies a maximum dimension M (default=100) for the number of SVD columns.
- The SVD algorithm produces M singular values in decreasing order.
- The sum of the M singular values acts as a metric for the amount of information in the document collection. If $M=N$ =the total number of kept terms, then the sum of the singular values is in some sense the amount of information in the collection. Treating the sum of the top M values as “total information” is useful for arriving at a so-called *optimal dimensionality*.
- SAS Text Miner uses heuristics to decide the actual number of SVD values to use as input variables for further analysis.

58

continued...

Dimensionality Reduction

Resolution

- High=100%
- Medium=5/6=83.3%
- Low=3/4=75%

The derived SVD dimension is the minimum number of SVD columns that produces the desired resolution.

59

continued...

Dimensionality Reduction

The screenshot shows the SAS Text Miner interface. At the top, there is a 'Transform' panel with 'SVD Resolution' set to 'Low' and 'Max SVD Dimensions' set to '100'. Below this is a 'Variables - EMCODE2' dialog box. The 'Basic' tab is selected in the 'Columns' section. A list of variables is shown, including 'TextCluster_SVD39' through 'TextCluster_SVD48', 'TextCluster_SVD49', and 'TextCluster_cluster_'. A callout box points to the 'TextCluster_SVD49' row with the text: 'Max=100 and low resolution produce 49 SVD dimensions.' At the bottom of the dialog box are 'Explore...', 'Update Path', 'OK', and 'Cancel' buttons.

60

The problem of dimensionality reduction is challenging in a text mining setting. The default settings in the SAS Text Cluster node might be adequate, but the first time that you encounter a new type of document collection, you should experiment with the maximum number of SVD vectors.

3.3 Chapter Summary

Text mining consists of the following steps:

1. preparing the data
2. parsing the text
3. converting the text to a numeric representation
4. transforming the numeric representation
5. reducing the dimensionality of the transformed representation
6. analyzing the text through the reduced dimension representation

SAS Text Miner nodes provide numerous strategies for completing the above steps.

The Linear Algebra approach to text mining, using the singular value decomposition, permits extraction of concepts and topics from a document collection.

For Additional Information

Albright, Russell. 2004. *Taming Text with the SVD*. SAS Institute White Paper.

Albright, R., J.A. Cox, and K. Daly. 2001. "Skinning the Cat: Comparing Alternative Text Mining Algorithms for Categorization." Proceedings of the 2nd Data Mining Conference of DiaMondSUG, Chicago, IL. DM Paper 113.

Cherniak, Eugene. 1993. *Statistical Language Learning*. Cambridge, Massachusetts: The MIT Press.

Evangelopoulos, Nicholas, Xiaoni Zhang, and Victor R. Prybutok. 2010. "Latent Semantic Analysis: Five methodological recommendations." *European Journal of Information Systems*. 1-17.

Jurafsky, Daniel, and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, New Jersey: Prentice Hall.

Konchady, Manu. 2006. *Text Mining Application Programming*. Boston: Charles River Media.

Manning, Christopher D., and Hinrich Schütze. 2002. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. New York: Cambridge University Press.

Shannon, C.E. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal*, Vol. 27, pp. 379-423 and 623-656.

Thisted, Ronald A. 1988. *Elements of Statistical Computing*. New York: Chapman and Hall.

Wakefield, Todd. 2004. "A Perfect Storm is Brewing: Better Answers are Possible by Incorporating Unstructured Data Analysis Techniques." *DM Direct*, August 2004.

3.4 Solutions

Solutions to Student Activities (Polls/Quizzes)

3.01 Multiple Choice Poll – Correct Answer

Which of the following conditions indicates that two documents are similar?

- a. a value of the phi coefficient near 0
- b. a value of cosine distance near 1
- c. a small value for Euclidean distance
- d. all of the above**

16

3.02 Multiple Choice Poll – Correct Answer

How many documents contain the term *cutting* as indicated in the setup slide?

- a. 18**
- b. 4
- c. 14
- d. 22

26

3.03 Multiple Choice Poll – Correct Answer

Which term weight is recommended for collections of small documents – for example, a paragraph or smaller?

- a. entropy
- b. IDF
- c. mutual information
- d. none ($G=1$)

Chapter 4 Applications of Text Mining to Pattern Discovery

4.1 Text Mining in Warranty Analysis.....	4-3
Demonstration: Text Analytics for Warranty Analysis.....	4-7
4.2 Processing and Categorizing Documents.....	4-25
Demonstration: Text Categorization for Identifying Potential Fraud Cases	4-33
4.3 Association and Sequence Discovery in Text Analytics	4-53
Demonstration: Association Discovery of Terms	4-56
Exercises	4-58
4.4 Chapter Summary.....	4-59
4.5 Solutions	4-60
Solutions to Exercises	4-60
Solutions to Student Activities (Polls/Quizzes)	4-67

4.1 Text Mining in Warranty Analysis

Objectives

- Analyze the Auto Warranty data and suggest strategies for warranty analysis using textual data.
- Show how chaining text mining in a series of analyses on filtered documents can improve text categorization results.

3

Automotive Warranty Claims

Primary Analytic Objective: Identify claims that might suggest safety problems (TREAD Act).

Secondary Analytic Objective: Find useful or “natural” categories for claims.

Challenges:

- Documents have not been assigned to categories.
- Mechanics use nonstandard technical language and abbreviations.

Data:

- There are 34,939 automotive warranty claims.
- Documents are notes describing the problem or warranty action.

4

continued...

Automotive Warranty Claims

Metadata

5

Automotive Warranty Claims

Navagational screen has lines across it fiagnsoe faulty display replace navigatioal display unit.

Release for gas cap is not working.

Noise over bumps.

Check engine light on. Fuel too lean. Added 5 gallons good gas.

Brakes grinding, replaced front rotors and pads.

continued...

6

Automotive Warranty Claims

Data Preparation

- Remove proper names and other identifying information.
- Convert abbreviations.
- Correct misspellings.

7

Auto Warranty Synonyms

EMWS16.TextParsing_synonymDS

Term	parent	CATEGORY
ft	front	Noun
ft	frt	Prop
inop	inoperative	Noun
inoper	inoperative	Noun
inoperatice	inoperative	Noun
inoperativ	inoperative	Noun
inoperative	inoperative	Noun
irc	if	Noun
mldg	molding	Noun
catafaction	catisfaction	Noun
sebelt	seatbelt	Noun
steering	steering	Noun
stearing	steering	Adj
steareing	steering	Verb
steering	steering	Noun
steering	steering	Noun
stability	stability	Noun
streering	steering	Noun
streering	steering	Noun
susp	suspension	Noun
suspenion	suspension	Noun
suspension	suspension	Noun
suspension	suspension	Noun
ystem	system	Noun
windshield	windshield	Noun

Import OK Cancel

Synonyms are primarily misspelled terms.

8

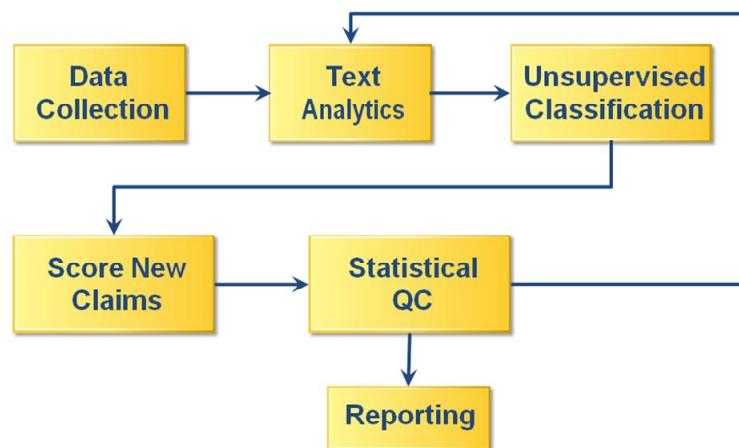
4.01 Multiple Choice Poll

How important do you think correcting misspellings and expanding abbreviations are for successful text mining?

- a. not important
- b. somewhat important
- c. very important
- d. critical for success

10

Warranty Text Analytic Workflow



11



Text Analytics for Warranty Analysis

This demonstration illustrates how to use SAS Text Miner nodes to analyze automotive warranty data.

An automotive manufacturer investigates warranty claims as required by the United States TREAD Act. In addition to looking for safety issues, the manufacturer also wants to address customer satisfaction and quality issues. A system has been created that classifies a warranty repair item into 11 categories as follows: (1) Body Electrical, (2) Body Mechanical, (3) Brakes, (4) Climate Control, (5) Drivetrain, (6) Engine Electrical, (7) Engine Mechanical, (8) Engine Performance, (9) Safety, (10) Suspension, and (11) Miscellaneous.

There are three supporting data sets for this analysis. A table of synonyms, **DMTXT.AutoSynonyms**, has been created using the Text Filter node in a preliminary analysis. The preliminary analysis also resulted in the creation of a stop list, **DMTXT.AWstop**. The steps in the preliminary analysis are sketched below. A third data set, **DMTXT.AutoSystemTopics**, contains custom topics created to facilitate automatic categorization of claims into topics.

The preliminary analysis is described below, but it can be skipped because the derived data sets already exist in the course folder.

1. Create an input data source for **LWDMTXT.WARRANTYCLAIMS** if necessary. The metadata shown earlier is reproduced below.

Name	Role	Level	Report	Order	Drop	Lower Li
ID	Nominal	No			No	
Mileage	Input	Interval	No		No	
RepairAmount	Input	Interval	No		No	
RepairDate	Time ID	Interval	No		No	
SalesDate	Time ID	Interval	No		No	
Text	Text	Nominal	No		No	

2. Create a diagram named Warranty Analysis. Drag the warranty data source into the diagram.

3. Create the process flow for a preliminary analysis as shown below.



The SAS code for frequency filtering is given below. The program source file is called `SCN_CreateStartStopList_AW.sas`.

```

/*----- SCN_CreateStartStopList_AW.sas -----*/
/*----- Create Start List using frequency filtering -----*/

%global LastParsing LastFilter TermData FTermData
      StartList MaxDocs MinDocs;

/*!!!!!! Edit the following 5 lines !!!!!/
%let StartList=DMTXT.AWstart;
%let StopList=DMTXT.AWstop;
%let MaxDocs=2000;
%let MinDocs=10;
%let AllowNumbers=N;

%let LastParsing= ;
%let LastFilter= ;

proc print data=&EM_IMPORT_DATA_EMINFO;
run;

proc sql noprint;
  select data into :LastFilter
    from &EM_IMPORT_DATA_EMINFO
    where key="LastTextFilter";
  select data into :LastParsing
    from &EM_IMPORT_DATA_EMINFO
    where key="LastTextParsing";
quit;

%put NOTE: Last SAS Text Parsing Node: &LastParsing;
%put NOTE: Last Text Filter Node: &LastFilter;

%let
TermData=%sysfunc(strip(&EM_LIB)).%sysfunc(strip(&LastParsing))_terms;
%let
FTermData=%sysfunc(strip(&EM_LIB)).%sysfunc(strip(&LastFilter))_terms_
data;

%EM_REGISTER(KEY=TERMS,TYPE=DATA);
%EM_REPORT(KEY=TERMS,VIEWTYPE=DATA);
%EM_REGISTER(KEY=TERMTRANS,TYPE=DATA);
%EM_REPORT(KEY=TERMS,VIEWTYPE=DATA);

```

```
%macro GetData();
%if %sysfunc(exist(&TermData)) & %sysfunc(exist(&FTermData))
%then %do;
%EM_REPORT(KEY=TERMS,VIEWTYPE=DATA,AUTODISPLAY=Y);
data &EM_USER_TERMS;
set &TermData;
run;
proc contents data=&EM_USER_TERMS;
run;

%EM_REPORT(KEY=TERMTRANS,VIEWTYPE=DATA,AUTODISPLAY=Y);
data &EM_USER_TERMTRANS;
set &FTermData(where=(Keep='Y' and
&MinDocs<=NumDocs<=&MaxDocs));
run;
proc contents data=&EM_USER_TERMTRANS;
run;
proc univariate data=&EM_USER_TERMTRANS;
var NumDocs Weight;
run;

proc sql;
create table &StartList as
select a.KEY, a.Role, a.TERM, b.Weight
from &EM_USER_TERMS a, &EM_USER_TERMTRANS b
where a.KEY=b.KEY and b.Weight>0;
quit;
%if (&AllowNumbers ne Y) %then %do;
data &StartList;
set &StartList;
if ('0'<=substr(TERM,1,1)<='9') then delete;
run;
%end;
proc sql;
create table &StopList as
select a.Term, a.Role
from &TermData a, &EM_USER_TERMTRANS b
where a.KEY=b.KEY and
a.KEY not in
(select c.KEY from &StartList c);
quit;
data &StartList;
set &StartList;
drop KEY;
run;
proc sort data=&StartList nodupkey;
by Term Role;
run;
proc contents data=&StartList;
run;
```

```

proc sort data=&StopList nodupkey;
  by Term Role;
run;
proc contents data=&StopList;
run;
%end;
%mend GetData;

%GetData();

```

Default settings are used for all of the Text Miner nodes, except that the default synonym list is omitted in favor of using no synonyms.

- Run the diagram through the Text Filter node. Open the Interactive Filter Viewer. Synonyms are created using the Term window. Select two terms, and then right-click and select **Treat As Synonyms**. An example using **innop** and **inop** appears below.

The screenshot shows the SAS Text Miner Interactive Filter Viewer. On the left, there is a tree view of terms under categories like 'inner fender line', 'innertie', 'innitial', and 'innop'. The term 'innop' is selected. A context menu is open over the 'innop' entry, listing options: 'Add Term to Search Expression', 'Treat as Synonyms', 'Remove Synonyms', 'Toggle KEEP', 'View Concept Links', 'Find', 'Repeat Find', 'Clear Selection', and 'Print...'. The main area displays a table titled 'Terms' with columns: TERM, FREQ, # DOCS, KEEP, WEIGHT, ROLE, and ATTRIBUTE. The table lists various terms with their respective values. The row for 'innop' is highlighted in blue.

	TERM	FREQ	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
+	injury	217	217	<input checked="" type="checkbox"/>	0.486	Noun	Alpha
	injurys	1	1	<input type="checkbox"/>	0.0	Noun	Alpha
	inlet	12	12	<input checked="" type="checkbox"/>	0.762	Noun	Alpha
	inline	1	1	<input type="checkbox"/>	0.0	Prop	Alpha
	inlocked	1	1	<input type="checkbox"/>	0.0	Noun	Alpha
	innaccurate	1	1	<input type="checkbox"/>	0.0	Noun	Alpha
	inner	84	82	<input checked="" type="checkbox"/>	0.58	Adj	Alpha
+	inner fender line	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
	innertie	1	1	<input type="checkbox"/>	0.0	Prop	Alpha
	innitial	1	1	<input type="checkbox"/>	0.0	Noun	Alpha
	innop				0.0	Noun	Alpha
	innovative				0.0	Adj	Alpha
+	innovative pro				0.0	Noun Group	Alpha
	infmration				0.0	Noun	Alpha
	inop				0.479	Noun	Alpha
	inop				0.846	Prop	Alpha
+	inop coin				0.0	Noun Group	Alpha
	inop fuse				0.0	Noun Group	Alpha
	inop light				0.0	Noun Group	Alpha
	inop repair				0.0	Noun Group	Alpha
	inop sop				0.0	Noun Group	Alpha

In this project, synonyms are primarily used to correct misspellings. You can use the synonym list that was created using the Text Filter node Check Spelling feature. The **DMTXT.AutoSynonyms** data set was derived using the Check Spelling feature and will be used as a synonym table for this analysis.

- You are now ready to run the SAS Code node (or nodes). The results have already been obtained, and the stop list is available as SAS table **DMTXT.AWstop**.

The formal analysis begins by attaching a Data Partition node to help validate the stability of the warranty categories that are derived. The project data has two variables with the role Time ID, so they can be used to examine results over time.

- Attach a **Data Partition** node to the original data source node in parallel to the preliminary analysis process flow. Use a train/validation/test split of 75/25/0. Optionally, if you did not perform the preliminary analysis, drag an input data source node into the diagram first, and then attach the **Data Partition** node.



When a date variable like repair date is populated with a valid date for each observation, you should consider partitioning your data using stratified random sampling applied to month or year depending on the frequency of the dates. This requires you to add a month or year variable to the data. While this is a recommended practice, it was not done for this example to simplify the demonstration for classroom purposes.

- Attach a Text Parsing node to the Data Partition node. Use the synonyms data set and the stop list table that were derived previously.
- Attach a Text Filter node to the Text Parsing node. Select the term weight **entropy**. This is the default term weight when no target variable is available. Run the Text Filter node.
- Open the Filter Viewer. The experts are aware of past issues related to automotive safety. For example, the document collection has entries from the 1990s describing air bag deployment when no collision occurred, resulting in injury to an occupant. This particular issue has been addressed and resolved. However, examining the older claims reveals keywords that might be useful for identifying new issues. If you examine the Term table, you can find terms that might be useful for uncovering safety issues. Select **Edit** ⇒ **Find** and type **injury**. This will take you to the portion of the Term table that contains the term *injury*.

TERM ▲	FREQ	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
injector pump	2	2	<input type="checkbox"/>	0.0	Noun Group	Alpha
injector rail	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
injector ring	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
+ injure	31	30	<input checked="" type="checkbox"/>	0.667	Verb	Alpha
+ injury	155	155	<input checked="" type="checkbox"/>	0.504	Noun	Alpha
injurys	1	1	<input type="checkbox"/>	0.0	Noun	Alpha
inlet	11	11	<input checked="" type="checkbox"/>	0.764	Noun	Alpha
inline	1	1	<input type="checkbox"/>	0.0	Prop	Alpha
inner	63	62	<input checked="" type="checkbox"/>	0.595	Adj	Alpha
+ inner fender line	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
innertie	1	1	<input type="checkbox"/>	0.0	Prop	Alpha
innovative	1	1	<input type="checkbox"/>	0.0	Adj	Alpha
+ innovative pro...	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
inofrmation	1	1	<input type="checkbox"/>	0.0	Noun	Alpha
+ inop	183	182	<input checked="" type="checkbox"/>	0.489	Noun	Alpha
inop fuse	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha

If you click + to the left of injury, you get the stemmed versions and the synonyms for injury.

<input checked="" type="checkbox"/> injury	155	155
<input type="checkbox"/> injury	41	41
<input type="checkbox"/> injuries	111	111
<input type="checkbox"/> iinjury	1	1
<input type="checkbox"/> inuries	1	1
<input type="checkbox"/> injuires	1	1

Spell-checking has done a good job of finding misspelled versions of injury.

10. Select **injury** by clicking on the cell containing the word, right-clicking, and selecting **Add Term to Search Expression**. The search window shows >**#injury**. This requests a search for *injury* and all of the synonyms of *injury*. Click **Apply**. The claims have been sorted by repair date, so you can scroll to the bottom of the documents window to see the most recent claims that mention an injury.

TEXT OF COMPLAINT	TEXTFILTER2_SNIPPET
While driving 60-70 mph air bags deployed on their own. There were no	... There were no injuries
While owner was opening the tailgate and the cables broke. The owners son	... son sustained minor injuries
The rear axle on the driver side broke while driving 65 mph. No injury was	... No injury was reported
While pulling into a driveway the dual air bags deployed inadvertently. No	... No injuries were reported .
When shutting the driver's side door the air bag deployed inadvertently. No	... No injuries reported
Owner was driving approximately 35 mph and suddenly the airbags deployed	... sustained a thumb injury
Loading a clothes dryer when tailgate cables broke, no injuries were noted.	... broke , no injuries were
While driving approximately 35 mph front axle literally broke. There were no	... There were no injuries
While driving at 60 mph suddenly the driver seat and door air bags deployed.	... owner sustained injuries
No injuries. Engine stops or stalls at various times. Nothing has been done to	... No injuries
Park to reverse injury.	... Park to reverse injury
Hood support struts can fail in cold weather without notice or warning. Could	... hand / arm injury
Power tailgate striker trapped finger, causing injury.	... finger , causing injury
Tension wheel to serpentine belt is corroding causing premature wearing out	... No accident no injury
The front driver seat airbag depolyed, and causes some injuries, on feb. 5,	... and causes some injuries ,
Badly rusted strut tower on drivers side front, needs attention or severe	... severe damage and injury
Trunk lid does not stay up - must be propped. Multitude of failures have	... failures have caused injuries
Sudden failure of parking brake which could lead to possible	... possible accident / injury /

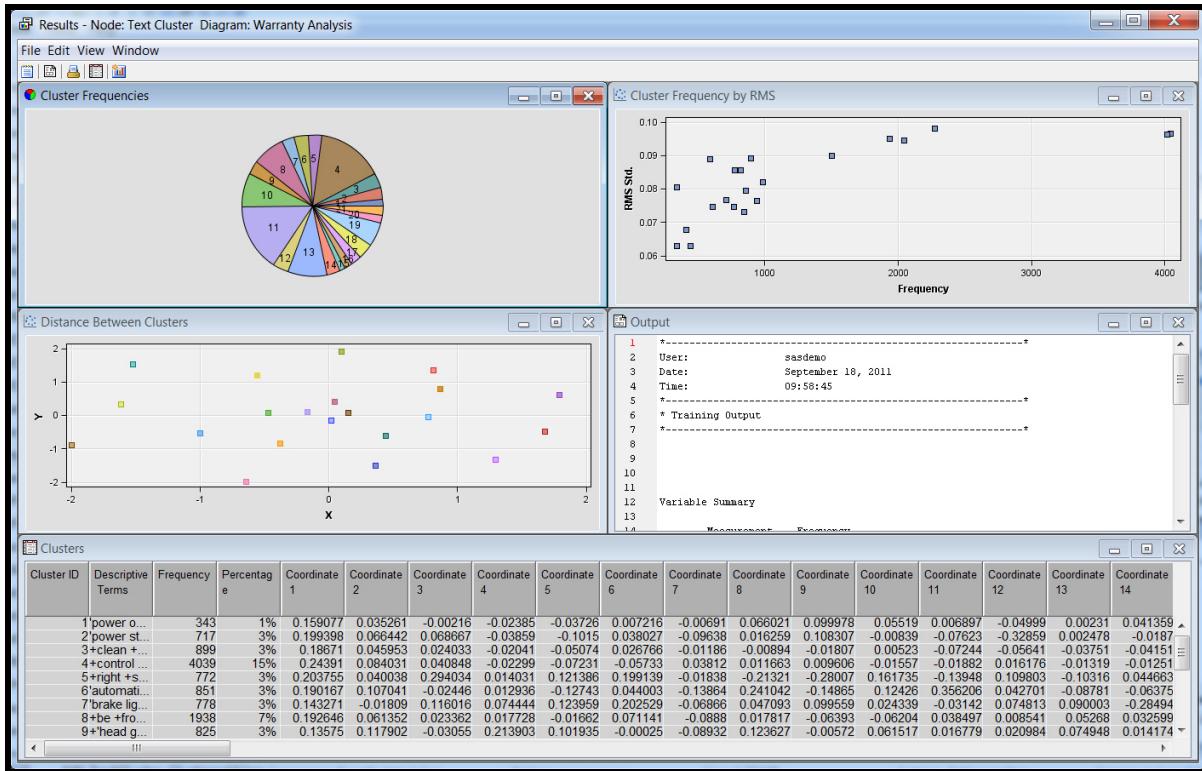
The last claim has a repair date in 2010, near the time the data was compiled. However, the previous claim is a 2008 claim, and the dates age quickly from there. For the TREAD Act, the older claims represent manufacturing issues that have already been resolved. However, the parking brake claim is worth investigating, even though no injury actually occurred.

When using historic data to develop an automated system, some data might seem outdated. Choosing a relevant date range is nontrivial. Even though a claim represents an issue that has been resolved, the words used to describe the issue can be relevant when processing new claims.

You can think of other terms that might be useful to identify possible safety issues. In a later exercise, you to try to find additional cases to review.

11. Attach a Text Cluster node to the Text Filter node. Use default settings, except set Maximum SVD Dimension to **200**. The value 200 is chosen to ensure that enough SVD dimensions are employed to capture most of the information in this document collection. Choosing the SVD dimension is more of an art than a science, and generally, small documents rarely need more than 100 dimensions. Because there are so many documents in this project, a larger value is supplied. Run the Text Cluster node.

12. Access the Text Cluster Results window.



Scrolling to the bottom of the Clusters window reveals that 21 clusters were derived. The Expectation-Maximization clustering algorithm was used (default), and the cubic clustering criterion selected 21 as the “optimum” number of clusters. To improve performance of subsequent runs, in case a predecessor node changes, it is common practice to change the Exact or Maximum Number property to **Exact** and to change the Number of Clusters property to the derived number, **21**. If the node must be re-run in the future, and a predecessor node has changed, then only one set of 21 clusters will be derived. With the maximum of 40 used, then 39 (2 through 40) sets of clusters will be derived, with a possible significant increase in processing time.

Another common practice is to set the cluster number based on domain knowledge. For example, you might request 11 clusters hoping to get clusters to conform to the 11 automotive systems. Unfortunately, if the derived clusters do not correlate well with the known systems, then you will probably want to revert back to using the cubic clustering criterion.

If you scroll to the right in the Clusters table, you see that 98 coordinates were used. These are the SVD columns. Thus, by specifying 200 dimensions and low resolution, you obtained fewer than 100 dimensions. Note that the heuristic employed would derive a smaller number of dimensions if the Max SVD Dimensions was set at 100. You can get the same results by setting SVD Resolution to **High** and Max SVD Dimensions to **98**. If the node has to be re-run, this will give slightly better performance than setting SVD Resolution to **Low** and Max SVD Dimensions to **200**.

13. Examine the descriptive terms for the clusters.

Cluster ID	Descriptive Terms	Frequency	Percentage	Coordinate 1	Coordinate 2
1	'power outlet' +inop +power fuse loss lost outlet stalled	343	1%	0.159077	0.0352
2	'power steering' +control +power +steer column pinion rack steering	717	3%	0.199398	0.0664
3	+clean +pull +vehicle align car left pulling pulls	899	3%	0.18671	0.0459
4	+control +fire +gear +pedal +stop +turn +vehicle driving	4039	15%	0.24391	0.0840
5	+right +separation +sidewall +tire +tread out separated tires	772	3%	0.203755	0.0400
6	'automatic transmission' +'transmission failure' +automatic +gear +transmission failure fluid gears	851	3%	0.190167	0.1070
7	'brake light' 'rear spoiler' +'rear window' +'third brake light' +rear +right +seal +window	778	3%	0.143271	-0.0180
8	+be +front +warning had has miles problem problems	1938	7%	0.192646	0.0613
9	'head gasket' +coolant +engine +head +leak +manifold +seal cracked	825	3%	0.13575	0.1179
10	+airbag +d+ +driver +not +passenger +properly +retract +speedometer	2044	8%	0.241701	-0.0880
11	+driver +driving +fire +passenger +window air broke car	4019	15%	0.256797	-0.0318
12	+fuel gauge' +'fuel pump' +fuel tank' +gas tank' +fuel +gauge +leak +pump	986	4%	0.186676	0.1353
13	+inoperative +out +speedometer cold from heater hood rusted	2273	9%	0.130747	0.0252
14	'head gasket' +'check engine light' +'engine head gasket' +'check + coolant +engine failure light	858	3%	0.219195	0.1665
15	'recall notice' +available +notice +repair owner parts performed recall	341	1%	0.219912	0.0211
16	'windshield wiper' 'windshield wipers' + poor visibility' +motor +rain +windshield +wiper +work	448	2%	0.139857	-0.0294
17	+loud noise' +loud grinding heard made make makes making	611	2%	0.182831	0.0788
18	+seat belt' +belt +driver +floor +not +retract +seat broke	945	4%	0.277611	-0.4504
19	'extended stopping distance' +control +floor +pedal +stop +system abs applying	1502	6%	0.186018	0.0920
20	+bulb +headlight +lens +night headlights lights low switch	414	2%	0.15413	0.0243
21	'center console' +rattle +side broken center clips console installed	592	2%	0.153658	-0.0794

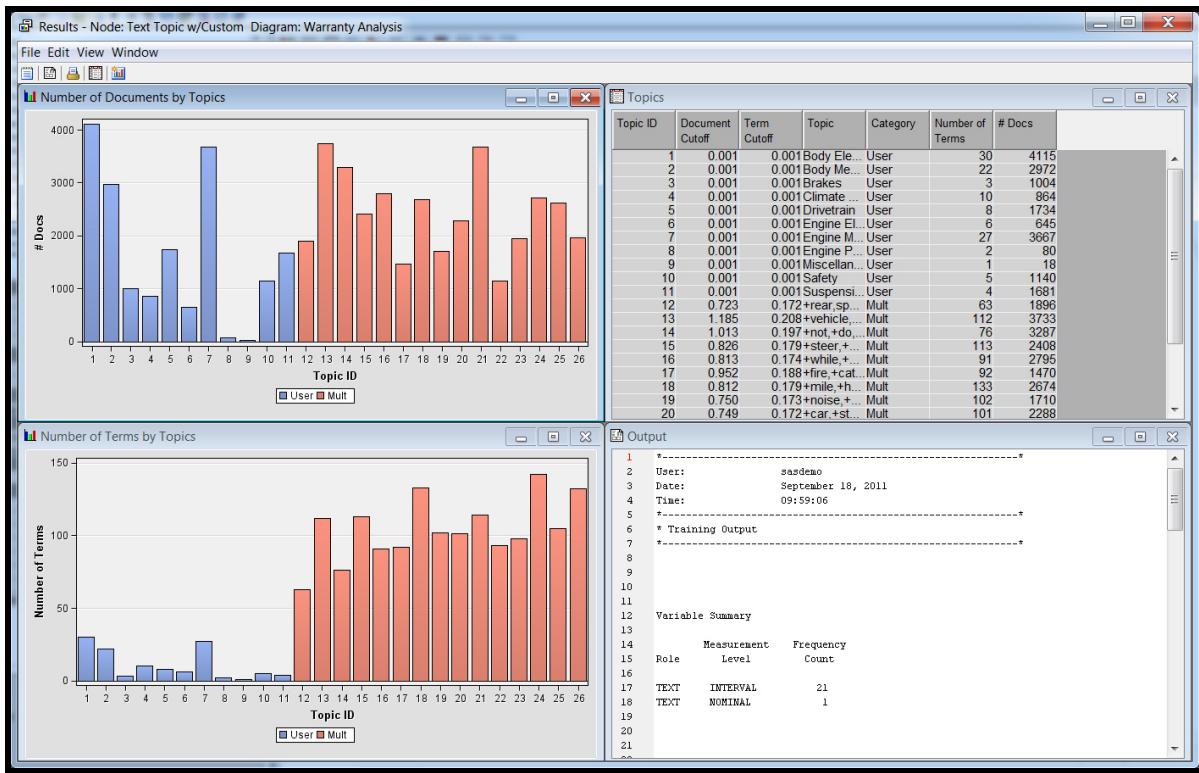
A quick scan of the clusters reveals that some clusters seem to match automotive systems, but some do not. For example, cluster 2 appears to be a pure suspension cluster, and cluster 6 appears to be a pure drivetrain cluster. Cluster 1 seems to be a mixture of engine electrical, body electrical, and engine performance.

Even if the clusters cannot be used exactly as they are derived, you can extract the “pure” clusters and use alternative methods on the remaining documents to try to get better separation and more interpretable results.

The SVD values will be part of the exported data. You can use the SVD values as inputs to other SAS Enterprise Miner nodes, such as the Cluster node or the SOM/Kohonen node. Unfortunately, performing unsupervised learning outside of the Text Miner nodes will make it more difficult to interpret the results relative to the document collection.

14. Add a Text Topic node to the Text Cluster node. Note that you could add the Text Topic node to the Text Filter node in parallel to the Text Cluster node. Adding the Text Topic node in series allows all derived columns to be exported in the same data set, since SAS Enterprise Miner nodes pass through data elements even if they are not used by the current node.
15. In the Properties panel of the Text Topic node, set the Number of Multi-term Topics property to **15**. While choosing 11 topics might be recommended, the number 15 is employed to investigate what the “natural” topics are in the collection. While you hope that the natural topics coincide with the automotive system topics, it is possible that the automotive system categorization is flawed or incomplete.
16. You also want to take advantage of domain knowledge. A table of custom topics has been created to try to facilitate automatic identification of automotive systems. As mentioned above, this table resides in **DMTXT.AutoSystemTopics**. Specify this table for the property User Topics. Run the Text Topic node.

17. Open the Results window.



The 11 custom topics are identified in the two bar charts using a blue color, and the 15 derived topics are identified using a red color. (In gray scale, the bars might be difficult to distinguish. The 11 custom topics are represented by the 11 rightmost bars.)

Custom topics 8 and 9 are relatively sparse. These are topics Engine Performance and Miscellaneous. Either there are few documents exhibiting these topics, or the custom definitions are inadequate to identify relevant documents.

Are there any documents that do not appear to have any of the custom topics? The Number of Documents by Topic suggests that many documents are not associated with any of the custom topics. To answer this question, a SAS Code node can be used. The finding is that 44.2% of the documents exhibit none of the custom topics. Evidence suggests that the custom topics need to be improved. Further analysis shows that 28.4% of the documents exhibit none of the derived topics. In addition, 12.6% of the documents do not exhibit any identified topics, custom or derived. Increasing the number of derived topics might reduce the percentage. Another possibility is that many documents contain only stop words, and thus are treated as being identical. Because these documents will not have any of the terms used to derive topics, adding derived topics or enhancing custom topics will not affect the classification of these documents.

18. Close the Results window.

19. Open the Text Topic Viewer.

The screenshot shows the 'Interactive Topic Viewer' application window with three main tabs: 'Topics', 'Terms', and 'Documents'.

- Topics:** This tab displays a table of topics with columns: Topic, Category, Term Cutoff, Document Cutoff, and N. The topics listed are 'Body Electrical', 'Body Mechanical', 'Brakes', and 'Climate Control'. All topics are categorized as 'User'.

Topic	Category	Term Cutoff	Document Cutoff	N
Body Electrical	User	0.001	0.001	30
Body Mechanical	User	0.001	0.001	22
Brakes	User	0.001	0.001	3
Climate Control	User	0.001	0.001	10

- Terms:** This tab displays a table of terms with columns: Topic Weight, +, Term, Role, # Docs, and Freq. The terms listed are 'light' (Noun), 'window' (Noun), 'windshield' (Noun), and 'wiper' (Noun). All terms have a Topic Weight of 1 and a '+' sign.

Topic Weight	+	Term	Role	# Docs	Freq
1	+	light	Noun	1009	1043
1	+	window	Noun	585	667
1	+	windshield	Noun	520	555
1	+	wiper	Noun	401	438

- Documents:** This tab displays a table of documents with columns: Topic Weight, Text of Complaint, Claim ID, Mileage at Repair, Repair Amount, and Repair Date. The documents listed are:

Topic Weight	Text of Complaint	Claim ID	Mileage at Repair	Repair Amount	Repair Date
1.455	Electrical system under	1689362601	96	49.0	1999-04-09
1.371	CD changer rejects cds	9100338723	11947	241.34	2008-10-02
1.23	Check engine light,	5049595532	1842	318.0	2003-06-23
1.229	CD/cassette clock radio	7224088458	6	15.0	2005-05-23
1.195	Check engine light on	0045590959	329	42.64	1996-08-19

The results for the custom topics look very promising. While the documents have no class labels for categorization, examination of the selected documents for each of the custom topics suggests that the precision statistic is probably high. Without extensive examination, it is difficult to determine how many documents were not selected that should have been selected, so you have no insight into the recall statistic.

20. Scroll down the Topics window until you see the topic **+fire,+catch,+vehicle,+park,+start**. Select this topic, right-click, and select **Select Current Topic**.

The screenshot shows the 'Interactive Topic Viewer' application window. It has three main sections: 'Topics', 'Terms', and 'Documents'.

Topics:

Topic	Category	Term Cutoff	Document Cutoff	Number of Terms	# Docs
+steer, +wheel, +power, +be, +lock	Mult	0.179	0.826	113	2408
+while, +driving, driving, mph, +headlight	Mult	0.174	0.813	91	2795
+fire, +catch, +vehicle, +park, +start	Mult	0.188	0.952	92	1470
+mile, +have, +rotor, +be, +time	Mult	0.179	0.812	133	2674
+noise, +make, +turn, +loud, +engine	Mult	0.173	0.75	102	1710
+car, +truck, +bus, +boat, +train	Mult	0.173	0.740	101	2289

Terms:

Topic Weight	+	Term	Role	# Docs	Freq
6.636	+	fire	Noun	773	841
4.864	+	catch	Verb	427	431
3.083	+	vehicle	Noun	5190	5782
2.505	+	park	Verb	432	448
1.833	+	start	Verb	843	879
1.600	+	ha	Noun	5719	6622

Documents:

Topic Weight	Text of Complaint	Claim ID	Mileage at Repair	Repair Amount	RepairDate	SalesDate
7.11	While parked vehicle caught on fire. it was	2641950176	2776	97.0	2000-04-17	2000-03-16
6.544	While vehicle was parked it caught on fire in the	1154797412	16956	552.0	1997-11-26	1997-09-20
6.413	Vehicle was parked, driver started vehicle up and	0502514657	919	38.0	1997-03-20	1993-11-07
6.105	Vehicle caught on fire while it was parked and	9461841056	270	52.0	2009-05-27	2008-12-25
6.074	Vehicle caught fire in the engine while driving. The	1592075364	4506	153.0	1999-01-07	1998-12-24

The chosen derived topic is a candidate for a TREAD Act topic. You can modify this derived topic to make it a custom topic.

21. To preserve the initial derived topics, make a copy of the Text Topic node. Attach it to the Text Cluster node in parallel with the original Text Topic node. Run the node. Open the Text Topic Viewer. Select topic **+fire, +catch, +vehicle, +park, +start** as the current topic.
22. First, remove terms that do not appear to be relevant. To do so, select the term, and give it a weight of zero (0). Irrelevant terms are *vehicle*, *be*, *while* (noun), *then*, *go*, *no*, *being*, *owner*, *have*, and *while* (verb). Second, add terms that might indicate a TREAD Act safety concern. Determining weights can be tricky, so for convenience, just assign a weight of 1. To add a term, use the **Edit** \Rightarrow **Find Term** feature. Terms to add are *injury*, *injure*, *hurt*, *risk*, *burning*, *burn* (noun), and *burn* (verb). There are many additional terms that could be used, but for educational purposes, the small set of words will be selected. When you finish making the weight modifications, change the topic name to TREAD Act. Then select **Recalculate**.

Interactive Topic Viewer

Topics

Topic	Category	Term Cutoff	Document Cutoff	N
+steer, +wheel, +power, +be, +lock	Mult	0.179	0.826	113
+while, +driving, driving, mph, +headlight	Mult	0.174	0.813	91
TREAD Act	User	0.188	0.952	89
+mile, +have, +rotor, +be, +time	Mult	0.179	0.812	133
+noise, +make, +turn, +loud, +engine	Mult	0.173	0.75	102
+car, +start, +have, +not, +turn	Mult	0.172	0.749	101
+owner +he +have +recall +problem	Mult	0.180	0.943	114

Terms

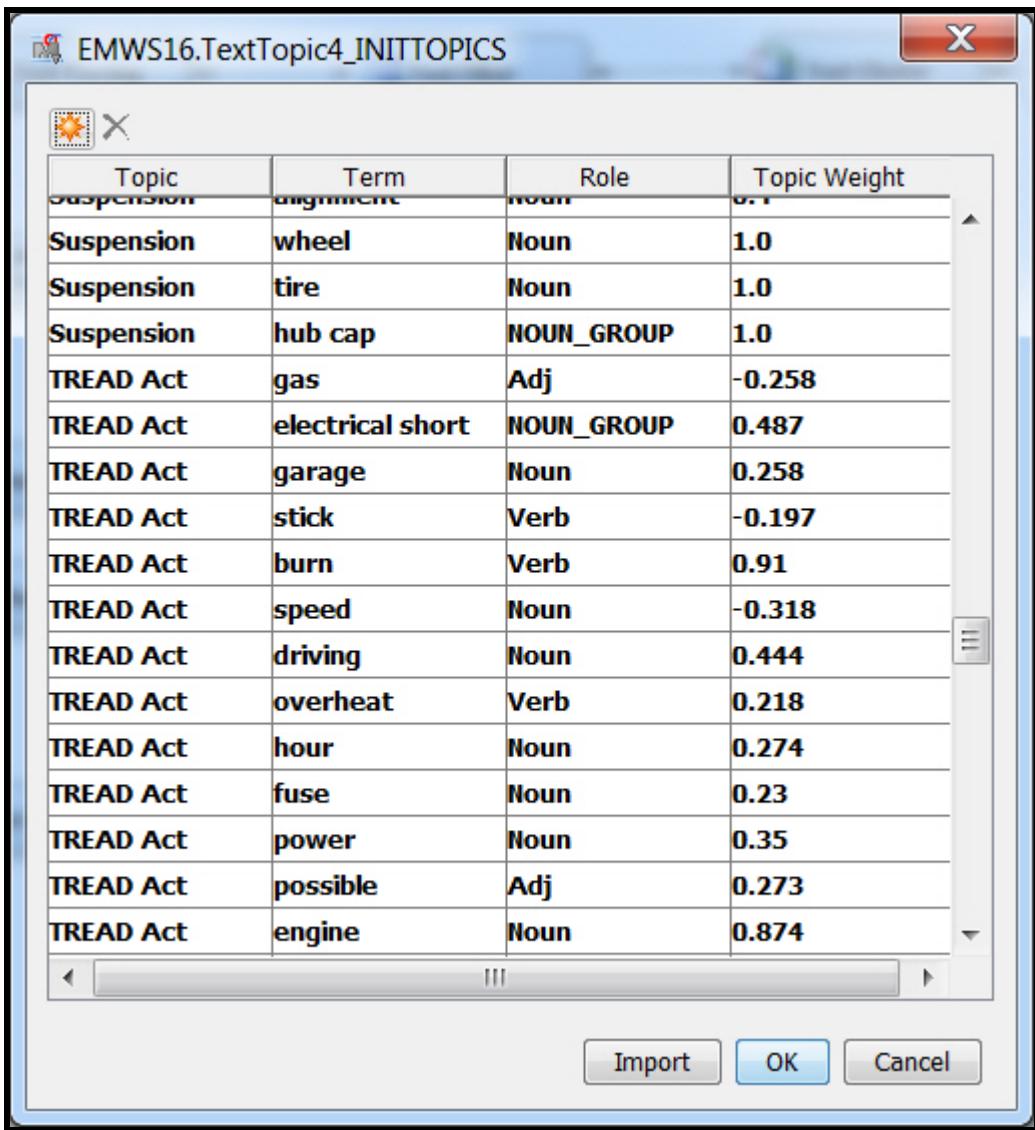
Topic Weight	+	Term	Role	# Docs	Freq
1	+	accident	Noun	273	274
1	+	injury	Noun	155	155
1		burning	Adj	33	34
1	+	injure	Verb	30	31
1	+	burn	Noun	17	17
1	+	hurt	Verb	9	10
1		rick	Noun	6	6

Documents

Topic Weight	Text of Complaint	Claim ID	Mileage at Repair	Repair Amount	Repair Date
6.169	While parked vehicle	2641950176	2776	97.0	2000-04-17
5.738	While vehicle was	1154797412	16956	552.0	1997-11-26
5.433	The while car was	3849144831	29	1086.0	2002-12-26
5.389	While parked vehicle	6659715523	272	56.0	2004-09-08
5.361	Vehicle was parked,	0502514657	919	38.0	1997-03-20
5.219	Vehicle caught fire.	6119387753	524	40.0	2004-02-26
5.183	While parked the	0509100160	7699	255.0	1997-03-20
5.070	Vehicle caught fire in	1502075261	1506	152.0	1999-01-07

The documents are re-scored based on the new term weights. Scanning the highest scoring documents suggests that the TREAD Act topic finder is somewhat successful.

23. Close the Text Topic Viewer, and select Yes to save the changes. When you access the User Topics window, you will see that the TREAD Act has been added as a custom topic. It has *not* been added to the original custom topic table. Instead, it has been added to a new table indicated at the top of the view window.

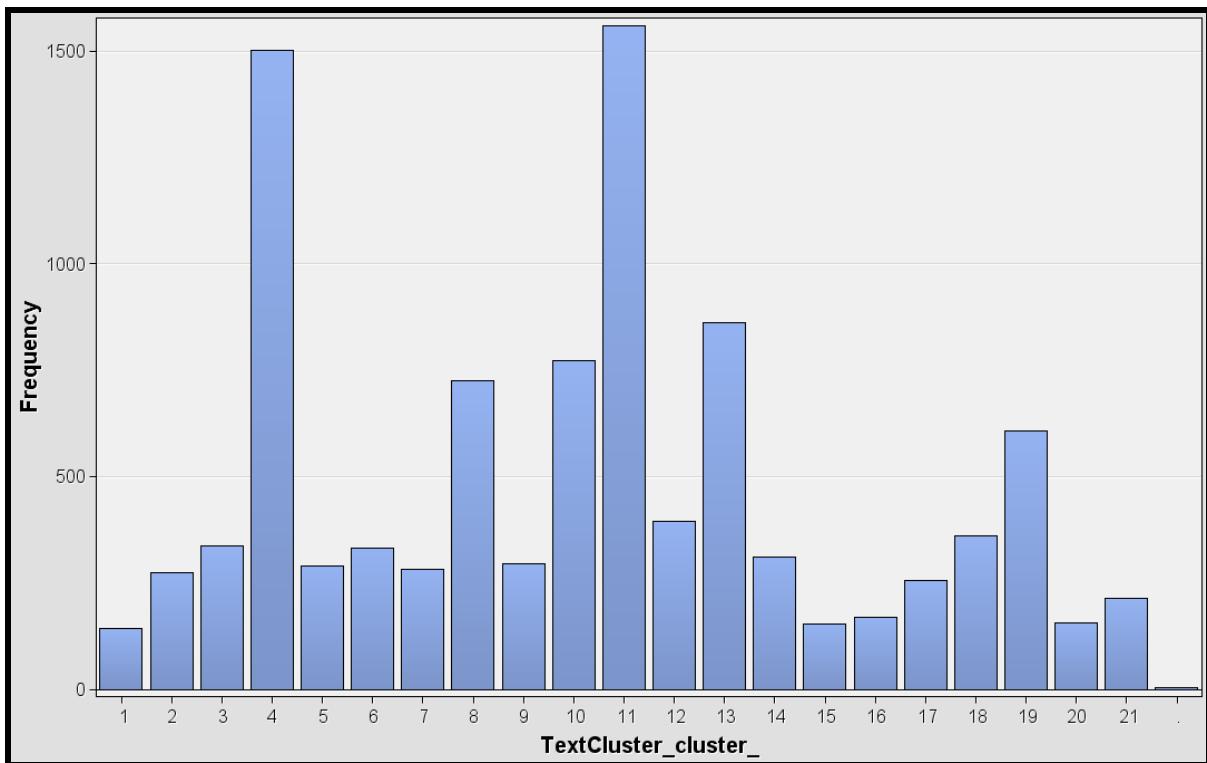


The screenshot shows a Windows dialog box titled "EMWS16.TextTopic4_INITTOPICS". The main area is a table with four columns: "Topic", "Term", "Role", and "Topic Weight". The table contains 16 rows of data. The "Topic" column alternates between "Suspension" and "TREAD Act". The "Term" column lists various automotive terms. The "Role" column indicates parts of speech or roles like "Noun", "Verb", "Adj", and "NOUN_GROUP". The "Topic Weight" column shows numerical values ranging from -0.318 to 0.874. At the bottom of the dialog box are three buttons: "Import", "OK", and "Cancel".

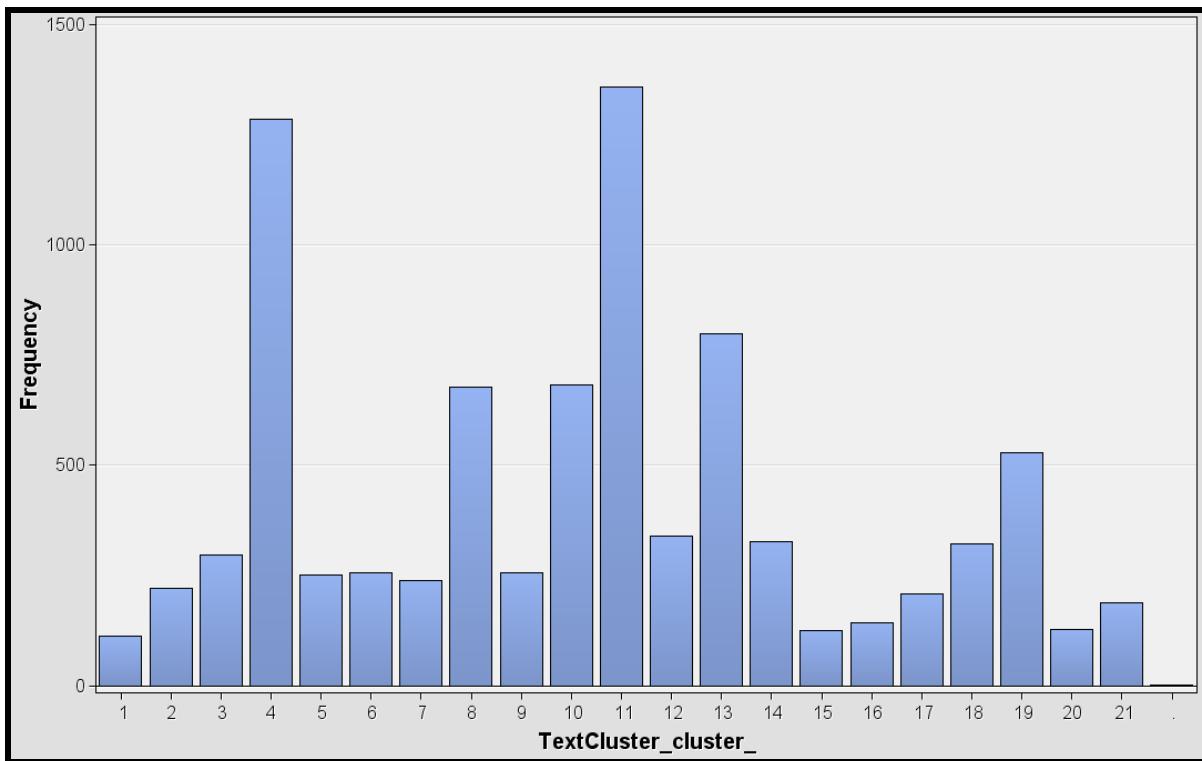
Topic	Term	Role	Topic Weight
Suspension	alignment	Noun	0.1
Suspension	wheel	Noun	1.0
Suspension	tire	Noun	1.0
Suspension	hub cap	NOUN_GROUP	1.0
TREAD Act	gas	Adj	-0.258
TREAD Act	electrical short	NOUN_GROUP	0.487
TREAD Act	garage	Noun	0.258
TREAD Act	stick	Verb	-0.197
TREAD Act	burn	Verb	0.91
TREAD Act	speed	Noun	-0.318
TREAD Act	driving	Noun	0.444
TREAD Act	overheat	Verb	0.218
TREAD Act	hour	Noun	0.274
TREAD Act	fuse	Noun	0.23
TREAD Act	power	Noun	0.35
TREAD Act	possible	Adj	0.273
TREAD Act	engine	Noun	0.874

24. Keep in mind that you are pursuing two analytic objectives: (1) automated classification for the TREAD Act, and (2) automatic classification into 11 automotive systems. The Text Topic step helped with both objectives. You will now focus on the second objective – namely, automatically scoring new claims into one of 11 categories. One additional strategy will be illustrated to complete this demonstration. This strategy is a “divide and conquer” strategy. The pure clusters arising from each clustering step will be set aside, and the remaining documents will be analyzed using different properties to try to improve the quality of the clusters that are derived.
25. Before proceeding, you need to verify that the good clusters are also stable. This was the purpose of the Data Partition node. The results that you have seen so far have all been for the training data. For the Text Cluster node, select the property **Exported Data**. Select the **TRAIN** data, and select **Explore**.

26. When the explore window appears, select the plot wizard, and select a bar chart. Assign the data role **Category** to the variable **TextCluster_cluster_**. Click **Finish**.



27. Repeat the steps, except select the **VALIDATE** data.

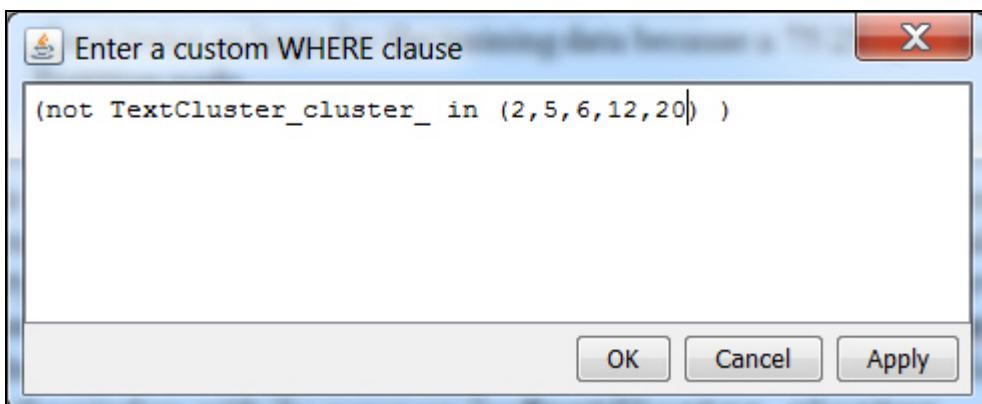


The distribution of the clusters is very similar, indicating stability of the clustering process. You should also consider checking stability as a function of time, but this would require use of a SAS Code node.

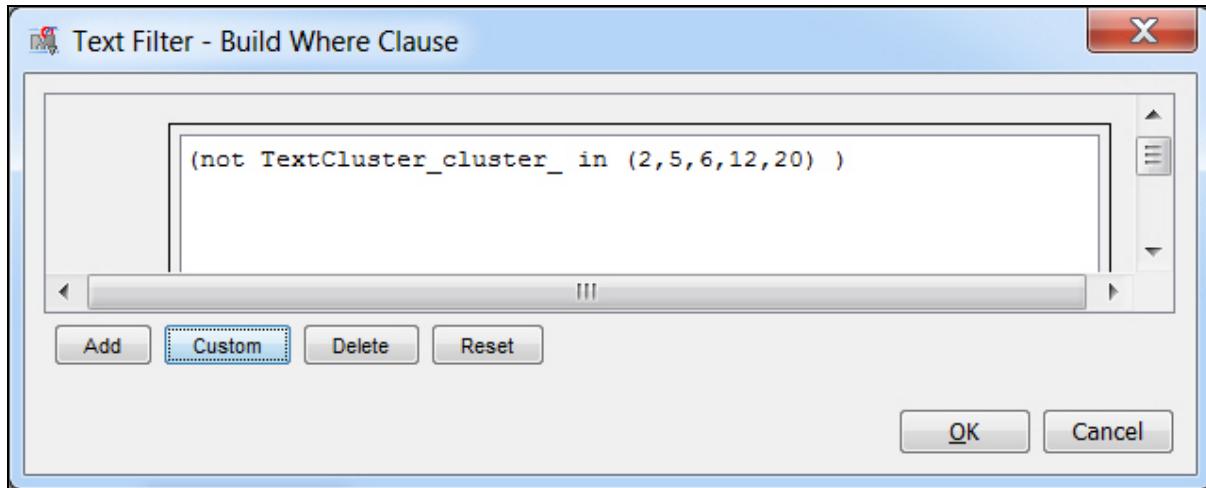


The frequencies in the two bar charts are similar because the training data is sampled to reduce it to 10,000 observations. Otherwise, you would expect the frequency counts to be about twice as large for the training data because a 75/25 split was employed by the Data Partition node.

28. Attach a Text Filter node to the Text Cluster node. Change the Term Weight property to **Inverse Document Frequency**. From the previous cluster analysis, you conclude that clusters 2, 5, 6, 12, and 20 are pure clusters, so you want to remove the documents in these clusters from analysis. In the Text Filter node Properties panel, select the **Subset Documents** property. Click the **Custom** button. Type the WHERE clause as shown below.

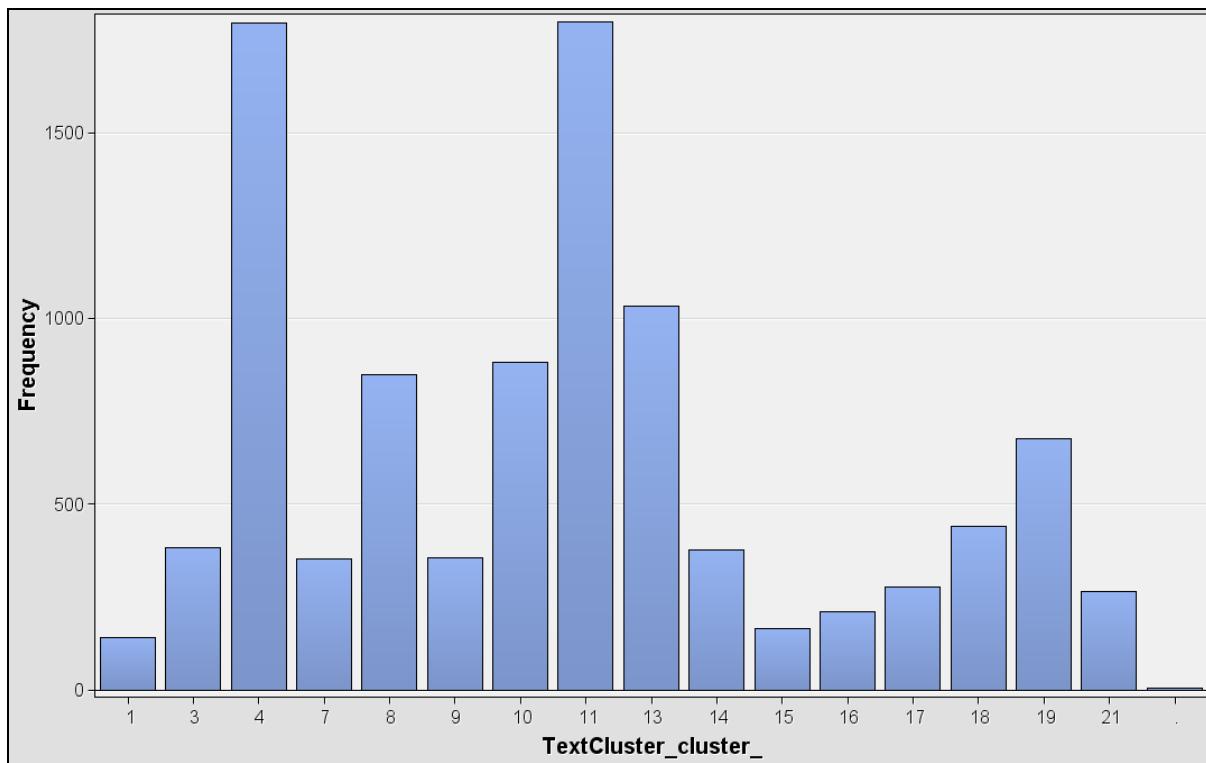


29. Click **OK**.



30. Click **OK**, and run the Text Filter node.

31. For the Text Filter node, select **Exported Data**, and select the data set **TRAIN**. Click the **Explore** button. Construct a bar chart for the variable **TextCluster_cluster_**.



You can see that the WHERE clause eliminated clusters 2, 5, 6, 12, and 20.

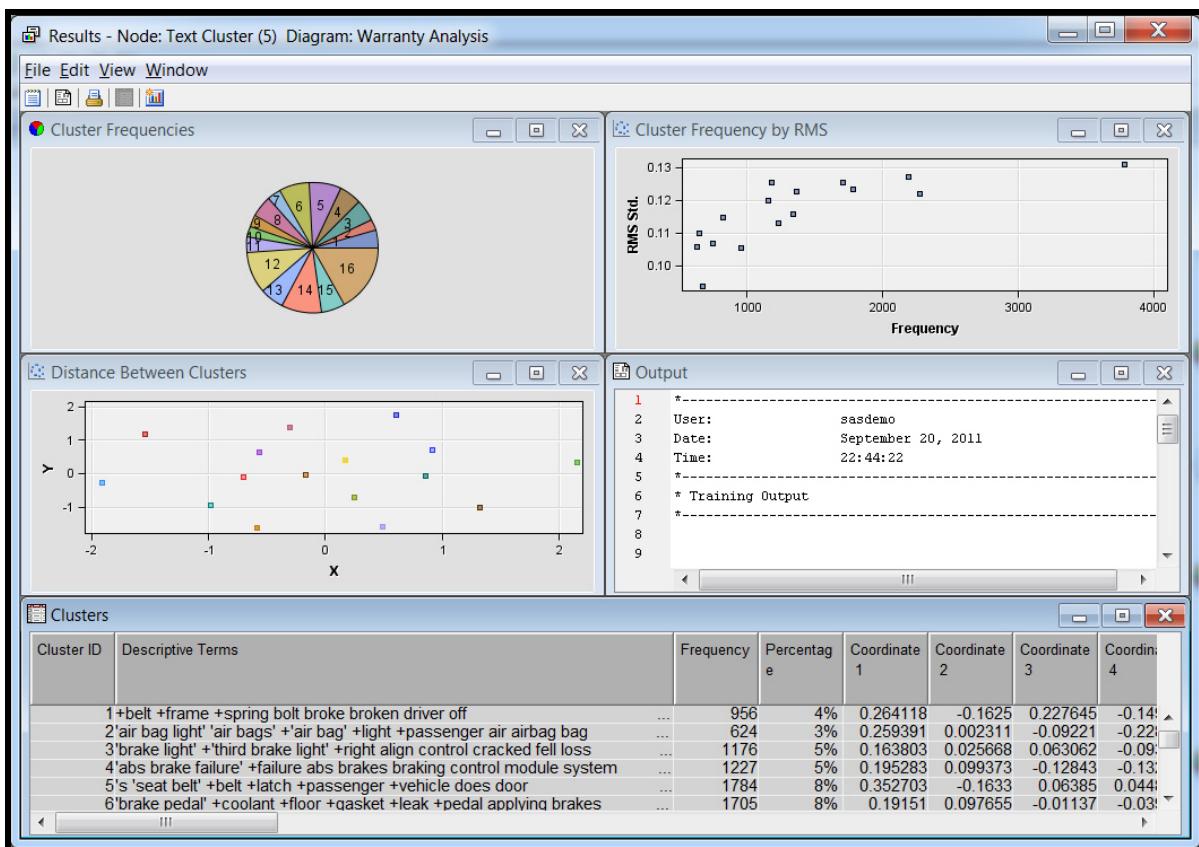
32. Before you continue, identify the changes that could result in better clusters.

- Some typical stop words, such as *be* and *had*, seem to have been left off of the stop list. Perhaps the frequency filtering stop list should be merged with the default stop list. (Text Parsing node)
- Entity identification was turned off in the first analysis. Could the identification of entities (person names, addresses, company names, locations, and so on) enhance clustering?

- The first analysis used entropy term weights. IDF terms weights could be used instead. (Text Filter node)
- The Text Cluster node chose to use 98 SVD dimensions. Could forcing a higher (or lower) number improve results?
- The Text Cluster node used Expectation-Maximization clustering. Perhaps hierarchical clustering would produce better results. You could also consider exporting the data and using a Cluster node or SOM/Kohonen (neural network) node.
- The Text Cluster node found 21 clusters using the cubic clustering criterion. Should you override this value and pick a larger or smaller number?

You cannot have two or more Text Parsing nodes connected in series in a process flow, so changes that affect the Text Parsing node will have to be tried in a separate process flow.

33. To illustrate actions that could improve results, change the term weight to **Inverse Document Frequency** (already changed above), and change the Text Cluster options to force exactly 16 clusters. Attach a Text Cluster node to the Text Filter node. Set the Exact or Maximum Number property to **Exact**, and set the Number of Clusters property to **16**. Run the Text Cluster node. Open the Results window.



The number of SVD dimensions chosen is 50. The clusters window shows that only a few additional “pure” clusters have been obtained.

Cluster ID	Descriptive Terms	Frequency
1	+belt +frame +spring bolt broke broken driver off	956
2	'air bag light' 'air bags' +'air bag' +light +passenger air airbag bag	624
3	'brake light' +'third brake light' +right align control cracked fell loss	1176
4	'abs brake failure' +failure abs brakes braking control module system	1227
5	'seat belt' +belt +latch +passenger +vehicle does door	1784
6	'brake pedal' +coolant +floor +gasket +leak +pedal applying brakes	1705
7	'windshield wipers' +'poor visibility' +motor +rain +speedometer +windshield +wiper intermittently	746
8	'rear spoiler' +'rear window' +fuse +inop +rear +wiper cracked driver	1152
9	+'spark plug' +blow +out +plug +spark +tire air blew	644
10	'brake pads' +prematurely +wear front miles only pads rotors	669
11	+available +notice +problem +recall owner parts problems received	819
12	's +be +car +have +vehicle had has is	2271
13	'steering wheel' +front +left +loud +noise +right end left	1367
14	+cause +driving +highway +vehicle +warning driving mph no	2194
15	+'check engine light' +check +fire +light +vehicle caught engine on	1338
16	+light +passenger airbag does driver n't not on	3783

These results suggest that clusters 2, 3, and 10 might be “pure” clusters. Chapter 1 discussed the signal versus noise problem. Perhaps the remaining documents are dominated by noise. One way to investigate is to employ a variety of clustering algorithms to see if there is consistency in the results. If the derived clusters are similar, and they are difficult to assign to a known category, then perhaps the documents are too noisy for additional classification. On the other hand, if derived clusters do not have much overlap, then perhaps there is a set of clusters that are of sufficient quality to aid in automation.

The goal of the demonstration was to reveal strategies for automating the classification of automotive warranty claims. Time constraints prevent a complete analysis.

If you decide to try to use one of the SAS Enterprise Miner unsupervised learning nodes, the Cluster node or the SOM/Kohonen (neural network) node, then you will have the added burden of associating terms with clusters in the way that the Text Cluster node assigns descriptive terms. Currently, assigning descriptive terms to derived segments requires an excess of manual intervention in the form of SAS Code nodes.

Here is the complete diagram for the above analysis:



The diagram uses only the training data and does not have a Data Partition node.

4.2 Processing and Categorizing Documents

Objectives

- Propose different scenarios that benefit from text analytics and text categorization.
- Use the Workers' Compensation Insurance data to illustrate how clustering can be used for preliminary predictive modeling when a target variable is not available.

20

Applications of Text Categorization

- Filtering e-mail
- Routing calls for a call center
- Classifying a news article
- Classifying a Web page
- Classifying a technical library based on similar content

21

4.04 Multiple Answer Poll

Which of the following text categorization problems are you interested in solving? Select all that apply.

- a. e-mail filtering
- b. call center routing
- c. news article classification
- d. Web page classification
- e. technical articles classification
- f. another problem not listed above
- g. I have no text categorization problems.

23

Text categorization plays an important role in fraud detection as a precursor to predictive modeling.

Predicting Fraud

- *Fraud* is intentional deception for personal gain.
- Credit card fraud provides a monetary gain through illegal purchases obtained from identity theft.
- Insurance fraud provides a monetary gain through benefits received from falsifying insurance claims.
- Warranty fraud provides monetary gain through compensation for services that are never performed or for services that are unnecessary.

The unique nature of fraud requires a unique approach to predictive modeling for each type of fraud.

24

Fraud permeates the business world. Credit card fraud is a part of the larger investigation area of transaction fraud, which has skyrocketed with the growth of e-commerce. A fraudulent transaction can be more than just credit card fraud. For example, a person opens an online store through one of the many companies that specialize in this, such as Amazon.com or eBay. The person sets up fake customer accounts, and then creates fake transactions. As a fake customer, she reports that an order that was paid for was never received. Through loopholes in the transaction guarantee, which could be through the parent e-commerce company, the credit card company, or another third party, the customer receives a refund of a payment that was never made, usually because the fake business owner has closed shop.

4.05 Poll

Insurance Company X contracts with your consulting firm to derive a fraud detection model. You are supplied with 800 paper reports that describe successful fraud investigations over the past five years. After constructing a data table based on the contents of the reports, you perform a quick exploratory analysis that reveals that 603 of the 800 fraud cases involve a soft-tissue back injury. Do you agree with the following statement?

The data is ***insufficient*** to suggest that soft-tissue back injuries are indicative of fraud.

- Yes
- No

26

Insurance Fraud versus Credit Card Fraud

Insurance	Credit Card
Fraud must be diagnosed or detected.	Fraud is established quickly based on billings and payments.
Many fraud cases are never discovered.	All cases are detected.
Predictive model false negatives are never quantified.	False positives and false negatives are easily quantified.
Data on fraud is limited and usually of poor quality.	Data is primarily transactional, and is of high quality and easily obtained.

28

Comparing insurance fraud to credit card fraud reveals how domain knowledge is important to the success of a fraud detection project. You cannot just stuff the data into a data mining software application and expect great results to come flowing out.

Workers' Compensation Insurance Fraud

Claimant Fraud: A worker files a false claim.

- The injury is not work-related.
- The severity of an injury is exaggerated.
- An accident never occurred.

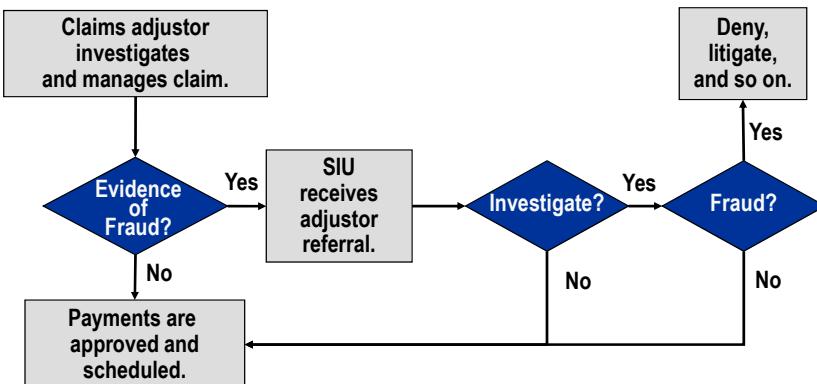
Provider Fraud: A health care or rehabilitation service provider bills an insurance company for services that were not provided or were unnecessary.

Internal Fraud: An insider files a false claim and routes payments to herself or a colleague.

29

Insurance fraud is often separated into three distinct categories. Knowledge of each category can provide insight for choosing analytical techniques.

Generic Insurance Fraud Process



30

The process flow for fraud reveals key steps where analytics can be beneficial.

A Sample of Typical Claim Fields

Field	Description
Claim ID	Primary index for extracting claimant, policy, and transactional data from the database
Birth Date	Claimant birth date
Injury Date	Date of injury
Employment Location	Location of employment (ZIP code)
Accident Location	Location of accident (ZIP code)
Gender	F/M
Body-Part Code	Part of body injured
Accident Code	Accident code (slip, fall, etc.)
Nature Code	Nature of injury (laceration, contusion, etc.)
Industry Code	Industry listed on insured company's policy
Occupation Code	Occupation
Prior Injury	Y/N for related prior injury
Risk Factors	Multiple Y/N fields for risk factors such as anemia, diabetes, etc.
ICD9 or CPT Codes	Standard medical procedure, disease, and condition codes, usually stored in multiple fields for primary, secondary, etc., codes
Medical Codes	Multiple Y/N fields for medical fields such as ER, inpatient hospital, outpatient hospital, radiology, physical therapy, etc.
Rehab Codes	Multiple rehabilitation fields such as vocational rehabilitation, education, and re-training
Adjustor Notes	Free-format text field with adjustor notes about the case

31

Insurance is a heavily regulated industry, so you can expect similarities in data that is collected. However, companies often differ in how they code the attributes related to a claim. International Classification of Disease (ICD) codes represent a universal standard for classifying insurance claims. Current Procedural Terminology (CPT) codes provide an alternative classification system. Either system has thousands of unique codes. ICD9 is a standard still employed in the U.S., but ICD10 is available, and ICD11 is expected to be released in 2015. The name ICD9 refers to ICD codes, Revision 9. Because there are thousands of ICD codes, companies seek to reduce the cardinality by adding codes of their own – for example, body part codes for the part of the body affected in the accident or illness that is the cause of the claim, and accident codes that categorize the cause of the accident (slip, fall, struck). The cardinality of the supplemental codes is typically less than 100, making these variables acceptable as inputs for predictive modeling.

Derived Variables

- Distance from claimant residence to health care provider
- Recency, frequency, and duration (RFD) statistics for health care provider payments
- Correlation of attorney contact date with chiropractor payments, claimant initiated provider change, onset of new symptoms, and so on
- Matched claim with relevant historical data:
 - Table of suspicious health care providers
 - Prior injuries
 - Provider primary practice and services provided

32

Insurance companies will excel in data mining after they recognize the value of adding derived variables to the existing insurance claims data. Transaction processing can provide many useful input variables in the form of RFD statistics.

Workers' Compensation Claims

Analytic Objective: Predict fraud for open claims.

Challenges:

- Rare target
- Claims adjusters using nonstandard technical language and abbreviations

Data:

- 3,037 documents extracted from Lotus Notes
- Documents merged with claims master data using claim ID (Most non-text fields are omitted.)

33

Workers' Compensation Claims

Name	Role	Level	Report	Order	Drop	Lower Limit
AdjusterNotes	Text	Nominal	No	No	-	
Body	Input	Nominal	No	No	-	
Cause	Input	Nominal	No	No	-	
ClaimNo	ID	Nominal	No	No	-	
FraudFlag	Rejected	Binary	No	No	-	
Nature	Input	Nominal	No	No	-	
SubroFlag	Rejected	Binary	No	No	-	
VEHflag	Input	Binary	No	No	-	

Metadata

34

The data set contains two target variables, **FraudFlag** and **SubroFlag**. The **FraudFlag** variable is questionable, and one goal of the analysis is to derive a proxy target variable using unsupervised learning techniques. The **SubroFlag** variable will be used in Chapter 5 in a project to create a model to predict subrogation. The role of both variables is Rejected.

Workers' Compensation Claims

An employee alleges that while lowering a machine down into the basement on planks and rollers, it fell off the planks and partially hit him on the shoulder and side.

The employee was terminated and the employer received knowledge of a claim via certified mail nine months later. The employee claims stress, stomach, back, and chest injuries.

Drink machine fell on employee.

The employee claims stress.

35

Workers' Compensation Claims

Data Preparation

- Remove proper names and other identifying information.
- Convert abbreviations.
- Correct misspellings.

36

Text Categorization for Predictive Modeling

Problem

Few if any cases are identified as fraudulent. Predictive modeling requires a reliable target variable.

Solution

Use clustering algorithms for outlier detection. Outliers are unusual cases and are potential indicators of fraud.

The unusual clusters serve as proxies for a target variable.

37

Outlier detection is an effective tool for identifying potentially fraudulent cases. As stated previously, fraud is committed by human beings, and human beings have a perceptual limitation of only a few dimensions. A person committing fraud wants to make the claim look like any other claim so it will not stand out and attract attention. However, what does not stand out when looking at one or a few attributes at a time can stand out when viewed in higher dimensional space. Because human beings can only view in three dimensions, analytics tools provide the resources to view data in higher dimensions. The analytic step of identifying outliers becomes a practical step of identifying unusual claims.



Text Categorization for Identifying Potential Fraud Cases

This demonstration illustrates how to use text categorization to derive a preliminary target variable that serves as a proxy for fraud.

- In this demonstration, the search for unusual cases uses only the **AdjusterNotes** field. In practice, you want to use as many text and numeric inputs as possible to derive clusters. The course data only has categorical inputs. The course data also lacks additional text fields such as physician reports, police reports, and accident reports.

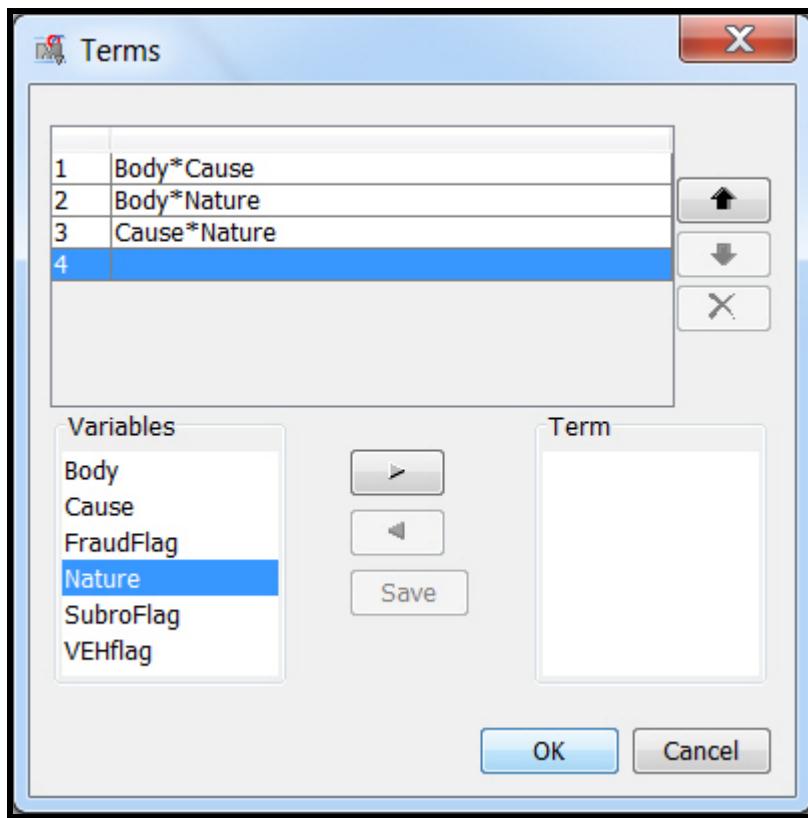
The SAS data set **LWDMTXT.WORKCOMP** contains 3,037 workers' compensation insurance claims. The data set has eight variables that are listed in a Metadata display above. The metadata table is reproduced below.

Variables - Ids						
(none)		<input type="checkbox"/> not	Equal to			
Columns:		<input type="checkbox"/> Label	<input type="checkbox"/> Mining	<input type="checkbox"/> Basic	<input type="checkbox"/> Statistics	
Name	Role	Level	Report	Order	Drop	Lower Limit
AdjusterNotes	Text	Nominal	No		No	.
Body	Input	Nominal	No		No	.
Cause	Input	Nominal	No		No	.
ClaimNo	ID	Nominal	No		No	.
FraudFlag	Rejected	Binary	No		No	.
Nature	Input	Nominal	No		No	.
SubroFlag	Rejected	Binary	No		No	.
VEHflag	Input	Binary	No		No	.

Note that the two target variables mentioned previously are rejected.

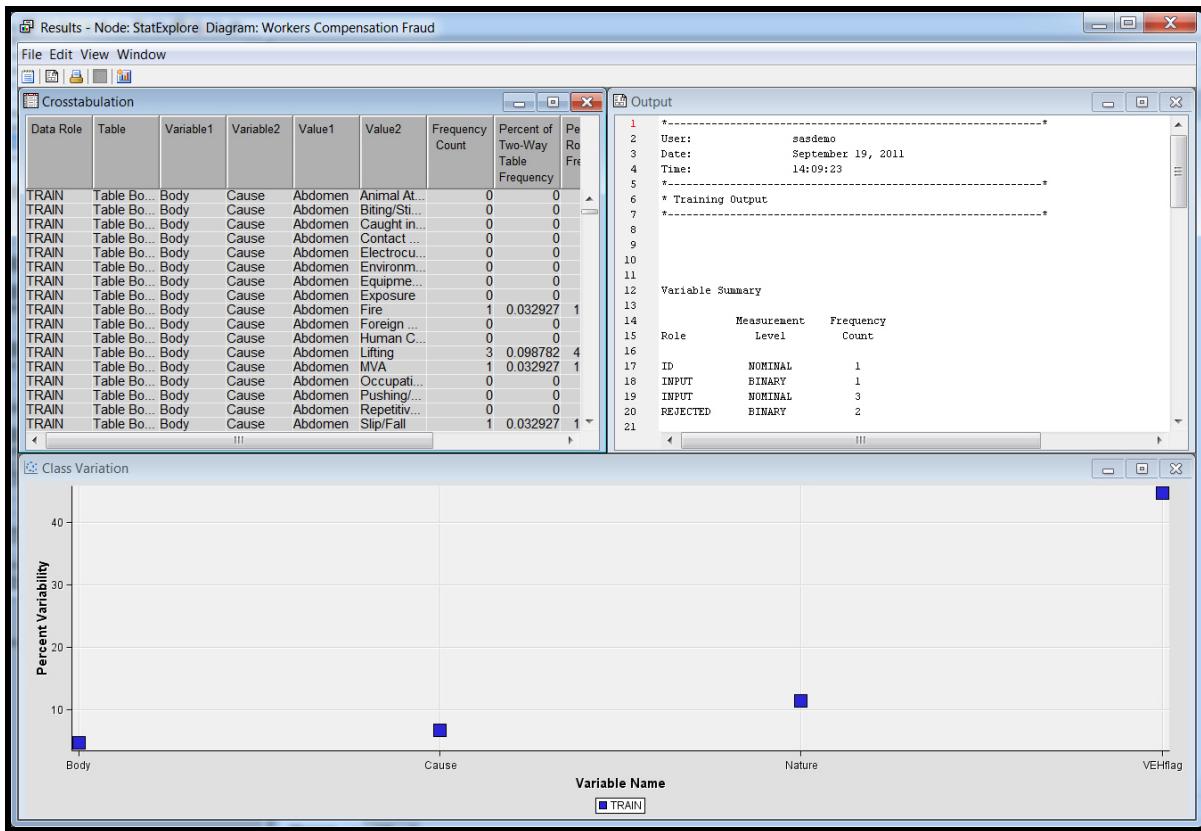
1. Create a new diagram and name it Workers Compensation Fraud.
2. Create an input data source for the **LWDMTXT.WORKCOMP** table if necessary. Be sure to change the role of the **FraudFlag** and **SubroFlag** variables to **Rejected**. This analysis assumes that the fraud target variable is unreliable. Drag the workers' compensation data source into the diagram.

3. Attach a StatExplore node to the Input Data Source node. Use the Cross-Tabulation property to get crosstabs for **Body**, **Cause**, and **Nature**.



For example, to get the **Body** by **Cause** crosstabulation, select **Body**, and then click on the right arrow. Select **Cause**, and click on the right arrow. Then select **Save**. Run the StatExplore node.

4. View the results.



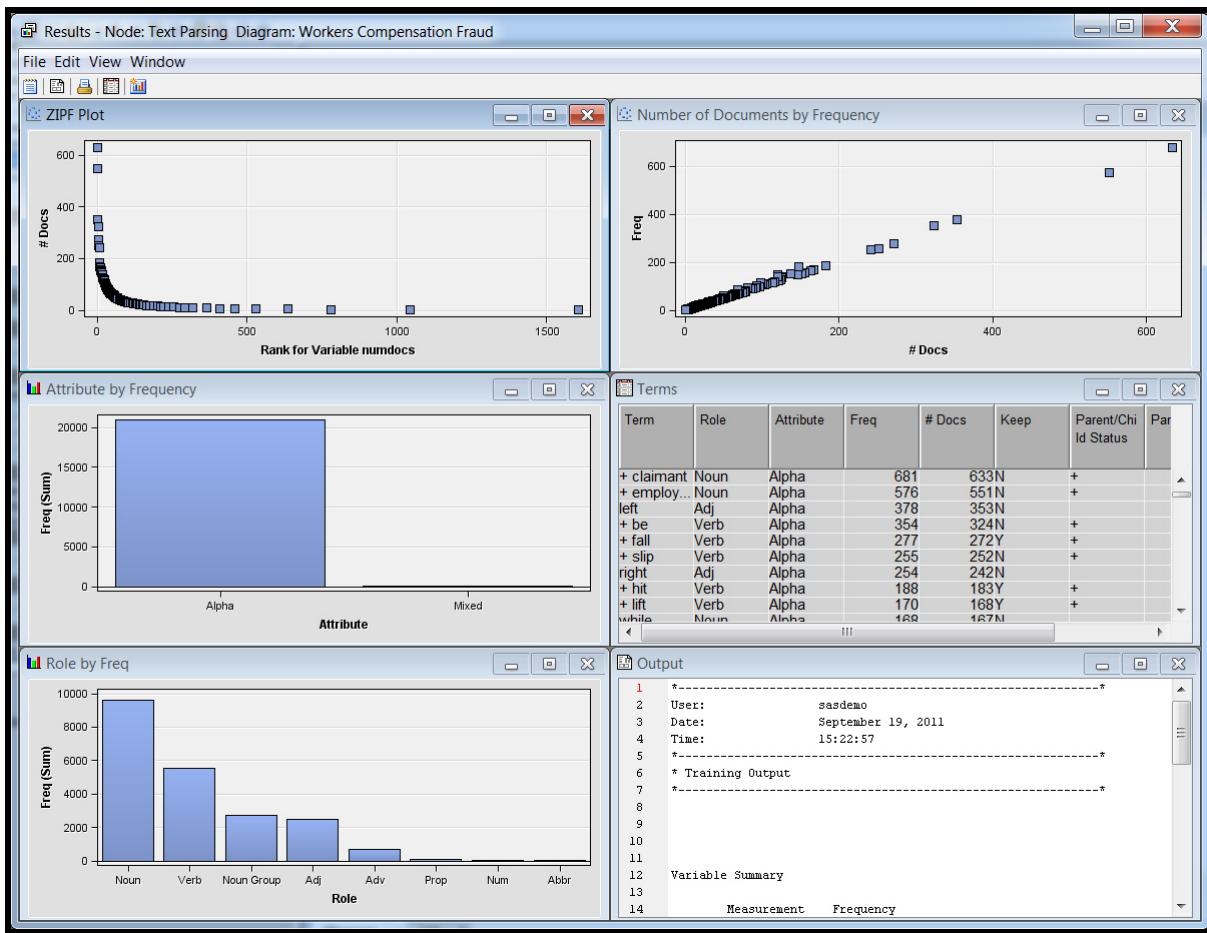
Here's a portion of the Crosstabulation table:

Data Role	Table	Variable1	Variable2	Value1	Value2	Frequency Count	Percent of Two-Way Table Frequency	Percent of Row Frequency	Percent of Column Frequency	Frequency Missing
TRAIN	Table Body * Cause	Body	Cause	Abdomen	Animal Attack	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Abdomen	Biting/Stinging Insect	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Abdomen	Caught in Machine	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Abdomen	Contact with Object	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Abdomen	Electrocution	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Abdomen	Environmental	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Abdomen	Equipment/Machinery	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Abdomen	Exposure	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Abdomen	Fire	1	0.032927	14.28571	2.222222	.
TRAIN	Table Body * Cause	Body	Cause	Abdomen	Foreign Body/Object	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Abdomen	Human Conflict	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Abdomen	Lifting	3	0.098782	42.85714	0.645161	.
TRAIN	Table Body * Cause	Body	Cause	Abdomen	MVA	1	0.032927	14.28571	0.266667	.
TRAIN	Table Body * Cause	Body	Cause	Abdomen	Occupational Illness	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Abdomen	Pushing/Pulling	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Abdomen	Repetitive Motion	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Abdomen	Slip/Fall	1	0.032927	14.28571	0.164204	.
TRAIN	Table Body * Cause	Body	Cause	Abdomen	Stress	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Abdomen	Struck Object	1	0.032927	14.28571	0.147493	.
TRAIN	Table Body * Cause	Body	Cause	Abdomen	Unknown	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Abdomen	Unusual Body Movement	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Ankle	Animal Attack	1	0.032927	0.862069	3.703704	.
TRAIN	Table Body * Cause	Body	Cause	Ankle	Biting/Stinging Insect	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Ankle	Caught in Machine	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Ankle	Contact with Object	1	0.032927	0.862069	2.380952	.
TRAIN	Table Body * Cause	Body	Cause	Ankle	Electrocution	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Ankle	Environmental	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Ankle	Equipment/Machinery	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Ankle	Exposure	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Ankle	Fire	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Ankle	Foreign Body/Object	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Ankle	Human Conflict	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Ankle	Lifting	14	0.460981	12.06897	3.010753	.
TRAIN	Table Body * Cause	Body	Cause	Ankle	MVA	1	0.032927	0.862069	0.266667	.
TRAIN	Table Body * Cause	Body	Cause	Ankle	Occupational Illness	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Ankle	Pushing/Pulling	6	0.197563	5.172414	6.25	.
TRAIN	Table Body * Cause	Body	Cause	Ankle	Repetitive Motion	1	0.032927	0.862069	0.763359	.
TRAIN	Table Body * Cause	Body	Cause	Ankle	Slip/Fall	47	1.54758	40.51724	7.71757	.
TRAIN	Table Body * Cause	Body	Cause	Ankle	Stress	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Ankle	Struck Object	11	0.3622	9.482759	1.622419	.
TRAIN	Table Body * Cause	Body	Cause	Ankle	Unknown	0	0	0	0	.
TRAIN	Table Body * Cause	Body	Cause	Ankle	Unusual Body Movement	34	1.119526	29.31034	18.18182	.
TRAIN	Table Body * Cause	Body	Cause	Arm	Animal Attack	3	0.098782	1.395349	11.11111	.
TRAIN	Table Body * Cause	Body	Cause	Arm	Biting/Stinging Insect	8	0.263418	3.72093	44.44444	.

Despite the lengthy table, you can find areas of high frequency, such as body part ankle and cause slip/fall. While categorical variables can be very useful in predictive modeling, unsupervised learning strategies do not typically work well with class input variables. The lack of data and the sparse results for many categories eliminates many of the strategies that might be employed to use class variables in clustering. Thus, you will have to rely on the textual data to facilitate grouping of claims.

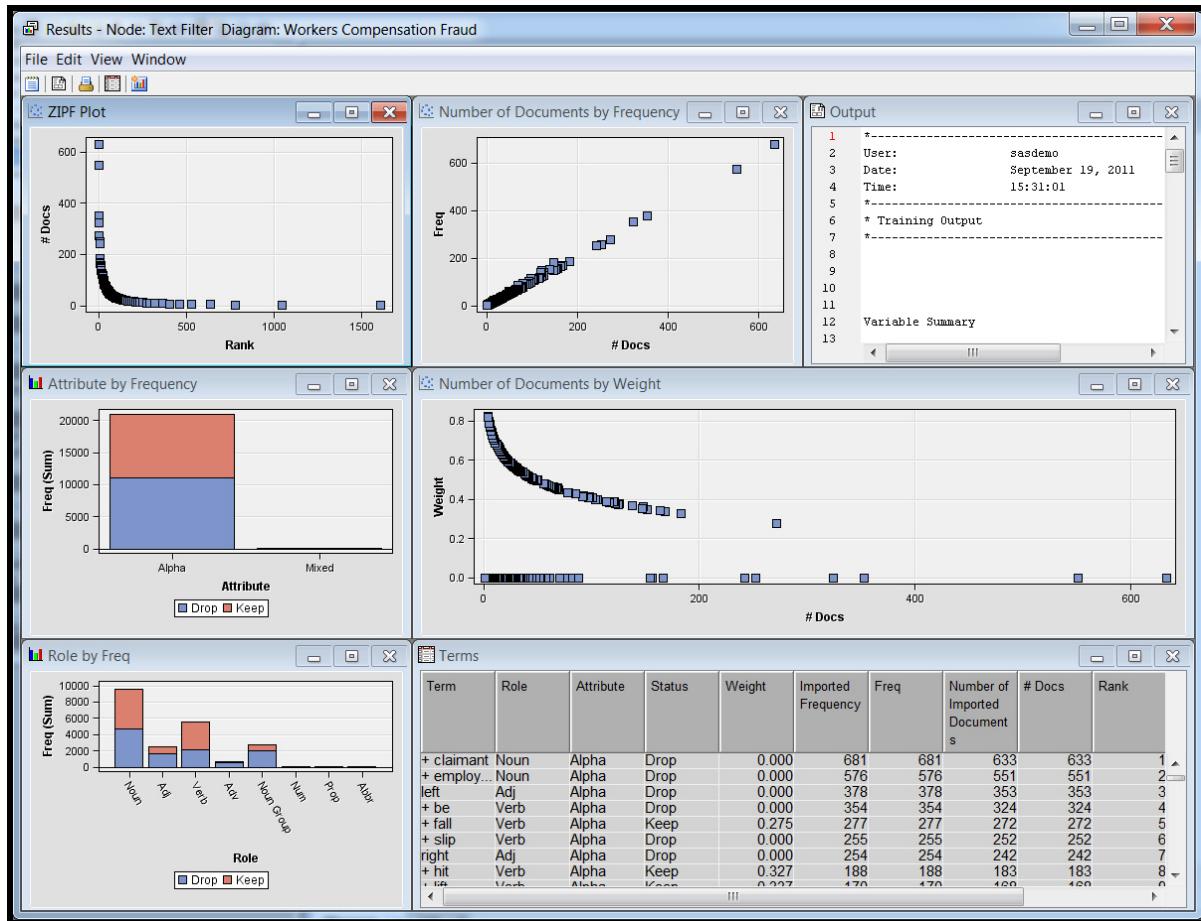
5. Attach a Data Partition node to the Input Data source node. Partitioning is required to ensure the stability of clusters that might be used operationally. Specify a train/validation/test split of 75/25/0. Run the Data Partition node.

6. Attach a Text Parsing node to the Data Partition node. Change Synonyms to **No Data Set to be Specified**. Change the stop list to **DMTXT.WORKCOMPSTOP**. Run the Text Parsing node. The Results window appears.

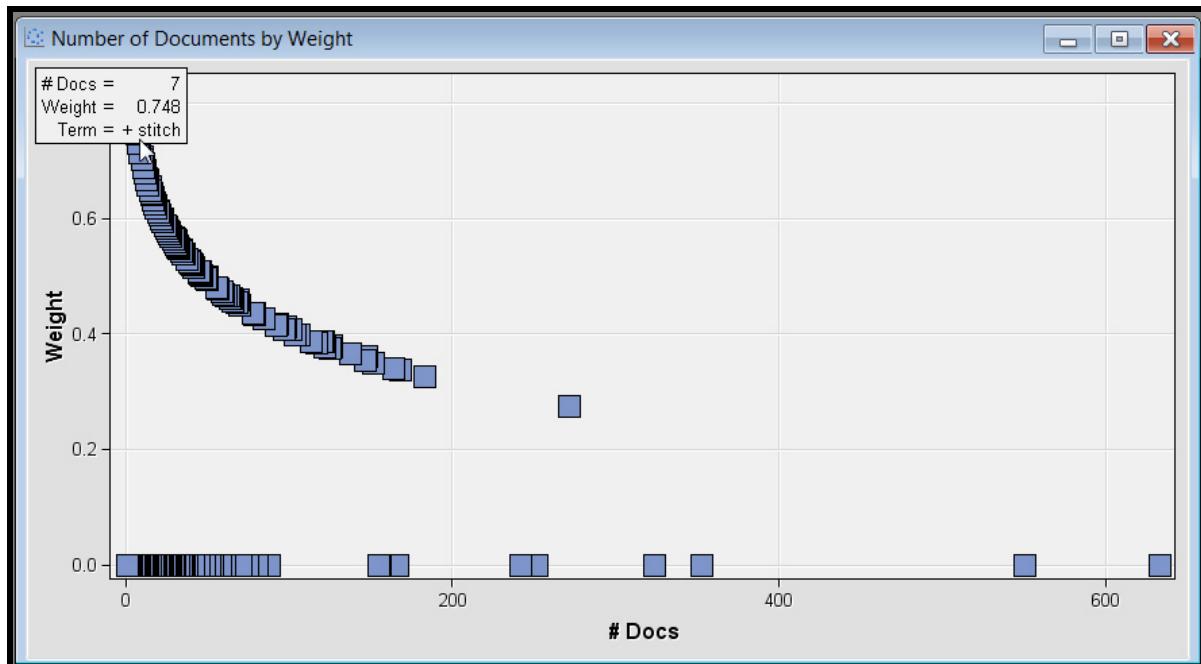


The results pass a “face validity” check. The ZIPF plot confirms to Zipf’s Law as described in the last chapter. Nouns dominate the parts of speech.

7. Attach a Text Filter node to the Text Parsing node. The default term weight is entropy when no target variable is present. For small documents like those in the insurance data, entropy is typically preferred over IDF weights. Run the Text Filter node. When the node completes successfully, access the Results window.



The terms are color coded in the bar charts: red for kept terms and blue for dropped terms. Examine the Number of Documents by Weight scatter plot.



If you position the cursor over a data point, you get statistics for that data point. For the term *+stitch*, (stitch and its synonyms and stemmed forms), you see that it appears in seven documents and has a term weight of 0.748, making it one of the more important terms in the collection. You can get more details by locating stitch in the Terms table.

Term	Role	Attribute	Status	Weight	Imported Frequency	Freq	Number of Imported Document s	# Docs	Rank
+ steer w...	Noun	Gro...	Alpha	Drop	0.000	1	1	1	1604
+ step	Verb	Alpha	Keep	0.478	59	59	57	57	57
+ step	Noun	Alpha	Keep	0.546	36	36	34	34	92
+ step door	Noun	Gro...	Alpha	Drop	0.000	1	1	1	1604
+ step walk	Noun	Gro...	Alpha	Drop	0.000	1	1	1	1604
+ stick	Verb	Alpha	Keep	0.658	14	14	14	14	222
+ sting	Verb	Alpha	Drop	0.000	3	3	3	3	781
+ stir hot ...	Noun	Gro...	Alpha	Drop	0.000	1	1	1	1604
+ stitch	Noun	Alpha	Keep	0.748	7	7	7	7	408
+ stitch	Verb	Alpha	Drop	0.000	1	1	1	1	1604
+ stitch ...	Noun	Gro...	Alpha	Drop	0.000	1	1	1	1604
+ stock	Verb	Alpha	Drop	0.000	3	3	3	3	781
+ stock c...	Noun	Gro...	Alpha	Drop	0.000	1	1	1	1604
+ stomach	Noun	Gro...	Alpha	Drop	0.000	1	1	1	1604
+ stone s...	Noun	Gro...	Alpha	Drop	0.000	1	1	1	1604
+ stop	Verb	Alpha	Drop	0.000	29	29	28	28	118
+ store	Verb	Alpha	Drop	0.000	2	2	2	2	1045
+ straight...	Verb	Alpha	Drop	0.000	5	5	5	5	528
+ strain	Verb	Alpha	Keep	0.470	60	60	60	60	53
+ straine...	Noun	Gro...	Alpha	Keep	0.748	7	7	7	408
+ strain	Noun	Alpha	Drop	0.000	1	1	1	1	625

The term *stitch* and its equivalent values appear seven times in seven documents, meaning that the term appears exactly one time in each of seven documents. The weight ranks 408 out of approximately 1,500 kept terms.

8. Close the Results window and open the Filter Viewer.

The screenshot shows the Interactive Filter Viewer interface with two windows:

- Main Window:** A table titled "ADJUSTOR NOTES" with columns: BODY P..., CAUSE..., CLAIM ..., FRAUD..., NATUR..., SUBRO..., VEHICL..., and _DATA... . The table contains numerous rows of accident descriptions and their corresponding codes.
- Terms Window:** A table titled "Terms" with columns: TERM, FREQ, # DOCS, KEEP ▾, WEIGHT, ROLE, and ATTRIBUTE. This table lists common words from the documents along with their frequency, document count, weight, part of speech, and attribute.

TERM	FREQ	# DOCS	KEEP ▾	WEIGHT	ROLE	ATTRIBUTE
fall	277	272	<input checked="" type="checkbox"/>	0.274	Verb	Alpha
hit	188	183	<input checked="" type="checkbox"/>	0.326	Verb	Alpha
lift	170	168	<input checked="" type="checkbox"/>	0.336	Verb	Alpha
back	167	164	<input checked="" type="checkbox"/>	0.34	Noun	Alpha
pain	152	152	<input checked="" type="checkbox"/>	0.349	Noun	Alpha
vehicle	183	148	<input checked="" type="checkbox"/>	0.361	Noun	Alpha
strain	148	147	<input checked="" type="checkbox"/>	0.353	Noun	Alpha
hand	153	138	<input checked="" type="checkbox"/>	0.365	Noun	Alpha
machine	135	126	<input checked="" type="checkbox"/>	0.377	Noun	Alpha
knee	139	126	<input checked="" type="checkbox"/>	0.378	Noun	Alpha
injure	127	125	<input checked="" type="checkbox"/>	0.375	Verb	Alpha
allege	123	122	<input checked="" type="checkbox"/>	0.377	Verb	Alpha
door	148	121	<input checked="" type="checkbox"/>	0.386	Noun	Alpha
finger	141	120	<input checked="" type="checkbox"/>	0.385	Noun	Alpha
accident	119	117	<input checked="" type="checkbox"/>	0.383	Noun	Alpha
cause	115	114	<input checked="" type="checkbox"/>	0.386	Verb	Alpha
pull	109	106	<input checked="" type="checkbox"/>	0.397	Verb	Alpha

Before embarking on the fraud investigation, you might want to use features of the Text Filter node to help assess the quality of the data. A quick example will show how this might be accomplished. In the Terms table, find the word *struck*. Filter on this word. Examine the fields **Body**, **Cause**, and **Nature**. Are the codes consistent with the use of the term *struck*?

The screenshot shows the 'Interactive Filter Viewer' application window. At the top, there's a menu bar with File, Edit, View, Window. Below the menu is a toolbar with icons for search, apply, and clear. A search field contains the word 'struck'. To the right of the search field are 'Apply' and 'Clear' buttons. The main area is divided into two panes. The left pane, titled 'Documents', lists adjuster notes. The right pane, titled 'Terms', lists terms with their frequency, document count, weight, role, and attribute.

TERM	FREQ	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
100lb	0	0	<input type="checkbox"/>	0.0	Noun	Mixed
100lbs	0	0	<input type="checkbox"/>	0.0	Noun	Mixed
10mo	0	0	<input type="checkbox"/>	0.0	Noun	Mixed
2x4	1	1	<input type="checkbox"/>	0.0	Noun	Mixed
2x4 block	0	0	<input type="checkbox"/>	0.0	Noun Group	Mixed
2x4 piece	0	0	<input type="checkbox"/>	0.0	Noun Group	Mixed
30lb	0	0	<input type="checkbox"/>	0.0	Noun	Mixed
4ft	0	0	<input type="checkbox"/>	0.0	Noun	Mixed
4th	0	0	<input type="checkbox"/>	0.0	Noun	Mixed
50lb	0	0	<input type="checkbox"/>	0.0	Noun	Mixed
50lbs	0	0	<input type="checkbox"/>	0.0	Noun	Mixed

The codes seem to be consistent with the description of the accident.

You would like SAS Text Miner to tell you what terms correspond to unusual cases, but you also have domain knowledge that suggests some terms might be related to fraud. Adjuster notes can be influenced by past investigations, and this particular data set includes data from the period where a lawyer/chiropractor/recruiter fraud ring was somewhat prevalent. This fraud ring was described in a previous chapter. If the medical provider field were included, you would see many chiropractor providers, because chiropractors are routinely employed for back injury cases. However, adjuster notes rarely mention the actual provider unless fraud is suspected.

- To investigate the lawyer/chiropractor/recruiter type fraud ring, you can search on lawyer or chiropractor. Begin the search by looking for notes that contain the word *chiropractor*. Sort the Terms table by the TERM column, and then locate the *chiropractor* term entry. The term *chiropractor* is not stemmed, and it appears exactly one time in each of 24 documents. Type **chiropractor** in the Search field, and click **Apply**.

The screenshot shows the 'Interactive Filter Viewer' application interface. At the top, there's a menu bar with File, Edit, View, Window. Below it is a toolbar with icons for search, apply, and clear. A search bar contains the text 'chiropractor'. There are 'Apply' and 'Clear' buttons. The main window has a title 'Documents' and displays a table of search results.

ADJUSTOR NOTES	TEXTFILTER_SNIPPET	TEXTFILTER_RELEVANCE	BODY P...	CAUSE...	CLAIM ...	FRAUD...	NAT...
Claimant opening his tool box, bending over, felt a pinch in his back.	... Chiropractor billing error .	1.0	Back	Unusual ...	217121...	0.0	Spr
The ladder began to slip, fell on back. Chiropractor on watch list.	... Chiropractor on watch list .	1.0	Back	Slip/Fall	2227063...	0.0	Spr
While pushing a keg, employee strained his back. Chiropractor more than 30	... Chiropractor more than 30	1.0	Back	Pushing/...	2233217...	0.0	Spr
Back pain while working in loading bay. Chiropractor prior record.	... Chiropractor prior record .	1.0	Back	Lifting	4420617...	1.0	Spr
Lawyer filed claim, worker no longer employed. Lawyer and chiropractor on	... Lawyer and chiropractor on	1.0	Back	Lifting	4422627...	1.0	Spr
Chiropractor prior infraction. Alleges lifting heavy plastic bins.	... Chiropractor prior	1.0	Back	Lifting	4422917...	1.0	Spr
Neck pain after alleged struck by grinder arm. Soft tissue, exam showed no	... Chiropractor double billed	1.0	Neck	Struck O...	4423257...	1.0	Cor
Chiropractor under investigation for fraud. Physician report requested.	... Chiropractor under	1.0	Back	Struck O...	4423617...	1.0	Cor
Lawyer filed claim, alleges trauma from fall on wet floor. Chiropractor on	... Chiropractor on watch list .	1.0	Head	Slip/Fall	4423697...	1.0	Cor
Chiropractor bill dated before injury date. Claim filed more than 30 days late.	... Chiropractor bill dated	1.0	Back	Lifting	4424850...	1.0	Spr
Installing closet doors. Chiropractor billing error.	... Chiropractor billing error .	1.0	Back	Pushing/...	4424917...	1.0	Spr
Chiropractor ICD-9 code error. Alleged back injury.	... Chiropractor ICD-9 code	1.0	Back	Lifting	4425017...	1.0	Spr
Transporting medical records to main office. Chiropractor possible fraud.	... Chiropractor possible fraud	1.0	Back	Lifting	4425017...	1.0	Spr
Rebar extended beyond faceplate. Chiropractor 30 miles.	... Chiropractor 30 miles . . .	1.0	Back	Struck O...	4425257...	1.0	Cor
Motor vehicle accident, no property damage. Chiropractor treatment for back	... Chiropractor treatment for	1.0	Neck	MVA	4425257...	1.0	Cor
Chiropractor suspicious, referred to SIU.	... Chiropractor suspicious ,	1.0	Back	Lifting	4425917...	1.0	Spr
Back, soft tissue, chiropractor more than 50 miles.	... soft tissue , chiropractor	1.0	Back	Pushing/...	4427617...	1.0	Spr
Back strain from lifting. Chiropractor/lawyer named	Chiropractor / lawyer	1.0	Back	Lifting	4427617...	1.0	Spr

Below the table, there's a section titled 'Terms' with a table:

TERM ▲	FREQ	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
chipper	0	0	<input type="checkbox"/>	0.0	Adj	Alpha
chipper frame	0	0	<input type="checkbox"/>	0.0	Noun Group	Alpha
chiropractor	24	24	<input type="checkbox"/>	0.0	Noun	Alpha
chisel	0	0	<input type="checkbox"/>	0.0	Noun	Alpha
chlorine	0	0	<input type="checkbox"/>	0.0	Noun	Alpha
chlorine water	0	0	<input type="checkbox"/>	0.0	Noun Group	Alpha
chop	0	0	<input type="checkbox"/>	0.0	Verb	Alpha
cide	0	0	<input type="checkbox"/>	0.0	Noun	Alpha
city	0	0	<input type="checkbox"/>	0.0	Adj	Alpha
city medical ce...	0	0	<input type="checkbox"/>	0.0	Noun Group	Alpha
city truck	0	0	<input type="checkbox"/>	0.0	Noun Group	Alpha
city vehicle	0	0	<input type="checkbox"/>	0.0	Noun Group	Alpha
claim	3	3	<input type="checkbox"/>	0.0	Noun	Alpha
claim	0	0	<input type="checkbox"/>	0.0	Verb	Alpha
claimant	5	5	<input type="checkbox"/>	0.0	Noun	Alpha
claimant assist...	0	0	<input type="checkbox"/>	0.0	Noun Group	Alpha
claimant back ...	0	0	<input type="checkbox"/>	0.0	Noun Group	Alpha

In lower dimensional space, the use of a chiropractor is consistent with each of the variables **Body**, **Cause**, and **Nature**. However, examining the **TEXTFILETR_SNIPPET** field reveals that all chiropractor references are suspicious in some way. The claims adjusters were following their training and noting suspicious behavior to help fraud investigations. These 24 cases should probably be investigated.

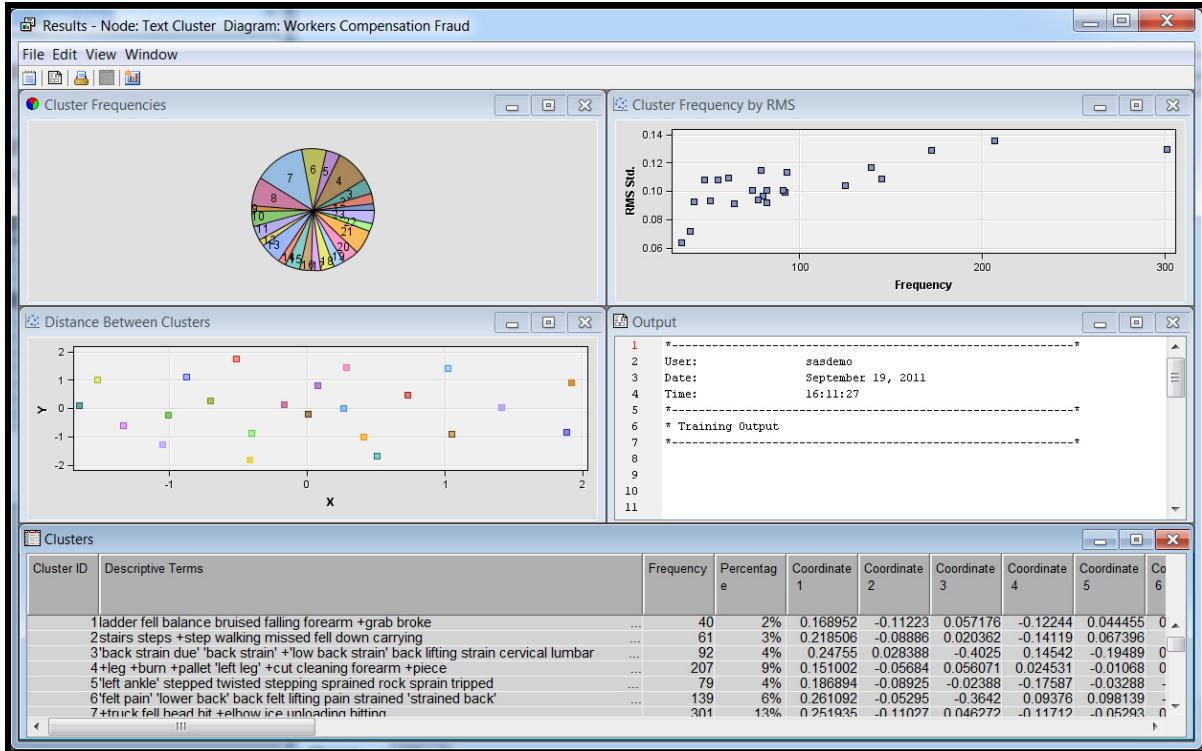
10. Examine which claims mention both *chiropractor* and *lawyer* by entering the search expression **lawyer & chiropractor**.

Without the **&**, the search would be performed as a logical OR operation.

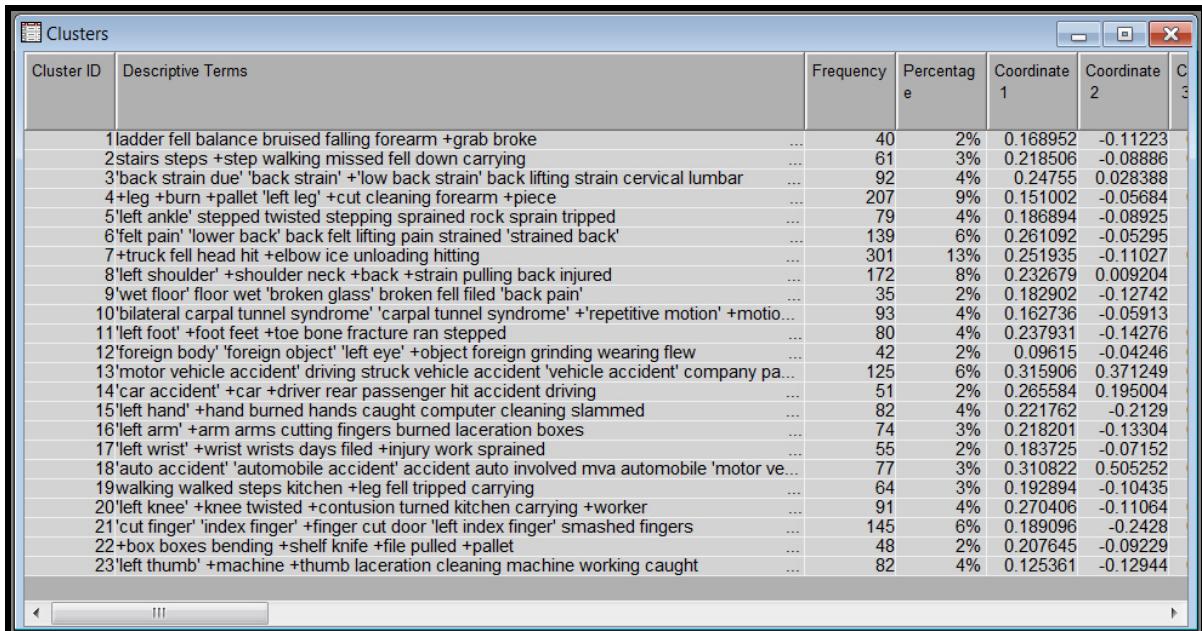
The screenshot shows the Interactive Filter Viewer interface. At the top, there is a search bar with the query "lawyer & chiropractor" and buttons for "Apply" and "Clear". Below the search bar is a table titled "Documents" with columns: ADJUSTOR NOTES, TEXTFILTER_SNIPPET, TEXTFILTER_RELEVANCE, BODY P..., CAUSE..., CLAIM..., FRAUD..., and VEHICL... . The table contains several rows of document snippets and their relevance scores. Below this is a section titled "Terms" with a table showing term frequency, document count, keep status, weight, role, and attribute for various terms like chipper, chiropractor, chisel, chlorine, etc.

TERM ▲	FREQ	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
chipper	0	0	<input type="checkbox"/>	0.0	Adj	Alpha
chipper frame	0	0	<input type="checkbox"/>	0.0	Noun Group	Alpha
chiropractor	3	3	<input type="checkbox"/>	0.0	Noun	Alpha
chisel	0	0	<input type="checkbox"/>	0.0	Noun	Alpha
chlorine	0	0	<input type="checkbox"/>	0.0	Noun	Alpha
chlorine water	0	0	<input type="checkbox"/>	0.0	Noun Group	Alpha
chop	0	0	<input type="checkbox"/>	0.0	Verb	Alpha
cidex	0	0	<input type="checkbox"/>	0.0	Noun	Alpha
city	0	0	<input type="checkbox"/>	0.0	Adj	Alpha
city medical ce...	0	0	<input type="checkbox"/>	0.0	Noun Group	Alpha
city truck	0	0	<input type="checkbox"/>	0.0	Noun Group	Alpha
city vehicle	0	0	<input type="checkbox"/>	0.0	Noun Group	Alpha
claim	2	2	<input checked="" type="checkbox"/>	0.0	Noun	Alpha
claims	n	n	<input type="checkbox"/>	n n	Verb	Alpha

11. Despite the use of logical AND, one claim does not have the word *chiropractor*. This is the nature of the Boolean query. The similarity of “Chiropractor/lawyer paired” and “Physician/lawyer paired” makes the documents appear very similar with respect to the query. Searching on *lawyer* alone would have uncovered the claim, but this example illustrates how filtering can aid in the fraud investigation. Close the Filter Viewer, and do not save any filter results because you want to export the complete document collection.
12. Attach a Text Cluster node to the Text Filter node. Deriving mutually exclusive clusters is often superior to deriving topics because fraud does not have partial membership in a group of documents. Furthermore, clustering algorithms provide powerful tools for outlier detection.
13. It might seem reasonable to request two clusters hoping to get good separation between fraud and non-fraud cases. This action is never recommended. Clustering as a tool for finding outliers works best by letting the algorithm break the data into well-separated groups. Unusual claims could be far apart from each other as well as far apart from normal clusters, so dividing the data into two pieces just forces outliers to team up with one of two clusters. It is common to have two or more unusual clusters and to have several clusters be chosen as proxies for fraud cases. Use the default setting for the Text Cluster node and run the node.



The cubic clustering criterion selected 23 clusters. The descriptive terms will help understand the clusters, but the initial focus is on cluster document frequency. Outliers tend to be in small clusters.

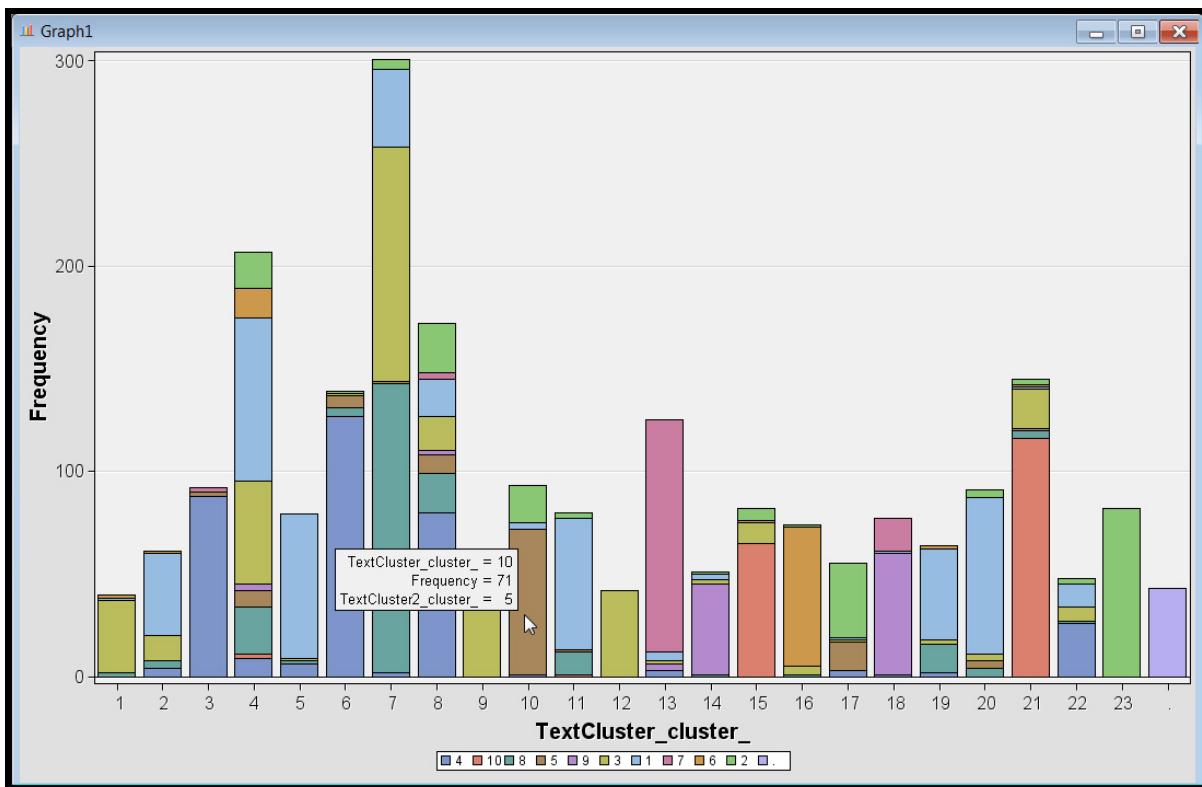


Unfortunately, the data set is so small that most clusters are small. If a cluster truly represents outliers, then combining clusters will not affect outlying clusters.

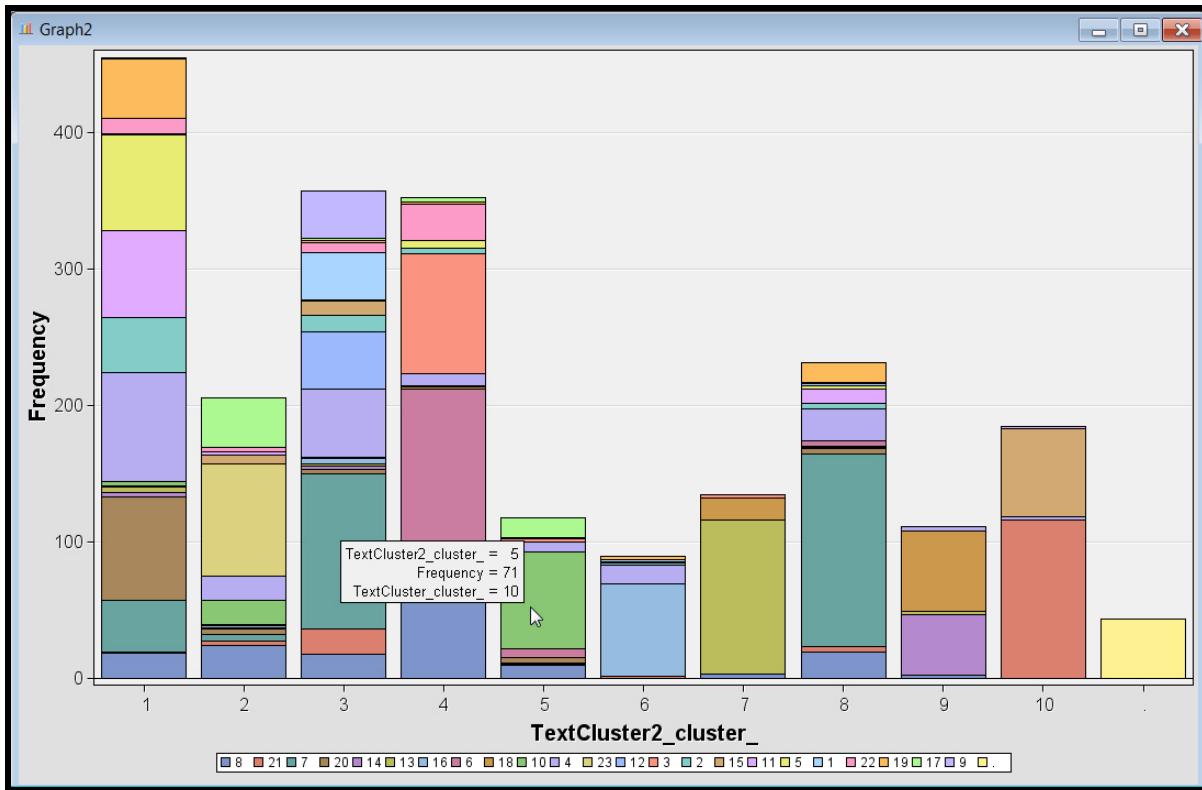
14. Attach a second Text Cluster node to the previous Text Cluster node. Rename the node 10 Clusters. Select **Text Cluster** properties to request exactly 10 clusters. Re-run the node.

Cluster ID	Descriptive Terms	Frequency	Percentag e	Coordinate 1	Coo 2
1	'left ankle' 'left foot' 'left knee' +foot +knee stepped tripped twisted	456	20%	0.200833	-0
2	'left thumb' 'left wrist' +machine +thumb +wrist repetitive +motion +'repetitive motion'	205	9%	0.166536	-0
3	'wet floor' +object fell floor foreign ladder wet door	357	16%	0.207832	-0
4	'back strain' 'lower back' back lifting pain strain strained groin	352	16%	0.240945	-0
5	'bilateral carpal tunnel syndrome' 'carpal tunnel syndrome' alleges syndrome filed computer alleging bilateral	117	5%	0.192902	-0
6	'left arm' +arm exposure hot burned arms cutting boxes	89	4%	0.194216	-0
7	'motor vehicle accident' 'vehicle accident' accident driving struck vehicle involved company	134	6%	0.330167	0
8	'left leg' +leg +truck hit hitting head unloading injured	230	10%	0.24759	-0
9	'auto accident' 'car accident' +car accident auto involved automobile neck	111	5%	0.281495	0.3
10	'index finger' 'left hand' 'left index finger' +finger +hand caught cut slammed	184	8%	0.189299	-0

15. Compare cluster 10 in the previous run to cluster 5 above. Close examination reveals that most of the 93 cases in cluster 10 also appear in cluster 5. You can investigate this by using the exploration features of SAS Enterprise Miner.
16. Close the Results window. In the Properties panel of the second Text Cluster node, select **Exported Data**, click on the **TRAIN** data, and select **Explore**. Use the plot wizard to construct a bar chart with **TextCluster_cluster_** as the category variable and **TexCluster2_cluster_** as the group variable.



17. Position the cursor over the cluster 10 bar. Observe that 71 of the 93 cases are also in cluster 5 from the second Text Cluster node. You can reverse the roles of the two segment variables to get another perspective.



The 71 cases in the overlap might be the best candidates for fraud related to soft tissue injuries like carpal tunnel where physical evidence of an actual injury might not be compelling. Fraud is common for soft tissue injuries because it can be impossible to verify the existence of an injury or pain using medical diagnostic equipment.

18. You can use SAS Enterprise Miner nodes to extract the 71 cases identified in the previous step. The steps to do so will be sketched here, but this step will be skipped to conserve class time. You can attach a Filter node (Sample tab) to the second Text Cluster node, and filter out all clusters except cluster 10 using **TextCluster_cluster_** and cluster 5 using **TextCluster2_cluster_**. You can attach a Drop node (Modify tab) to the Filter node and drop all of the variables derived from text mining. You can then attach a Text Parsing node followed by a Text Filter node to the Drop node. You would use the same settings for these nodes as you used before. Running these nodes then allows you to use the Text Filter node to examine the 71 documents.

Interactive Filter Viewer

File Edit View Window

Search : Apply Clear

Documents

ADJUSTOR NOTES		BODY P...	CAUSE...	CLAIM ...	FRAUD...
Employee alleges from heavy typing, filing and phones a repetitive motion causing both wrists to be sore. Alleges tendonitis and carpal tunnel syndrome.		Wrist	Repetitiv...	0074294...	
Employee alleges bilateral wrist pain due to typing.		Wrist	Repetitiv...	0904168...	
Claimant alleges stress in the work place. Denied pending appeal.		Multiple	Stress	1433125...	
Employee alleges right shoulder pain.		Shoulder	Lifting	1495052...	
Alleging bilateral carpal tunnel syndrome, due to handling labels, hands started to cramp and become numb.		Hand	Repetitiv...	1884258...	
Due to claimant using a computer typewriter, she has numbness/tingling in both wrist, swelling, pain, coldness and hard time to bend fingers, possible carpal tunnel syndrome.		Finger	Repetitiv...	1977125...	
Due to constant keying into the computer, claimant is having problems with numbness and tingling.		Hand	Repetitiv...	2016222...	
Employee alleges closing door she twisted ankle.		Ankle	Struck O...	2020293...	
Employee alleges she reached over to printer and felt a muscle pull.		Back	Pushing/...	2076018...	
Claimant alleges aggravation of carpal tunnel syndrome due to excessive writing at work. She reported after taking disability absence for 30+ days.		Wrist	Repetitiv...	2502163...	
Employee alleges he strained right knee while moving jack.		Knee	Lifting	2542016...	
Repetitive duties causing carpal tunnel syndrome on right side with shoulder strain.		Shoulder	Repetitiv...	2911280...	
Employee alleges disk aggravated by bending and lifting.		Spine	Lifting	2912025...	
Alleges bilateral wrist pain from repetitive work.		Wrist	Repetitiv...	2965239...	
Claimant alleges right carpal tunnel syndrome from data entry.		Wrist	Repetitiv...	3002246...	
Employee alleges that from continuous lifting of objects he developed carpal tunnel syndrome of his arms and wrists.		Arm	Lifting	3013175...	
Claimant alleges carpal tunnel syndrome and back injury from working in the cashiers cage from lifting and moving 25 lbs containers of coin.		Back	Lifting	3280278...	

Terms

TERM	FREQ	# DOCS	KEEP ▾	WEIGHT	ROLE	ATTRIBUTE
allege	47	46	<input checked="" type="checkbox"/>	0.104	Verb	Alpha
syndrome	25	25	<input checked="" type="checkbox"/>	0.245	Noun	Alpha
carpal tunnel s...	14	14	<input checked="" type="checkbox"/>	0.381	Noun Group	Alpha
computer	11	11	<input checked="" type="checkbox"/>	0.437	Noun	Alpha
repetitive	10	10	<input checked="" type="checkbox"/>	0.46	Adj	Alpha
work	10	10	<input checked="" type="checkbox"/>	0.46	Noun	Alpha
pain	9	9	<input checked="" type="checkbox"/>	0.485	Noun	Alpha
cause	8	8	<input checked="" type="checkbox"/>	0.512	Verb	Alpha
lift	8	8	<input checked="" type="checkbox"/>	0.512	Verb	Alpha
bilateral	7	7	<input checked="" type="checkbox"/>	0.544	Adj	Alpha
bilateral carpal...	6	6	<input checked="" type="checkbox"/>	0.58	Noun Group	Alpha

Two of the chiropractor cases fall in this overlapped cluster.

Interactive Filter Viewer

File Edit View Window

Search : chiropractor Apply Clear

Documents

ADJUSTOR NOTES		TEXTFILTER2_SNIPPET	TEXTFILTER2_RELEVANCE
Chiropractor prior infraction. Alleges lifting heavy plastic bins.		... Chiropractor prior	1.0
Claimant alleges foreman required too many units in plastic bins. Chiropractor 40 miles.		... Chiropractor 40 miles . . .	1.0

Terms

TERM ▲	FREQ	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
catch	0	0	<input type="checkbox"/>	0.0	Verb	Alpha
cause	0	0	<input type="checkbox"/>	0.0	Verb	Alpha
chiropractor	2	2	<input type="checkbox"/>	0.0	Noun	Alpha
claimant	1	1	<input type="checkbox"/>	0.0	Noun	Alpha
closing door	0	0	<input type="checkbox"/>	0.0	Noun	Alpha
coin	0	0	<input type="checkbox"/>	0.0	Noun	Alpha
coldness	0	0	<input type="checkbox"/>	0.0	Noun	Alpha
compromise	0	0	<input type="checkbox"/>	0.0	Verb	Alpha

19. The Text Topic node can also help with fraud investigation, even though topics are different than clusters and are not constrained to be mutually exclusive. Attach a Text Topic node to the second Text Cluster node. Use the default settings. You can benefit from using domain knowledge to specify custom fraud topics, but this task is not included in the demonstration. Run the Text Topic node. When the node has completed execution, access the Topic Viewer.

The screenshot shows the "Interactive Topic Viewer" application window with three main tabs:

- Topics:** Displays a table of topics with columns: Topic, Category, Term Cutoff, and Doc. The topics listed include "vehicle, +strike, motor vehicle accident, accident, passenger", "accident, auto, +involve, auto accident, motor vehicle accident", "strain, back, +lift, back strain, +box", "+finger, +laceration, cut, index finger, +cut", "+machine, +clean, +catch, +work, machine", and "+door, +catch, +hand, +close, +open".
- Terms:** Displays a table of terms with columns: Topic Weight, +, Term, Role, # Docs, and Freq. The terms listed include "vehicle" (Noun, 148 docs, 183 freq), "strike" (Verb, 104 docs, 106 freq), "motor vehicle accident" (Noun Group, 29 docs, 29 freq), "accident" (Noun, 117 docs, 119 freq), "passenger" (Noun, 24 docs, 24 freq), and "neck" (Noun, 70 docs, 70 freq).
- Documents:** Displays a table of documents with columns: Topic Weight, Adjustor Notes, Body Part, Cause of Injury, Claim Number, and Fraud Flag (1=Yes). The documents listed include "Employee was Multiple MVA 351911195108 0.0", "Involved in a motor Back MVA 138915357608 0.0", "Deliver driver had Back MVA 916414913508 0.0", "Employee says that Back MVA 940710496308 0.0", "Vehicle accident. Neck MVA 723919881108 0.0", "Claimant driving Multiple MVA 368719488108 0.0", and "Claimant passenger Multiple MVA 375400400008 0.0".

20. Scroll down to find some soft tissue injury topics.

The screenshot shows the 'Interactive Topic Viewer' application window with three main sections: 'Topics', 'Terms', and 'Documents'.

Topics:

Topic	Category	Term Cutoff	Doc
+door, +catch, +hand, +close, +open	Mult	0.282	0.587
back,strained, +pull, +lift, +low back	Mult	0.285	0.609
pain,back, +feel, +low back,back	Mult	0.29	0.607
+foot,left foot, +cause, +fall, +step	Mult	0.277	0.562
ladder, +fall, +bruise,head, +walk	Mult	0.269	0.579
+knee,left knee, +contusion, +pull, +twist	Mult	0.273	0.55

Terms:

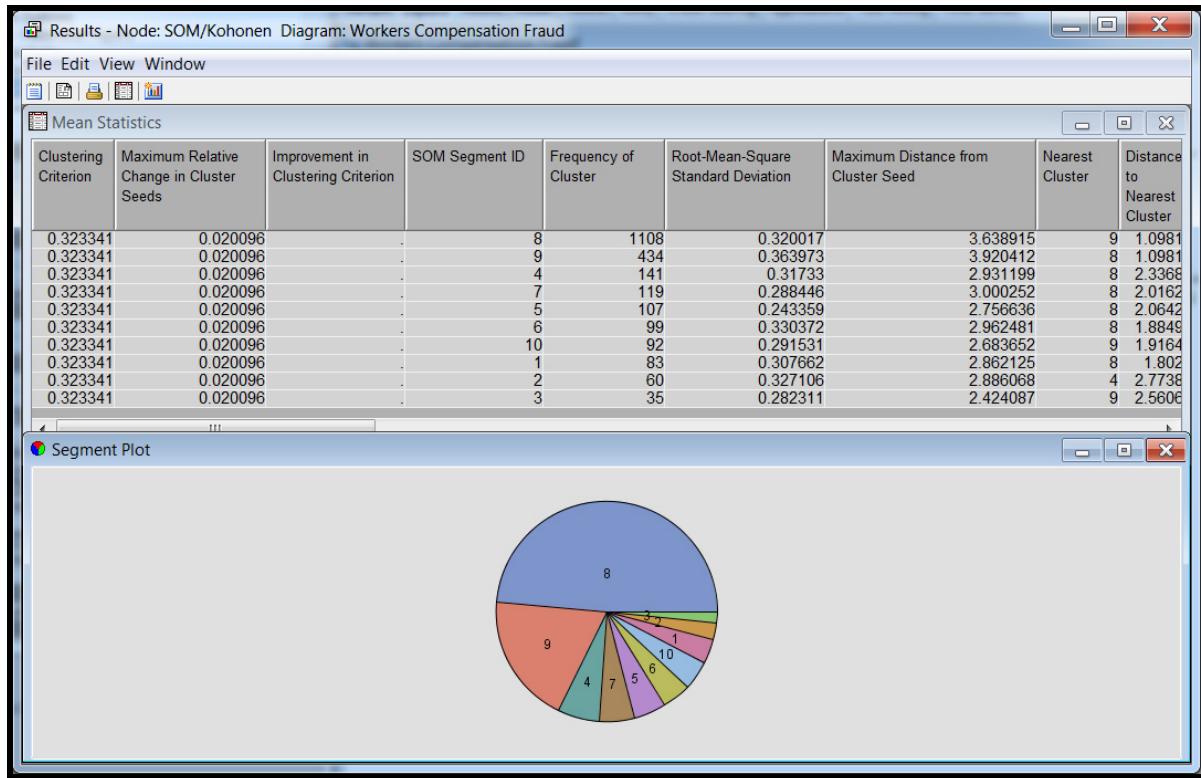
Topic Weight	+	Term	Role	# Docs	Freq
2.879		back	Noun	164	167
1.664		strained	Adj	45	45
1.519	+	pull	Verb	106	109
1.425	+	lift	Verb	168	170
1.2	+	low back	Noun Group	47	47
0.777		strained back	Noun Group	11	11

Documents:

Topic Weight	Adjustor Notes	Body Part	Cause of Injury	Claim Number	Fraud Flag (1=Yes)
2.883	Employee was lifting	Back	Lifting	030809186508	0.0
2.816	Pulling cart, strained	Back	Pushing/Pulling	665007553208	0.0
2.728	Lifted baydoor not	Back	Lifting	349520757408	0.0
2.728	Employee was lifting	Back	Lifting	451510558408	0.0
2.64	Lifting a keg, and	Back	Lifting	247620729908	0.0
2.64	Claimant lifted keg	Back	Lifting	318420911808	0.0
2.564	Claimant states he	Back	Lifting	288100907908	0.0

Farther down the Topics table is a carpal tunnel syndrome topic. Documents exhibiting one or more of these topics might be worth investigating. A useful benefit to the Text Topic node is that it exports the rotated SVD values used to derive the topics. These SVD columns can be used as inputs to an unsupervised learning node. While the demonstration can end here, the next few steps will sketch how to use SAS Enterprise Miner with SAS Text Miner inputs to search for outliers.

21. Attach an SOM/Kohonen node to the Text Topic node. Select all of the **TextTopic_raw** variables as input variables. Set everything else to have a Use status of **No**. Select the **Kohonen VQ** method (Vector Quantization), and leave the default of a maximum of **10** clusters. Set the Internal Standardization property to **None** because the inputs are already standardized. Run the SOM/Kohonen node. View the results.



Cluster 3 has 35 claims, so it is a candidate for an outlier cluster. You can use the same methods outlined above to extract these 35 cases and feed them back to the Text Filter node so they can be examined.

Here are some of the documents in cluster 3.

The screenshot shows the 'Interactive Filter Viewer' application interface. At the top is a menu bar with File, Edit, View, Window. Below it is a toolbar with a magnifying glass icon, a search input field containing 'Search :', and buttons for Apply and Clear. The main area has two windows:

- Documents**: A table titled 'ADJUSTER NOTES' with columns: BODY P..., CAUSE..., CLAIM ..., NATUR..., VEHICL... and rows of accident reports. Some examples include: 'Employee slipped on a wet floor and fell landing on her left shoulder.', 'Alleges left shoulder pain from slip and fall on wet floor.', 'Employee slipped and fell on wet bathroom floor from overflowing toilet.', etc.
- Terms**: A table with columns: TERM, FREQ, # DOCS ▾, KEEP, WEIGHT, ROLE, ATTRIBUTE. It lists common terms like 'floor', 'slip', 'wet', 'wet floor', 'fall', 'employee', 'walk', 'claimant', 'on', 'shoulder', 'right', 'left', 'contusion', 'hand', 'hit', and 'knee'. The 'KEEP' column contains checkboxes, many of which are checked. The 'ATTRIBUTE' column is labeled 'Alpha' for all terms.

Cluster 3 might not be an outlier cluster. The claims look like ordinary slip-and-fall accidents. You should proceed to the next sparse cluster, namely cluster 2 with 60 cases. You would continue in this fashion trying to identify any outlier clusters and see if the cases are candidates for fraud.

Text categorization identifies unusual injuries, mostly soft-tissue injuries that are prevalent in fraud cases. The expectation-maximization clustering algorithm can be used to score new cases if the unusual clusters are pure. Because domain expertise almost always selects a subset of cluster members, predictive modeling of the tagged cases is a precursor to predictive modeling using actual fraud cases. You could build a predictive model using the 71 cases identified above as proxies for fraud. As actual fraud cases are discovered and added to the data, you will transition from an unsupervised learning approach to a supervised learning approach. However, there should always be an unsupervised learning aspect to the analytics, because criminals are always thinking of new and innovative ways to commit fraud.

4.06 Poll

Have you ever used clustering techniques to find unusual observations in a data set?

- Yes
- No

40

Applications of Outlier Detection

In this area...	...outliers could indicate...
customer support queries	phishing by competitors
school district student essays	plagiarism or teacher misconduct
warranty reports	fraud
electronic communication	terrorist threats
college applications	creative applicants

41

Too often, outlier detection is viewed as “finding bad data.” This section has shown that outlier detection is more appropriately characterized as “finding **unusual** data.” Unusual data can be bad data arising from coding errors, or it could be an indication of anomalous behavior.

4.3 Association and Sequence Discovery in Text Analytics

Objectives

- Examine the SAS Text Miner transaction data set exported by several of the SAS Text Miner nodes.
- Show how transaction-based analysis, such as association analysis, can be applied to text mining transactional data.

43

Text Mining

A solution in search of a problem....



44

Market Basket Analysis

- A document is a market basket.
- Terms are the items in the market basket.



45

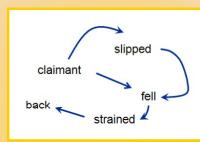
Association and Sequence Discovery

Association Discovery

- Terms are in a market basket (document).

Sequence Discovery

- Terms are arranged in a certain order, a sequence, in a document.
- The order of the terms represents a path traveled by the process to retrieve information.



46

The SAS Text Miner Transaction Data Set

Variables

- **_DOCUMENT_**: Document ID
- **_TERMINUM_**: Term ID
- **_COUNT_**: Relative frequency of term for given document

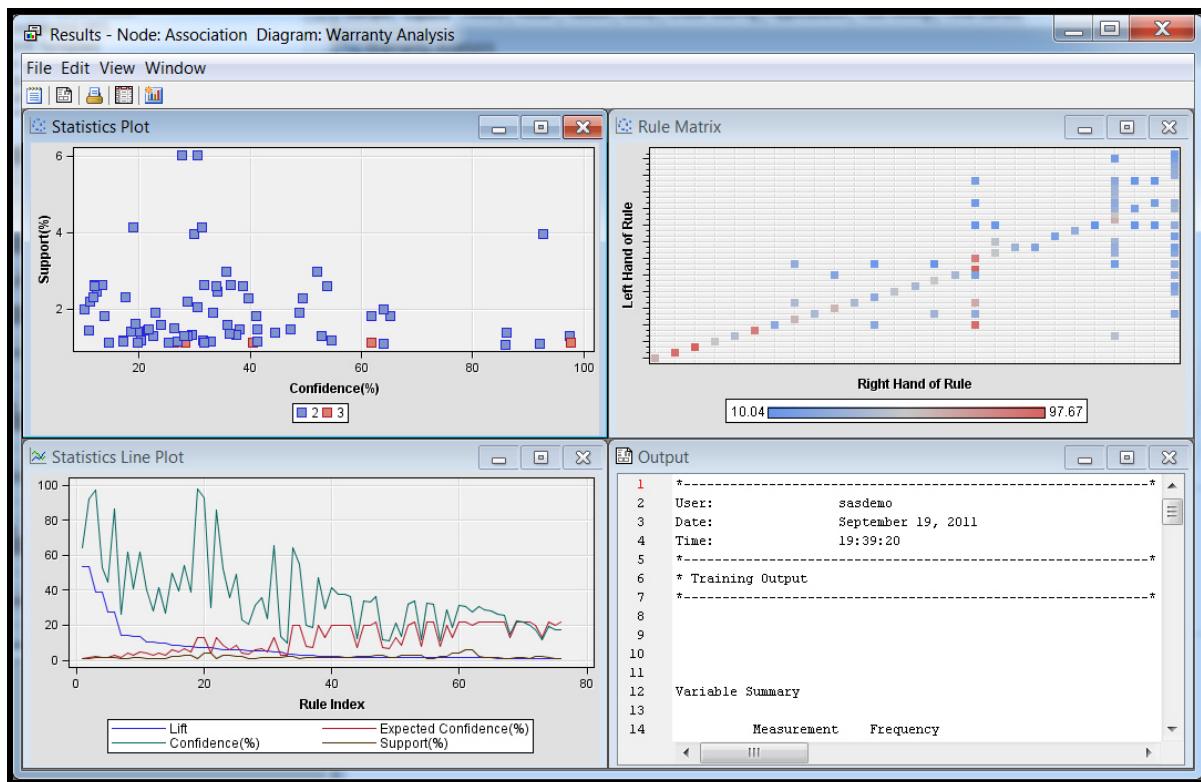
Sequence information is not available.



Association Discovery of Terms

This demonstration illustrates how to use the SAS Text Miner transaction data set to perform an association discovery analysis.

1. Use the Warranty Analysis diagram. Attach a SAS Code node to the Text Topic node. Open the Code Editor. In the Editor pane, open the program **SCN_MergeTerms.sas**. This program converts the term *ID* to the term *name*. Otherwise, the association rules would not be very useful – for example, term 37 \Rightarrow term 129. Save the code and close the editor.
2. Attach an Association node from the Explore tab to the SAS Code node. Run the Association node.



3. Expand the Output window and scroll down to the association rules.

Association Report											
Relations	Expected			Transaction			Left Hand Rule				Right Hand Rule
	Confidence (%)	Confidence (%)	Support (%)	Lift	Count	Rule	of Rule	of Rule	Item 1	Rule Item 2	
23	2	1.20	63.94	1.10	53.34	289.00	cause ==> unknown	cause	unknown	cause	=====
24	2	1.73	92.04	1.10	53.34	289.00	unknown ==> cause	unknown	cause	unknown	=====
25	2	2.50	97.46	1.32	39.04	345.00	bag ==> air	bag	air	bag	=====
26	2	1.35	52.75	1.32	39.04	345.00	air ==> bag	air	bag	air	=====
27	2	1.63	44.50	1.40	27.30	368.00	fire ==> catch	fire	catch	fire	=====
28	2	3.16	86.18	1.40	27.30	368.00	catch ==> fire	catch	fire	catch	=====
29	3	1.81	26.27	1.12	14.49	294.00	do ==> work & not do	work & not do	work & not do	do	=====
30	3	4.27	61.89	1.12	14.49	294.00	work & not ==> do	work & not do	work	not	=====
31	2	2.97	41.08	1.84	13.83	481.00	seat ==> belt	seat	belt	seat	=====
32	2	4.47	61.83	1.84	13.83	481.00	belt ==> seat	belt	seat	belt	=====
33	3	3.96	40.38	1.12	10.20	294.00	work ==> not & do work	not & do work	not & do	work	=====
34	3	2.78	28.35	1.12	10.20	294.00	not & do ==> work	not & do work	not	do	=====
35	2	4.27	41.35	1.15	9.68	301.00	work ==> do	work	do	work	=====
36	2	2.78	26.90	1.15	9.68	301.00	do ==> work	do	work	do	=====
37	2	5.78	49.42	2.29	8.55	600.00	on ==> light	on	light	on	=====
38	2	4.63	39.60	2.29	8.55	600.00	light ==> on	light	on	light	=====
39	2	6.73	53.91	2.61	8.01	683.00	side ==> driver	side	driver	side	=====
40	2	4.84	38.72	2.61	8.01	683.00	driver ==> side	driver	side	driver	=====
41	3	13.24	97.67	1.12	7.38	294.00	work & do ==> not work & do	not work & do	not work	do	=====
42	2	13.24	92.67	3.96	7.00	1037.0	do ==> not	do	not	do	=====
43	2	4.27	29.90	3.96	7.00	1037.0	not ==> do	not	do	not	=====
44	2	13.24	85.84	1.09	6.48	285.00	properly ==> not properly	not properly	not	properly	=====
45	2	8.34	52.00	2.97	6.23	779.00	while ==> driving	while	driving	while	=====
46	2	5.72	35.65	2.97	6.23	779.00	driving ==> while	driving	while	driving	=====
47	2	8.34	48.92	1.91	5.87	500.00	mph ==> driving	mph	driving	mph	=====

Nothing exciting leaps out of the analysis. This example is not compelling, but association discovery is a useful tool to have in your text analytics bag of tricks.



Exercises

1. Warranty Analysis

Use the diagram from the automotive warranty demonstration. Use the Text Filter node Interactive Filter Viewer to search for the following terms: *danger, fire, accident*. Do any claims arise that appear to be relevant for an early warning system as required by the U.S. TREAD Act?

2. Fraud Detection

Using the same methodology employed in the warranty analysis demonstration, verify that the 10 derived clusters are stable.

3. Text Categorization

The movies data set **LWDMTXT.MOVIESGENRE** was introduced in a previous exercise.

- a. Using the Text Topic node, define two topics: macho movies and date movies. Macho movies are movies that would be enjoyed by “macho” males, such as action and adventure movies. Date movies are movies that typically might be appropriate for a first date, such as romantic comedies. Here dates and dating refer to a romantic social engagement between two persons.
- b. Identify the three top macho movies and the three top date movies using your definitions. Recall that topic creation is often an interactive activity that proceeds through several iterations. The iterative process is necessary to fine-tune definitions to the actual corpus being studied. You can use any properties that you think are appropriate for this problem.

4.4 Chapter Summary

Text mining can be applied to numerous pattern discovery problems.

The analysis of text fields from call center data can lead to improved response times. The analysis can also reveal trends in call center operations.

The SAS course description project illustrates how text mining can be used to create automated referral systems.

The ability to chain Text Miner nodes enables more complex strategies for grouping documents.

The *bag-o'-words* approach to text mining suggests an analogy to market basket analysis. The transaction data set produced by Text Miner nodes can feed directly into the Association node for an association discovery analysis.

For Additional Information

Jurafsky, Daniel, and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, New Jersey: Prentice Hall.

Manning, Christopher D., and Hinrich Schutze. 2002. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.

Sanders, Annette, and Craig DeVault. 2004. “Using SAS® at SAS: The Mining of SAS Technical Support.” paper 010-29, SUGI 29, Montréal, Quebec.

SAS Institute Inc. 2009. *Getting Started with SAS® Text Miner 4.1*. Cary, North Carolina: SAS Institute Inc.

Wakefield, Todd. 2004. “A Perfect Storm is Brewing: Better Answers are Possible by Incorporating Unstructured Data Analysis Techniques.” *DM Direct*, August 2004.

Wallace, John, and Tracy Cermack. 2004. “Text Mining Warranty and Call Center Data: Early Warning for Product Quality Awareness.” paper 003-29, SUGI 29, Montréal, Quebec.

4.5 Solutions

Solutions to Exercises

1. Warranty Analysis

The result for danger (just the top part of the data):

The screenshot shows two windows from the Interactive Filter Viewer application.

Documents Window:

TEXT OF COMPLAINT	TEXTFILTER2_SNIPPET	TEXTFILTER2_RELEVANCE	CLAIM ID	MILEAGE	REF
Concerned with dangers in suv's.	... Concerned with dangers in	1.0	4849828...	21241	
Right, rear seat belt in 7 passenger mini-van will not retract, therefore puts an	... row seat in danger	1.0	7256150...	112	
Complaint concerning the dangers on and off highway ramp in durant and	... Complaint concerning the	1.0	7669101...	3815	
Erratic fuel gauge, danger of running out of fuel on a dangerous situation.	... fuel gauge , danger of	1.0	7853636...	14730	
Transmission slipped while the car running on freeway, make it slow down,	... down , very danger , or get	1.0	8555675...	380	
Wheel bearings eroded after only 75k miles; tires in danger of falling off.	... ; tires in danger of falling	1.0	8985536...	4204	
Steering terminal came off making the vehicle go in danger.	... vehicle go in danger	1.0	9642173...	2624	

Terms Window:

TERM ▲	FREQ	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
dance	0	0	<input type="checkbox"/>	0.0	Noun	Alpha
dand	0	0	<input type="checkbox"/>	0.0	Noun	Alpha
danger	7	7	<input checked="" type="checkbox"/>	0.0	Noun	Alpha
dangerious	0	0	<input type="checkbox"/>	0.0	Prop	Alpha
dangerous	1	1	<input type="checkbox"/>	0.0	Adj	Alpha
dangerous am...	0	0	<input type="checkbox"/>	0.0	Noun Group	Alpha
dangerous bra...	0	0	<input type="checkbox"/>	0.0	Noun Group	Alpha
dangerous con...	0	0	<input type="checkbox"/>	0.0	Noun Group	Alpha
dangerous des...	0	0	<input type="checkbox"/>	0.0	Noun Group	Alpha
dangerous drive	0	0	<input type="checkbox"/>	0.0	Noun Group	Alpha
dangerous due	0	0	<input type="checkbox"/>	0.0	Noun Group	Alpha

The result for fire:

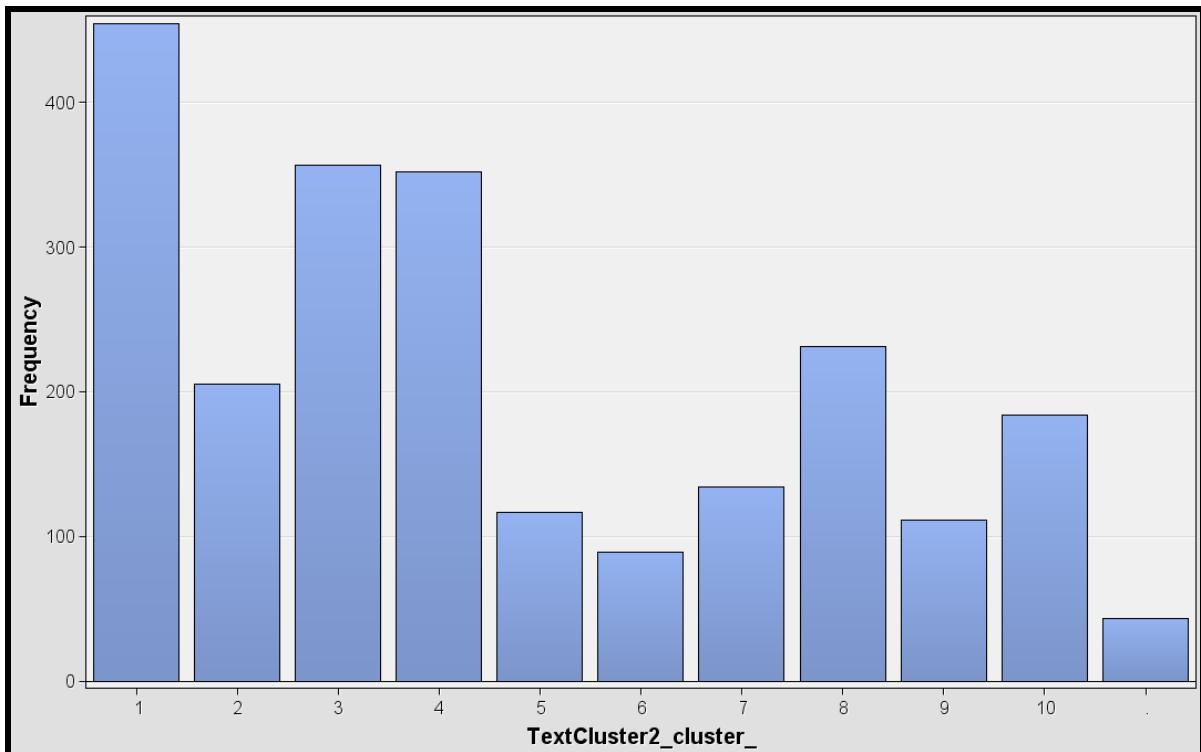
The screenshot shows the 'Interactive Filter Viewer' application window. At the top, there is a menu bar with File, Edit, View, Window, and a toolbar with a magnifying glass icon, a search input field containing 'fire', and buttons for Apply and Clear. Below the toolbar is a section titled 'Documents' with a table. The table has columns: TEXT OF COMPLAINT, TEXTFILTER2_SNIPPET, TEXTFILTER2_RELEVANCE, CLAIM ID, and MILEA... (partially visible). The data in the table includes various complaints related to fires, such as 'Alarm went off and truck caught a fire; local fire department put fire out.' and 'Underhood fire due to wires overheating /burning. Fire department came out'. The relevance score for most entries is 1.0. In the bottom half of the window, there is another table titled 'Terms' with columns: TERM ▲, FREQ, # DOCS, KEEP, WEIGHT, ROLE, and ATTRIBUTE. The 'fire' entry is highlighted with a checked 'KEEP' checkbox and a weight of 0.019.

The result for accident:

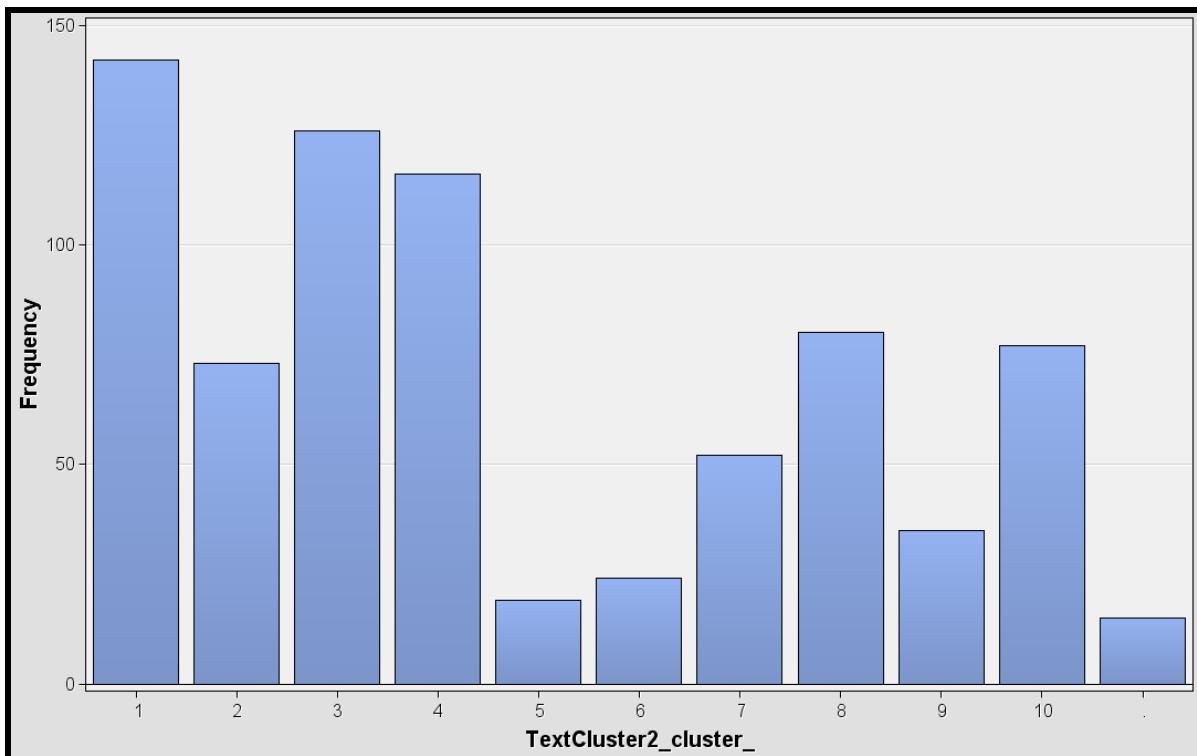
The screenshot shows the 'Interactive Filter Viewer' application window with a search term of '>#accident' entered. The interface is similar to the previous one, with a 'Documents' table and a 'Terms' table below it. The 'Documents' table lists various accidents, such as 'Passenger's side airbag deployed by itself, no accident involved.' and 'Air bag deployed without accident or collision.', along with their corresponding snippets and relevance scores. The 'Terms' table shows the term 'accident' with a frequency of 274, appearing in 273 documents, and a weight of 0.0.

2. Fraud Detection

Produce a bar plot of the cluster distribution for the training and validation data. From the property panel for the Text Cluster node that produces 10 clusters, select **Exported Data**. Select the **TRAIN** data, and click the **Explore** button. Using the plot wizard, construct a bar chart for the variable **TextCluster2_cluster_**. Make sure that **TextCluster2_cluster_** is the category variable.



Follow the same steps for the **VALIDATE** data.



There are slight discrepancies, most notable for cluster 5, but overall the clustering seems to be stable.

3. Text Categorization

- a. A sample topic table appears below. This table is part of the course data, stored as SAS table DMTXT.MDTOPICS. Answers will vary.

	topic	_term_	_role_	_weight_
1	Date	affectionate	noun	0.8
2	Date	bride	noun	0.7
3	Date	couple	noun	0.5
4	Date	family	noun	0.2
5	Date	friend	noun	0.2
6	Date	friendship	noun	0.5
7	Date	kiss	verb	0.5
8	Date	love	noun	0.9
9	Date	love	verb	0.9
10	Date	married	adj	0.4
11	Date	relationship	noun	0.4
12	Date	romance	noun	0.7
13	Date	romantic	adj	0.7
14	Date	true love	NOUN_GROUP	1
15	Date	wedding	noun	0.5
16	Macho	alien	noun	0.5
17	Macho	angry	adj	0.2
18	Macho	battle	noun	0.8
19	Macho	cowboy	noun	0.6
20	Macho	crime	noun	0.7
21	Macho	criminal	noun	0.7
22	Macho	detective	noun	0.7
23	Macho	fight	noun	0.8
24	Macho	fight	verb	0.8
25	Macho	gun	noun	0.8
26	Macho	lethal	adj	0.4
27	Macho	monster	noun	0.7
28	Macho	murder	noun	0.7
29	Macho	police	noun	0.7
30	Macho	prison	noun	0.4
31	Macho	shoot	verb	0.3
32	Macho	war	noun	0.8
33	Macho	weapon	noun	0.6

- b. The topic viewer for macho movies:

The screenshot shows the "Interactive Topic Viewer" application window with three main panels:

- Topics:** A table showing topics and their characteristics. The columns are Topic, Category, Term Cutoff, and Document Cutoff.

Topic	Category	Term Cutoff	Document Cutoff
Date	User	0.001	0.001
Macho	User	0.001	0.001
+show,+rate,+recommend,+script,+sex	Mult	0.084	0.801
hollywood,+order,later,+hand,+cinema	Mult	0.073	0.678
+viewer,+moment,+minute,+relationship,+i	Mult	0.072	0.649

- Terms:** A table showing terms and their properties. The columns are Topic Weight, +, Term, Role, # Docs, and Freq.

Topic Weight	+	Term	Role	# Docs	Freq
0.8	+	war	Noun	149	264
0.8	+	fight	Verb	146	179
0.8	+	battle	Noun	105	159
0.8	+	gun	Noun	98	132
0.8	+	fight	Noun	80	88

- Documents:** A table showing document details. The columns are Topic Weight, Synopsis, Title, Action, Comedy, and Crime.

Topic Weight	Synopsis	Title	Action	Comedy	Crime
0.398	Since buddy movies, cop	Showtime	1.0	1.0	0.0
0.347	"My mommy always said	Alien Resurrection	1.0	0.0	0.0
0.334	School shootings and similar	Past Perfect	1.0	0.0	0.0
0.296	THE LAST SAMURAI stars	Last Samurai, The	1.0	0.0	0.0
0.291	In ENEMY AT THE GATES, by	Enemy at the Gates	1.0	0.0	0.0
0.289	'Remember "Catch 22" and	Three Kings	0.0	0.0	0.0

The topic viewer for data movies:

The screenshot shows the "Interactive Topic Viewer" application window with three main panels:

- Topics:** A table showing topics with their category, term cutoff, and document cutoff. The topics listed are Date, Macho, +show,+rate,+recommend,+script,+sex, hollywood,+order,later,+hand,+cinema, and +viewer,+moment,+minute,+relationship,+i.

Topic	Category	Term Cutoff	Document Cutoff
Date	User	0.001	0.001
Macho	User	0.001	0.001
+show,+rate,+recommend,+script,+sex	Mult	0.084	0.801
hollywood,+order,later,+hand,+cinema	Mult	0.073	0.678
+viewer,+moment,+minute,+relationship,+i	Mult	0.072	0.649

- Terms:** A table showing terms with their topic weight, role, number of documents, and frequency. The terms listed are true love, love, love, romance, and romantic.

Topic Weight	+	Term	Role	# Docs	Freq
1		true love	Noun Group	18	19
0.9	+	love	Noun	372	557
0.9	+	love	Verb	263	331
0.7	+	romance	Noun	161	204
0.7		romantic	Adj	130	141

- Documents:** A table showing movie titles with their synopsis, action, comedy rating, and other details. The movies listed are When casting a light comedy, Wedding Planner, The; In Mira Nair's MONSOON; You've had a hard day, and Don Juan DeMarco; When Coleman Silk (Anthony; From the start, director Nick; and BEFORE SUNRISE is a.

Topic Weight	Synopsis	Title	Action	Comedy	C
0.377	When casting a light comedy,	Wedding Planner, The	0.0	1.0	0.0
0.373	In Mira Nair's MONSOON	Monsoon Wedding	0.0	0.0	0.0
0.359	You've had a hard day, and	Don Juan DeMarco	0.0	1.0	0.0
0.273	When Coleman Silk (Anthony	Human Stain, The	0.0	0.0	0.0
0.262	From the start, director Nick	Notebook, The	0.0	0.0	0.0
0.261	BEFORE SUNRISE is a	Before Sunrise	0.0	1.0	0.0

Solutions to Student Activities (Polls/Quizzes)

4.03 Multiple Answer Poll – Correct Answers

Select all items that represent properties of the Score node.

- a. can score the following data set types: raw, train, validate, test, score
- b. produces score code written in the SAS language
- c. produces text mining score code that will run on any platform that has a valid Base SAS license
- d. exports SAS data sets containing columns with scores and scoring related values
- e. scores cases into clusters using clustering algorithms even if predictive modeling was not performed

18

4.05 Poll – Correct Answer

Insurance Company X contracts with your consulting firm to derive a fraud detection model. You are supplied with 800 paper reports that describe successful fraud investigations over the past five years. After constructing a data table based on the contents of the reports, you perform a quick exploratory analysis that reveals that 603 of the 800 fraud cases involve a soft-tissue back injury. Do you agree with the following statement?

The data is **insufficient** to suggest that soft-tissue back injuries are indicative of fraud.

- Yes
- No

27

Chapter 5 Applications of Text Mining to Predictive Modeling

5.1 Predictive Modeling with SAS Enterprise Miner.....	5-3
5.2 Using Adjustor Notes to Predict Recovery Potential in Insurance Claims	5-17
Demonstration: Predicting Workers' Compensation Recovery Potential.....	5-20
5.3 Text Categorization via Predictive Modeling.....	5-25
Demonstration: Text Categorization of the ASRS Data	5-28
Exercises	5-41
5.4 Chapter Summary.....	5-42
5.5 Solutions	5-43
Solutions to Exercises	5-43
Solutions to Student Activities (Polls/Quizzes)	5-44

5.1 Predictive Modeling with SAS Enterprise Miner

Objectives

- Describe predictive modeling data sets.
- Explain predictive modeling projects and features of SAS Enterprise Miner related to predictive modeling.

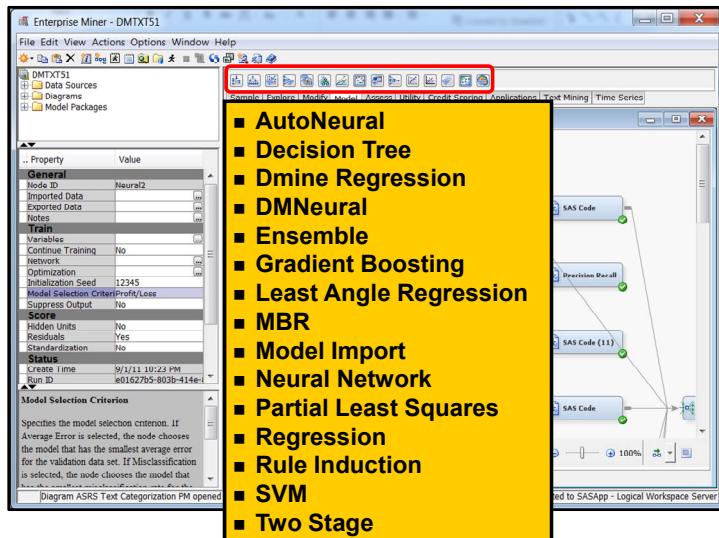
5.01 Multiple Choice Poll

Which choice most closely matches your interest in text mining predictive modeling?

- My primary project is a predictive modeling project.
- Predictive modeling might become important as I gain experience with text mining.
- Others in my organization are responsible for predictive modeling.
- I do not anticipate using text mining for predictive modeling.

5

SAS Enterprise Miner Model Tab



6

SAS Enterprise Miner has 15 model nodes. Some nodes are general purpose, such as the Decision Tree, Neural Network, and Regression nodes. Some nodes are specialized, such as the MBR, Rule Induction, and Partial Least Squares nodes. This course illustrates use of the Decision Tree, MBR, Neural Network, and Regression nodes.

Predictive Modeling Training Data

Training Data

	inputs		target

Training data case: categorical or numeric input and target measurements

7

The minimum requirement for data mining predictive modeling is at least one target variable and at least one input variable. A predictive model is constructed using a training data set. The model attempts to predict the value of the target variable using only the values of a set of input variables. For example, input variables can measure customer attributes such as gender, age, income, location of primary residence, and average purchases to try to estimate the probability that a customer will respond to a particular promotion, such as a 20% off discount on purchases of \$100 or more.

Predictive Model

Training Data

	inputs		target

Predictive model: a concise representation of the input and target association

8

After a model is constructed using training data, the accuracy of the model can be assessed using a holdout data set. If the model is judged to be accurate, it can be used to score new data to determine, for example, which customers should be selected for a promotional offer. The term *score* is synonymous with *predict*.

Predictive Model

Training Data

Validation Data

Test Data

Score Data

	inputs	

	prediction

Predictions: output of the predictive model given a set of input measurements

9

To choose from a variety of models, a holdout data set called a *validation* data set is used to determine how well models will extrapolate to new data. This helps overcome the problem of *overfitting*, which occurs when a model is constructed to fit the training data set so well that it does not fit any other data well. For a model that has been selected for deployment, a second holdout data set, called a *test* data set, is employed to get an unbiased estimate of the accuracy of the model in the live environment. A predictive

model can score any data set that has the inputs used by the model. It is important to ensure that models are applied to data commensurate with how a model was constructed. For example, a model constructed using only customers who reside in California might not be appropriate for scoring customers in Florida.

SAS Enterprise Miner Source Data

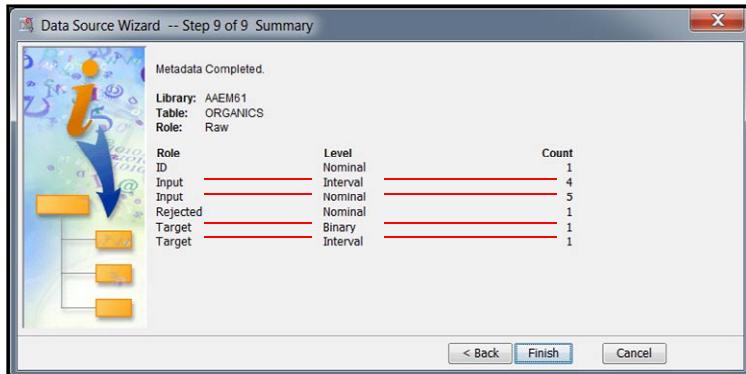
Name	Role	Level	Report	Order	Drop	Lower Limit
DemAffl	Input	Interval	No	No	No	.
DemAge	Input	Interval	No	No	No	.
DemCluster	Rejected	Nominal	No	No	No	.
DemClusterGrp	Input	Nominal	No	No	No	.
DemGender	Input	Nominal	No	No	No	.
DemReq	Input	Nominal	No	No	No	.
DemTVReq	Input	Nominal	No	No	No	.
ID	ID	Nominal	No	No	No	.
PromClass	Input	Nominal	No	No	No	.
PromSpend	Input	Interval	No	No	No	.
PromTime	Input	Interval	No	No	No	.
TargetAmt	Target	Interval	No	No	No	.
TargetBuy	Target	Binary	No	No	No	.

For predictive modeling:

- at least one variable with the role **Target**
- at least one variable with the role **Input**

On the slide above, the table has nine input variables and two target variables. If you want to predict two or more target variables, you might have to use multiple model nodes. For example, you could use a decision tree to predict one target, and a neural network to predict the other target. Some model nodes, such as the Neural Network node, can build models for multiple targets. This is usually a positive feature, but caution must be exercised, because there are many situations where the inputs used to predict one target variable might not be appropriate for predicting another target variable.

SAS Enterprise Miner Source Data



The Data Source Wizard summarizes the metadata.
There are nine inputs and two targets.

11

When you create a data source in SAS Enterprise Miner, the summary window alerts you to variable roles so you can make sure that you have the target and input variables you need to build a predictive model.

SAS Enterprise Miner Predictions Data

Name	Use	Report	Role	Level
AdjusterNotes	Default	No	Text	Nominal
BP_SUBROFLAG	Default	No	Assessment	Interval
Body	Default	No	Input	Nominal
CP_SUBROFLAG	Default	No	Assessment	Interval
Cause	Default	No	Input	Nominal
Clamino	Default	No	ID	Nominal
D_SUBROFLAG	Default	No	Decision	Nominal
EP_SUBROFLAG	Default	No	Assessment	Interval
F_SubroFlag	Default	No	Classification	Nominal
FraudFlag	Default	No	Rejected	Binary
L_SubroFlag	Default	No	Classification	Nominal
Nature	Default	No	Input	Nominal
P_SubroFlag0	Default	No	Prediction	Interval
P_SubroFlag1	Default	No	Prediction	Interval
Q_SubroFlag0	Default	No	Input	Interval
Q_SubroFlag1	Default	No	Input	Interval
R_SubroFlag0	Default	No	Residual	Interval
R_SubroFlag1	Default	No	Residual	Interval
SubroFlag	Default	No	Target	Binary
U_SubroFlag	Default	No	Classification	Nominal
VEHflag	Default	No	Rejected	Binary
V_SubroFlag0	Default	No	Prediction	Interval
V_SubroFlag1	Default	No	Prediction	Interval
NODE	Default	No	Segment	Nominal
WARN	Default	No	Assessment	Nominal
dataobs_	Default	No	ID	Interval

Decision Tree

12

A predictive model will provide a predicted value. In the case of a binary target variable, the predicted value is the posterior probability of the primary event given the inputs. You can use this probability to derive a decision rule, for example, if the probability exceeds 0.37, send the promotion to the customer; otherwise do nothing. SAS Enterprise Miner model nodes will add a variety of columns to the imported data when creating the exported data. The nature of the added columns depends on the model node employed. The above table is displayed by selecting the Variables... property in a SAS Code node attached to a Decision Tree node. The **Q_** and **V_** variables are unique to the Decision Tree node.

SAS Enterprise Miner Predictions Data

Regression

13

The Regression node adds some of the same columns as the Decision Tree node to the exported data, but notice that in particular there are no **Q_** or **V_** columns.

The following table, obtained from **Help** \Rightarrow **Contents**, summarizes columns that can appear in a prediction data set. Note that a prediction data set in SAS Enterprise Miner retains the data role of Train, Validate, Test, or Score.

Prefixes Commonly Used in Scored Data Sets

Prefix	Root	Description	Target Needed?
BL_	Decision data set	Best possible loss of any of the decisions, $-B(i)$	Yes
BP_	Decision data set	Best possible profit of any of the decisions, $B(i)$	Yes
CL_	Decision data set	Loss computed from the target value, $-C(i)$	Yes
CP_	Decision data set	Profit computed from the target value, $C(i)$	Yes
D_	Decision data set	Label of the decision chosen by the model	No
E_	Target	Error function	Yes
EL_	Decision data set	Expected loss for the decision chosen by the model, $-E(i)$	No
EP_	Decision data set	Expected profit for the decision chosen by the model, $E(i)$	No
F_	Target	Normalized category that the case comes from	Yes
I_	Target	Normalized category that the case is classified into	No
IC_	Decision data set	Investment cost $IC(i)$	No
M_	Variable	Missing indicator dummy variable	—
P_	Target or dummy	Outputs (predicted values and posterior probabilities)	No
R_	Target or dummy	Plain residuals: target minus output	Yes
RA_	Target	Anscombe residuals	Yes
RAS_	Target	Standardized Anscombe residuals	Yes
RAT_	Target	Studentized Anscombe residuals	Yes
RD_	Target	Deviance residuals	Yes
RDS_	Target	Standardized deviance residuals	Yes
RDT_	Target	Studentized deviance residuals	Yes
ROI_	Decision data set	Return on investment, $ROI(i)$	Yes
RS_	Target	Standardized residuals	Yes
RT_	Target	Studentized residuals	Yes
S_	Variable	Standardized variable	—
T_	Variable	Studentized variable	—
U_	Target	Unformatted category that the case is classified into	No

Prediction Types for Binary Response Models

Decisions:

- I_Target is 1 if P_Target1>P_Target0. It is 0 (zero) otherwise. Thus, I_Target decisions are equivalent to using a posterior probability cutoff of 50%.
- D_Target is 1 if Profit(Target=1)>Profit(Target=0). It is 0 (zero) otherwise. If no profit information is provided, then D_Target is equivalent to I_Target.

Estimates:

- P_Target1 is the estimated posterior probability that Target=1.
- P_Target0 is the estimated posterior probability that Target=0.

14

The above slide summarizes the prediction variables that are usually of interest. A Decision Tree node would produce, for example, Q_Target1, which would be the posterior probability derived from the tree prior to correcting for oversampling. Note that if the binary target variable has the name **FLOYD**, then the posterior probability that FLOYD=1 is given by the variable **P_FLOYD1**.

Inputs Derived Using SAS Text Miner

- Text Cluster node SVD variables
- Text Cluster node segment variable changed to a nominal input
- Text Topic node raw SVD variables
- Text Topic node binary topic class labels, changed from segment to input variables

15

There are four different types of input variables that can be produced by SAS Text Miner. Text mining can improve predictive modeling results when few input variables are available. If text information is available, such as descriptions of warranty claims or physician reports of injury, then the text can be used to add input variables to an analysis.

SAS Enterprise Miner Input Selection

Explore Tab

- Variable Clustering node
- Variable Selection node

Model Tab

- Decision Tree node
- Regression node

16

While lack of input variables can be a problem, another problem is having too many inputs. Input selection, or variable selection, is an important topic in predictive modeling. SAS Enterprise Miner facilitates input selection in a number of different ways.

SAS Enterprise Miner Dimensionality Reduction

Explore Tab

- Variable Clustering node

Modify Tab

- Principal Components

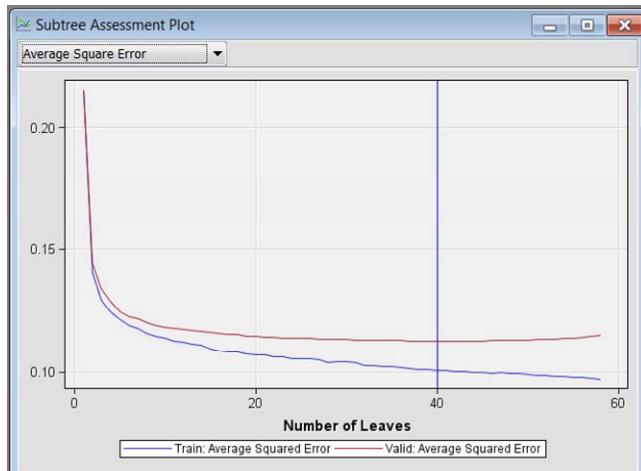
Model Tab

- Partial Least Squares node

17

Dimensionality reduction is like input selection in that models will have fewer parameters. However, dimensionality reduction strives to preserve as much information as was present in the original input variables while reducing the size of a model that uses the inputs.

Model Selection=Input Selection

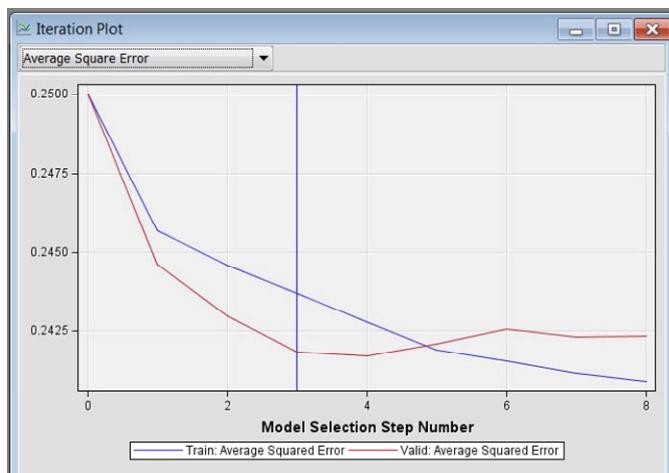


Decision Tree Subtree Assessment Plot

18

A Decision Tree node performs input selection by deciding which subset of input variables will be used to partition the data into separate leaves. If a variable is not useful in separating the data into more pure leaf nodes, then the variable is discarded. In the above plot, a tree with 40 leaves is derived. A 59-leaf tree was pruned to remove leaves that did not improve overall model accuracy. The pruned subtree is used to pick the input variables that are useful for prediction. These variables will be passed to successor nodes, while the other input variables will have their roles changed to Rejected.

Model Selection=Input Selection



Regression Iteration Plot

19

The Regression node also has options for performing variable selection. The above iteration plot reveals that the Regression node tried a series of models culminating in an eight-variable model, but it chose as a final model one that has only three variables.

Model Assessment

- Fit Statistics
 - Average square error
 - Misclassification rate
 - Information criteria
 - Others
- Charts and Plots
 - ROC chart
 - Gains chart
 - Lift chart
 - Others

20

Model assessment is performed using results from the model nodes and using the Model Comparison node.

Model Assessment: Fit Statistics

Target	Fit Statistics	Statistics Label	Train	Validation	Test
Target19	_NOBS_	Sum of Frequencies	16139	5380	.
Target19	_MISC_	Misclassification Rate	0.13204	0.150743	.
Target19	_MAX_	Maximum Absolute Error	0.975649	0.975649	.
Target19	_SSE_	Sum of Squared Errors	3250.704	1213.841	.
Target19	_ASE_	Average Squared Error	0.10071	0.112811	.
Target19	_RASE_	Root Average Squared Error	0.317348	0.335873	.
Target19	_DIV_	Divisor for ASE	32278	10760	.
Target19	_DFT_	Total Degrees of Freedom	16139	.	.

Decision Tree Node Fit Statistics Table

21

Fit statistics, such as average square error (ASE), are defined in the SAS Enterprise Miner documentation, which you can access by selecting **Help** ⇒ **Contents**.

Model Assessment: Fit Statistics

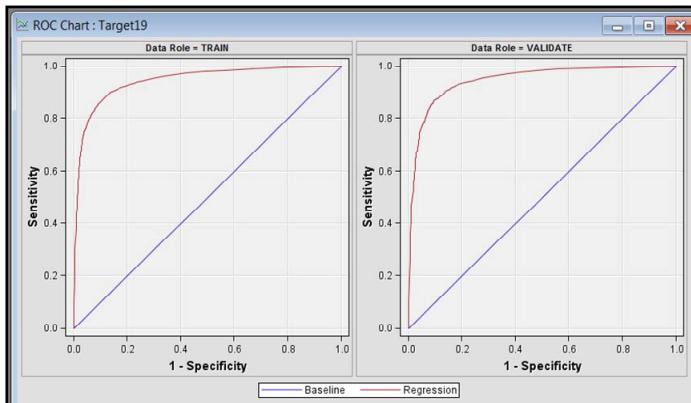
Target	Fit Statistics	Statistics Label	Train	Validation	Test
Target19	DFT_	Total Degrees of Freedom	16139	.	.
Target19	DFE_	Degrees of Freedom for Error	15985	.	.
Target19	DFM_	Model Degrees of Freedom	154	.	.
Target19	NW_	Number of Estimated Weights	154	.	.
Target19	AIC_	Akaike's Information Criterion	9423.025	.	.
Target19	SBC_	Schwarz's Bayesian Criterion	10607.13	.	.
Target19	ASE_	Average Squared Error	0.08377	0.096703	.
Target19	MAX_	Maximum Absolute Error	0.994556	0.993259	.
Target19	DIV_	Divisor for ASE	32278	10760	.
Target19	NOBS_	Sum of Frequencies	16139	5380	.
Target19	RASE_	Root Average Squared Error	0.28943	0.310971	.
Target19	SSE_	Sum of Squared Errors	2703.924	1040.523	.
Target19	SUMW_	Sum of Case Weights Times Freq	32278	10760	.
Target19	FPE_	Final Prediction Error	0.085384	.	.
Target19	MSE_	Mean Squared Error	0.084577	0.096703	.
Target19	RFPE_	Root Final Prediction Error	0.292205	.	.
Target19	RMSE_	Root Mean Squared Error	0.290821	0.310971	.
Target19	AVERR_	Average Error Function	0.282391	0.320264	.
Target19	ERR_	Error Function	9115.025	3446.043	.
Target19	MISC_	Misclassification Rate	0.111593	0.129554	.
Target19	WRONG_	Number of Wrong Classifications	1801	697	.

Neural Network Node Fit Statistics Table

22

Different model nodes tend to produce different fit statistics.

Model Assessment: Charts and Plots

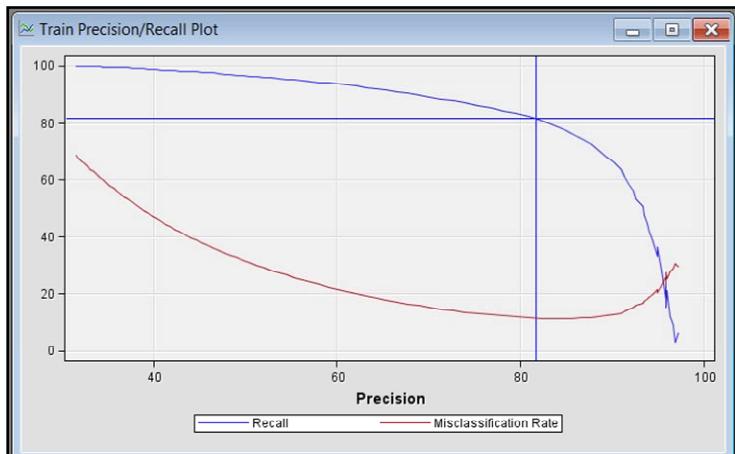


Model Comparison Node ROC Chart

23

The Model Comparison node produces charts and plots that might not have been produced by a model node.

Model Assessment: Charts and Plots



SAS Code Node Custom Plot

24

Some plots are not available, but SAS Enterprise Miner provides functionality for creating custom results. The above plot was produced using a SAS Code node. The plot shows break-even points for precision and recall for a text categorization problem.

The rest of this chapter provides examples of predictive modeling projects using text mining inputs.

5.2 Using Adjustor Notes to Predict Recovery Potential in Insurance Claims

Objectives

- Describe predictive modeling challenges associated with using text-based inputs.
- Illustrate how to build predictive models with text mining inputs using the Workers' Compensation Insurance data.

26

Problems and Pitfalls in Using Text-Based Inputs

- Deriving text-based inputs can be computationally intensive, making real time scoring difficult or impossible.
 - Stemming and part-of-speech tagging are difficult to program for generic real time scoring.
 - SAS Text Miner does not produce pure SAS language score code.
- The free-format nature of text can lead to temporal inconsistencies:
 - turnover in text writers (claims adjusters, loan officers, and so on)
 - new reporting requirements

27

Despite possible pitfalls, when compared to pattern discovery, predictive modeling with text mining inputs is relatively easy. Add text mining inputs to your other input variables, and then use variable-selection techniques to arrive at a good model. Predictive modeling can be challenging and time-consuming. The point is that adding text mining inputs does not add much to the challenge of finding a good model.

Workers' Compensation Claims

Analytic Objective: Predict recovery potential for open claims.

Challenges:

- Rare target
- Claims adjusters using nonstandard technical language and abbreviations

Data:

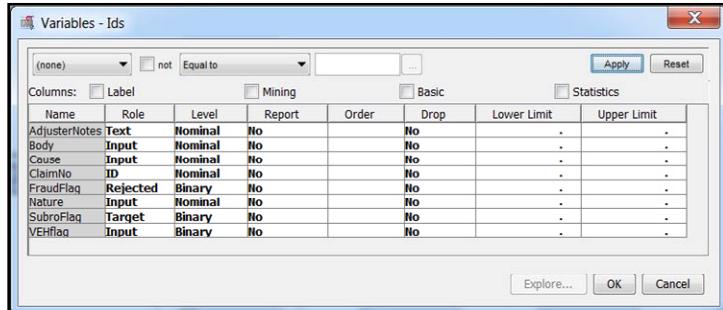
- 3,037 documents extracted from Lotus Notes
- Documents merged with claims master data using claim ID

28

continued...

This data set was collected, screened, and cleaned primarily using Base SAS software. The data was collected before SAS Text Miner existed.

Workers' Compensation Claims



Metadata

29

continued...

The target variable **FraudFlag** is set to **Rejected** for the recovery model. A preliminary analysis related to fraud detection was conducted as a project in text categorization in a previous chapter.

Workers' Compensation Claims

An employee alleges that while lowering a machine down into the basement on planks and rollers, it fell off the planks and partially hit him on the shoulder and side.

The employee was terminated, and the employer received knowledge of a claim via certified mail nine months later. The employee claims stress as well as stomach, back, and chest injuries.

Drink machine fell on employee.

The employee claims stress.

30

continued...

Workers' Compensation Claims

Data Preparation

- Remove proper names and other identifying information.
- Convert abbreviations.
- Correct misspellings.

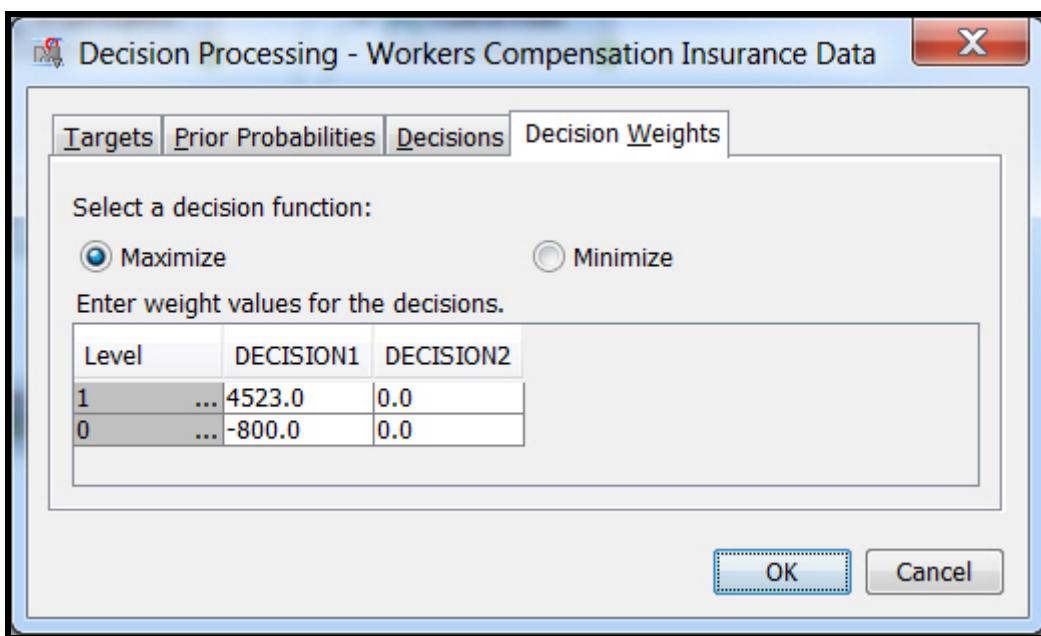
31



Predicting Workers' Compensation Recovery Potential

This demonstration illustrates how to use SAS Text Miner and SAS Enterprise Miner to predict recovery potential for Workers' Compensation Insurance claims.

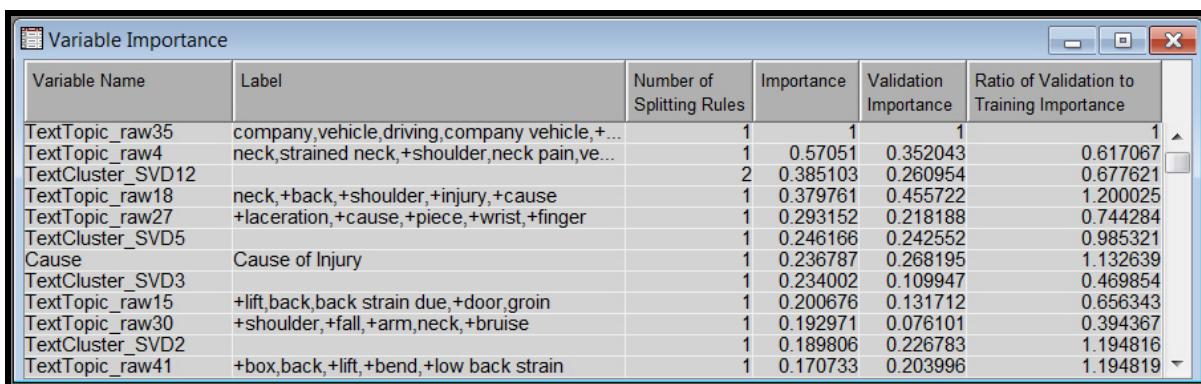
1. If a data source does not exist for **DMTXT.WORKCOMP**, then create one.
2. Create a diagram named Workers' Compensation Recovery Potential.
3. Drag the **WORKCOMP** input data source into the diagram. Select the **Decisions** property and select **Build**. The original data was oversampled to create the modeling data set. Specify prior probabilities of 0.07 for level 1 and 0.93 for level 0. Set decision weights as indicated in the following table, and then run the Input Data Source node.



The decision weights are based on past experience with investigating recovery cases.

4. Attach a Data Partition node to the Input Data Source node. Set the properties in the Data Set Allocations section as follows: Training=75%, Validation=25%, Test=0%. Run the Data Partition node.
5. Attach a Text Parsing node to the Data Partition node. Select **DMTXT.WORKCOMPSTOP** for the Stop List property. Set to Synonyms property to **No data set to be specified**. Run the Text Parsing node.
6. Attach a Text Filter node to the Text Parsing node. Set the Term Weight property to **Mutual Information**. This is the default setting for SAS Text Miner 5.1, but Entropy was the default setting for previous releases of SAS Text Miner even when a target variable was available. Run the Text Filter node.

7. Attach a Text Cluster node to the Text Filter node. Set the Exact or Maximum Number property to **Exact**. Set the Number of Clusters property to **6**. Run the Text Cluster node.
8. Attach a Text Topic node. Set the Number of Multi-term Topics property to **50**. Run the Text Topic node.
9. Attach a Metadata node to the Text Topic node. Change the variable role of the first 20 TextTopicM_N binary variables from Segment to **Input**. Run the Metadata node.
10. Attach an MBR (Memory Based Reasoning) node to the Metadata node. Change the Number of Neighbors property in the Train section to **8**. Select the **Variables** property, and change the Use status of all input variables to **Rejected**. Then change the Use status of all TextCluster_SVDn variables to **Yes**. These variables are orthogonal, and hence can be used as inputs to the MBR node. Otherwise, you would need to use a method such as principal components to convert inputs to orthogonal inputs. Run the MBR node.
11. Attach a decision tree node to the Metadata node. Change the Assessment Measure property in the Subtree section to **Average Square Error**. Run the Decision Tree node. Our primary goal is to build a successful scoring model, but it might be informative to examine how the decision tree chose to partition the data. Open the results window for the Decision Tree node. Select **View** \Rightarrow **Model** \Rightarrow **Variable Importance**.

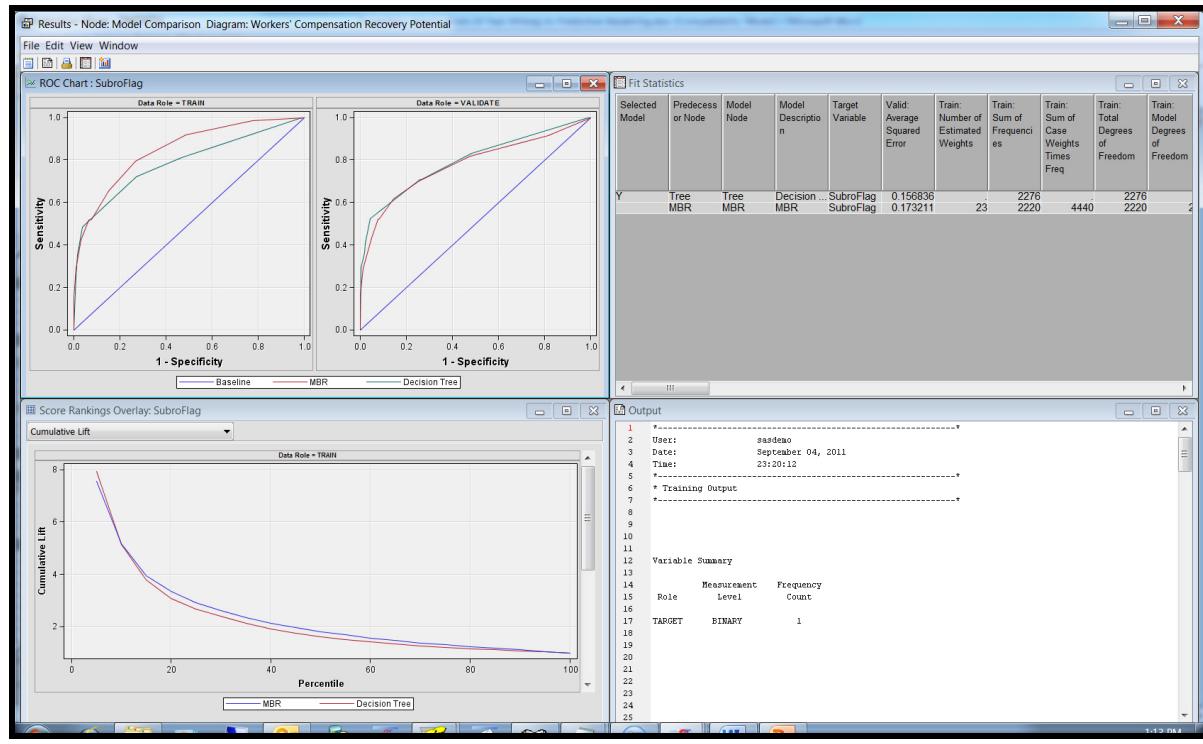


The screenshot shows a Windows application window titled "Variable Importance". The window contains a table with the following columns: Variable Name, Label, Number of Splitting Rules, Importance, Validation Importance, and Ratio of Validation to Training Importance. The table lists various variables and their corresponding values. The most important variable is "TextTopic_raw35" with an importance of 0.57051 and a validation importance of 0.352043. The "Cause" variable is also listed with an importance of 0.236787 and a validation importance of 0.268195.

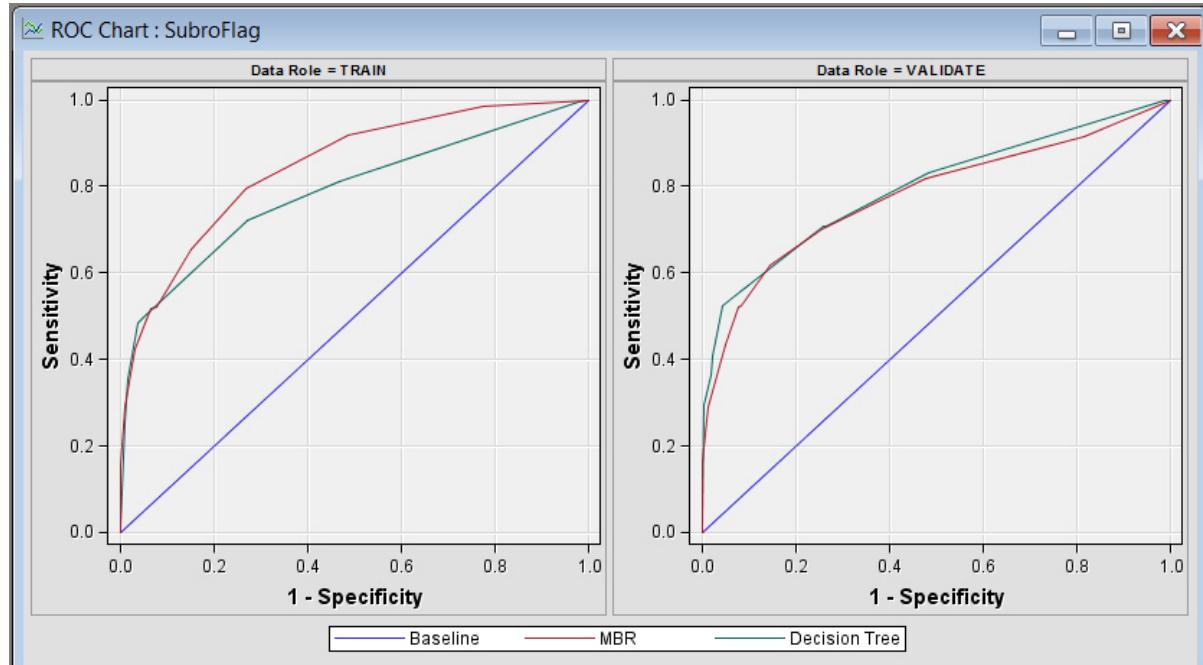
Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
TextTopic_raw35	company,vehicle,driving,company vehicle, +... neck,strained neck,+shoulder,neck pain,ve...	1	0.57051	0.352043	0.617067
TextCluster_SVD12		2	0.385103	0.260954	0.677621
TextTopic_raw18	neck,+back,+shoulder,+injury,+cause +laceration,+cause,+piece,+wrist,+finger	1	0.379761	0.455722	1.200025
TextTopic_raw27		1	0.293152	0.218188	0.744284
TextCluster_SVD5		1	0.246166	0.242552	0.985321
Cause	Cause of Injury	1	0.236787	0.268195	1.132639
TextCluster_SVD3		1	0.234002	0.109947	0.469854
TextTopic_raw15	+lift,back,back strain due,+door,groin	1	0.200676	0.131712	0.656343
TextTopic_raw30	+shoulder,+fall,+arm,neck,+bruise	1	0.192971	0.076101	0.394367
TextCluster_SVD2		1	0.189806	0.226783	1.194816
TextTopic_raw41	+box,back,+lift,+bend,+low back strain	1	0.170733	0.203996	1.194819

The only input that was not derived using text mining is CAUSE (Cause of Accident). The Text Topic node variables seem to contribute the most to the tree. The most important variable, **TextTopic_raw35**, is characterized by keywords such as “vehicle” and “driving,” suggesting a motor vehicle accident possibly caused by a third party, a common accident resulting in subrogation.

12. Attach the MBR node and the Decision Tree node to a Model Assessment node. Change the Selection Statistic property to **Average Square Error** and the Selection Table property to **Validation**. Run the Model Assessment node. Open the Results window.



The Tree model produces a smaller ASE value, and hence it is selected by the Model Comparison node. The ROC chart follows.

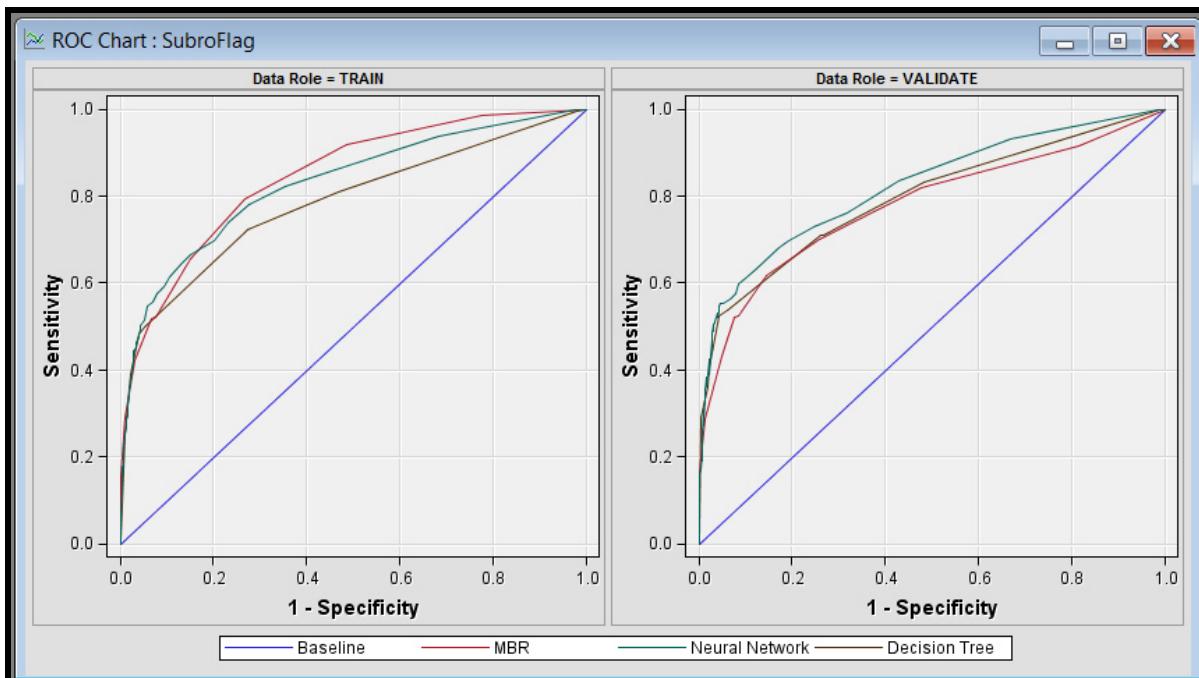


While the MBR model seems to be superior for the training data, both models are very similar when applied to the validation data. It is common practice in predictive modeling to use a neural network model as a *benchmark* model. Neural networks can be shown theoretically to be *universal approximators*, so they tend to produce the best results with the least effort. Even when professionals do not plan to deploy neural network models, they use them to determine what level of accuracy can be achieved, which establishes a baseline or a benchmark for other models to try to beat.

13. Attach a Neural Network node to the Metadata node. Change the Model Selection Criterion property to **Average Error**. Without any predecessor nodes to perform input selection, the Neural Network node will use all 98 input variables. This will require the estimation of 478 parameters, or almost one-fourth of the training data degrees of freedom. Run the Neural Network node.
14. Attach the Neural Network node to the Model Comparison node. Rerun the Model Comparison node. Open the Results window. The Fit Statistics table and the ROC chart for the three models follow.

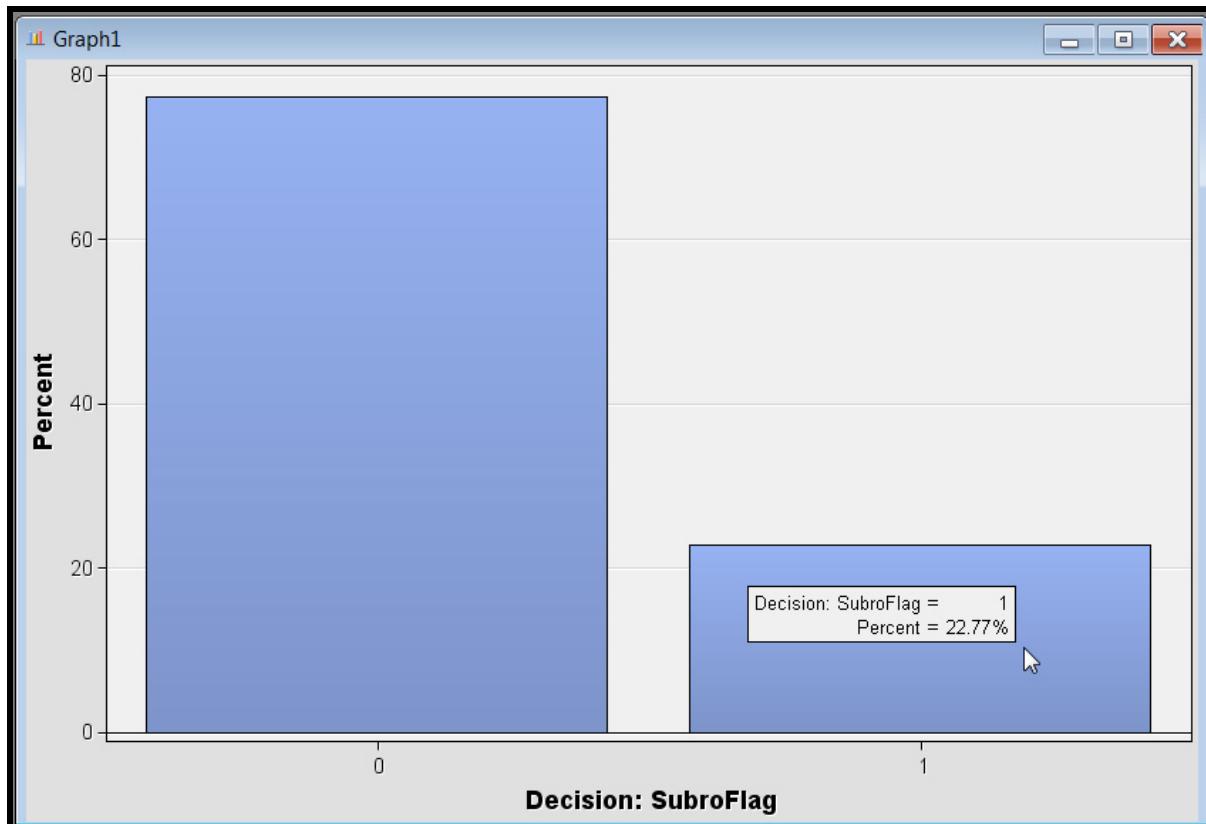
Selected Model	Predecessor or Node	Model Node	Model Description	Target Variable	Valid: Average Squared Error
Y	Neural	Neural	Neural Network	SubroFlag	0.150087
	Tree	Tree	Decision Tree	SubroFlag	0.156836
	MBR	MBR	MBR	SubroFlag	0.173211

The neural network model slightly beats the other two models with respect to ASE.



The neural network model appears to be superior to the other two models for any reasonable decision cutoff.

15. Because the model is used to make a decision whether to investigate an insurance claim for recoveries, there are no regulations that prohibit the use of a neural network model for this purpose. Thus, the model will be used to score new cases to decide whether to investigate a case for recoveries. Attach a Score node to the Model Comparison node.
16. Create a data source for the data set **DMTXT.WORKCOMPSCORE**. Make sure that the data role is set to **Score**. Drag the data source into the diagram and attach it to the Score node as an input – that is, connect from the data source to the Score node. Run the Score node. Select the **Exported Data** property, and then select the score data set. Explore the score data set. Create a bar chart for the **D_SUBROFLAG** variable.



Based on the specified decision criteria, the neural network model recommends investigating 22.8% of the new cases for possible subrogation.

5.3 Text Categorization via Predictive Modeling

Objectives

- Review the topic of text categorization.
- Illustrate predictive modeling as a tool for text categorization using the Aviation Safety Reporting System (ASRS) data.

34

Text categorization was introduced in a previous chapter. The focus of this chapter is *supervised* text categorization.

Text Categorization

- Unsupervised
 - a special case of the pattern discovery classification problem
 - might be a first step for deriving labels that can then be used for supervised categorization
- Supervised
 - Class labels assigned by domain experts.
 - Supervised classification.
 - Classes can be mutually exclusive.
or
 - One document can belong to two or more classes.
 - Models are trained to automatically assign class labels to new documents.

35

Improving Text Categorization Accuracy

- Refine the dictionary/vocabulary.
 - Revise the start/stop list.
 - Add synonyms.
 - Add entities.
- Use different frequency weights, term weights, or both.
- Add SVD dimensions.
 - Add raw dimensions in the Text Cluster node.
 - Add topics in the Text Topic node.
- Improve SVD weights.
 - Customize term weights in the Text Topic node.

36

continued...

There are many ways to try to improve accuracy. When the improvements are negligible, it might be time to quit trying and settle on a final model.

Improving Text Categorization Accuracy

- Improve the predictive model.
 - Try several predictive models or algorithms.
 - Try different options for each predictive model or algorithm.
- Add inputs that are linked to the documents.
 - Origination codes (for example, physician, adjuster, agent)
 - Time stamps
 - Ontologies

37

Categorizing ASRS Reports

Analysis goal:

A government organization seeks to develop computer software to automatically categorize incoming safety/hazard reports with respect to 22 pre-determined safety/hazard categories.

Analysis data:

- Safety reports are extracted from the ASRS.
- Reports are pre-processed to promote uniform use of aviation terms and to remove personal references.
- Class labels have been assigned for 22 aviation safety/hazard categories.



Text Categorization of the ASRS Data

This demonstration illustrates how to use text mining and predictive modeling techniques to categorize aviation safety reports.

The Aviation Safety Reporting System (ASRS) was introduced in Chapter 2. For convenience, details of the ASRS are reviewed below.



The ASRS can be accessed from the following link:

asrs.arc.nasa.gov/

From the Web site:

“ASRS captures confidential reports, analyzes the resulting aviation safety data, and disseminates vital information to the aviation community.”

“More than 850,000 reports have been submitted (through October, 2009) and no reporter’s identity has ever been breached by the ASRS. ASRS de-identifies reports before entering them into the incident database. All personal and organizational names are removed. Dates, times, and related information, which could be used to infer an identity, are either generalized or eliminated.”

As with other data sets used in this course, data sets derived from ASRS have been modified. The original data for this demonstration was extracted from the ASRS, pre-processed, and provided to competitors in a text mining competition sponsored by SIAM and the NASA Ames Research Center. The competition results were presented at the Seventh SIAM International Conference on Data Mining held in 2007 in Minneapolis, Minnesota. Participants were prohibited from using the R language, SAS software, and most commercial software. A link that provides access to the original data follows.

c3.ndc.nasa.gov/dashlink/resources/138/

A single report in the ASRS database might be a composite derivation of two or more reports filed for the same incident. For example, one runway incursion incident can result in three reports: one from the pilot, one from the copilot, and one from an air traffic controller. An incident involving two or more aircraft can have reports filed from pilots of all aircraft involved as well as from air traffic controllers. In both examples, there will only be one ASRS report, but that report will be prepared by NASA professionals based on all reports submitted.

Reports can be submitted by an aviation professional, such as pilots, flight attendants, and mechanics. They can also be submitted by non-professionals, such as private pilots.

A report in the ASRS database has many fields, with one field representing a primary narrative describing the incident. This primary narrative is stored in the **Text** variable. All of the other fields have been omitted to simplify the text mining component of the analysis. In practice, an automated labeling system would attempt to use all fields.

NASA manually assigns to each report one or more of 54 anomalies, one or more of 32 results, one or more of 16 contributing factors, and one or more of 17 primary problems. For example, the report might describe an event that was a “runway ground incursion” anomaly, with a “took evasive action” result, that was a “human factor” contributing factor, and a “human factor” primary problem. These fields are not available in the contest data. Instead, the contest data has 22 labels, with a value of 1 “if document i has label j.” Otherwise the label has value -1. Labels correspond to the topics identified by NASA to aid in the analysis of the reports. The labels are not defined in the competition. For the course data, the 22 labels are named **Target01**, **Target02**, up through **Target22**. An original coding of (-1,1) has been changed to (0,1), with a code of 1 indicating the presence of the label in the document. A document can be associated with one or more labels.

There are two data sources for the ASRS: ASRS Training Data and ASRS Test Data. The training data contains columns indicating which of the 22 manually assigned labels relates to a given report. The test data represents new reports that are not classified. However, for evaluation purposes, classification of the 22 labels has been added. In the competition data, the labels were originally masked in the test data. The goal is to develop a system to automatically detect topics to avoid the time, cost, and error associated with manually labeling the reports.

A preliminary analysis shows that categorizing a report as exhibiting the **Target01** topic is difficult. The following table shows results for three predictive models using entropy term weights. A cutoff equal to the percentage of documents having Target01=1 is used, which is 6.7%.

Target01					
Data	Text Mining	Predictive Model	Precision	Recall	Misclassification
Original	Entropy	Regression	36.98	91.11	10.99
Original	Entropy	Decision Tree	38.62	83.33	9.98
Original	Entropy	Neural Network	38.41	88.98	10.29

Results for **Target19** are more promising. A value of 31.45% of all documents exhibit Target19=1.

The table below summarizes results for Target19 using both entropy term weights and mutual information term weights.

Target19					
Data	Text Mining	Predictive Model	Precision	Recall	Misclassification
Original	Entropy	Regression	80.26	85.73	11.12
Original	Entropy	Decision Tree	80.35	81.30	12.13
Original	Entropy	Neural Network	81.41	86.62	10.43
Original	Mutual Info	Regression	74.60	86.50	13.51
Original	Mutual Info	Decision Tree	70.16	86.42	15.83
Original	Mutual Info	Neural Network	77.35	85.40	12.45

The data provided for the SIAM competition has the NASA edits, which produce keywords in all uppercase. Because the parsing engine uses rules to determine noun groups, proper names, and so on, using all uppercase letters thwarts the correct identification of term roles. The following table is produced after converting many of the uppercase terms to lowercase.

Target19—Edited Data					
Data	Text Mining	Predictive Model	Precision	Recall	Misclassification
Original	Entropy	Regression	80.52	85.58	11.05
Original	Entropy	Decision Tree	75.98	83.13	13.57
Original	Entropy	Neural Network	80.89	85.89	10.82
Original	Mutual Info	Regression	74.88	86.58	13.35
Original	Mutual Info	Decision Tree	73.20	83.17	14.86
Original	Mutual Info	Neural Network	77.72	85.93	12.17

The edited data produced slightly better results for the neural network model compared to the unedited data when mutual information was selected as the term weighting method. In general, editing the uppercase entries had very little impact on the analysis.

Using a cutoff of 31.45% produces good results for all three models. However, one strategy for picking a cutoff is to select a value that produces equal precision and recall statistics, which is called the *break-even point*. The following table shows results for deriving a break-even point for a regression model.

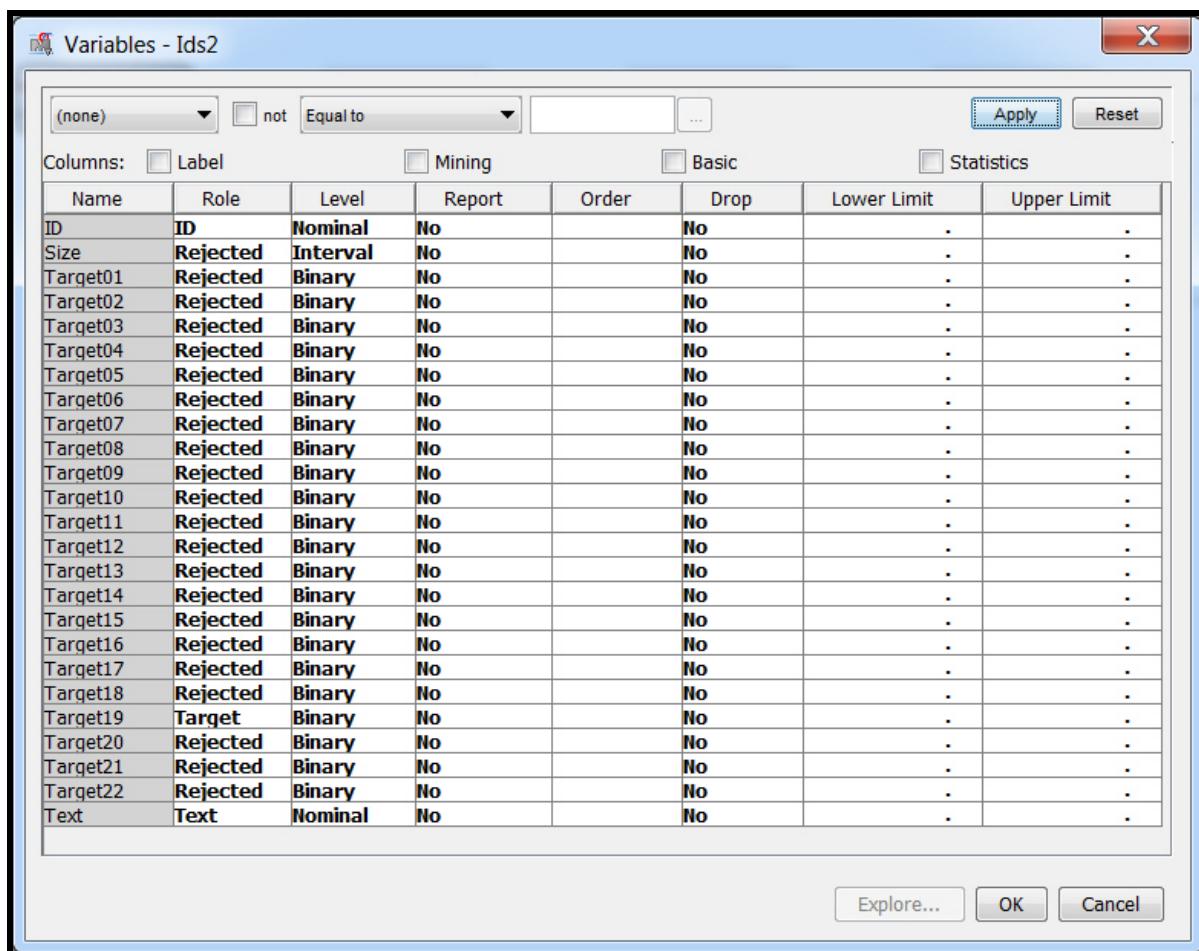
Break Even Precision/Recall	Break Even Percentile	Break Even Misclassification	Cutoff	Data Role
82.98	31.45	10.71	0.376	Train
83.27	31.46	10.53	0.420	Validate

Because the training data is used to derive the model and decision rule, a cutoff of 0.376 would be used to try to achieve the same precision and recall.

Predictive Modeling Results

To obtain the above results, an analysis was performed that included many parallel process branches in the process flow. The following summarizes an analysis that uses the default mutual information term weight and three predictive models. You should recall that the default term weight depends on whether a target variable is present in the training data. If there is no target variable, entropy is the default. If there is a target variable, mutual information is the default. Because target variable **Target19** is used for predictive modeling, mutual information term weights are used.

1. Use the input data source for the ASRS data created in Chapter 2 using the SAS data set **DMTXT.ASRSTRAIN**. Create the data source if necessary.
2. Create a diagram and name it ASRS Text Categorization PM. (PM=Predictive Modeling.)
3. Drag the ASRS data source into the diagram.
4. Select the input data source property **Variables** in the Columns subsection of the Train section. Assign variable roles as indicated in the following table:



5. Attach a Data Partition node to the Input Data Source node. Set the data set allocations as follows: Training 75%, Validation 25%, Test 0%. Run the Data Partition node.
6. Attach a Text Parsing node to the Data Partition node. Set the Synonyms property to **No data set to be specified**. Set the Stop List property in the Filter section to the data set **DMTXT.ASRS_STOP**. Run the Text Parsing node. (You can wait to run all of the SAS Text Miner nodes in sequence by running the last node in the sequence.)
7. Attach a Text Filter node to the Text Parsing node. Change the Term Weight property from Default to **Mutual Information**. Because a target variable is present, mutual information is the default term weight. However, in previous releases of SAS Text Miner, the default term weight was entropy even if a target variable was present. Setting the term weight as indicated will produce the appropriate results no matter which version you are using. Run the Text Filter node.
8. Attach a Text Cluster node to the Text Filter node. Change the Exact or Maximum Number property to **Exact**. Change the Number of Clusters property to **10**. Run the Text Cluster node.
9. Attach a Text Topic node to the Text Cluster node. Set the Number of Multi-term Topics property to **50**. Run the Text Topic node.
10. Attach a Metadata node to the Text Topic node. TextTopicM_N variables (for example, **TextTopic2_3**) are assigned a role of Segment by the Text Topic node. If you want to use these binary variables as inputs to a predictive model, then you must change the role to Input. The **TextTopicM_rawN** variables contain the actual weights used to derive the binary segment variables, and hence have a default role of Input. You can use the Metadata node to manipulate variable roles. For illustration, the first 10 TextTopic binary variables are given a role of Input.
11. Attach a Decision Tree node to the Metadata node. Change the Assessment Measure property in the Subtree section to **Average Square Error**. Run the Decision Tree node. Examine the results. The Fit Statistics table appears below.

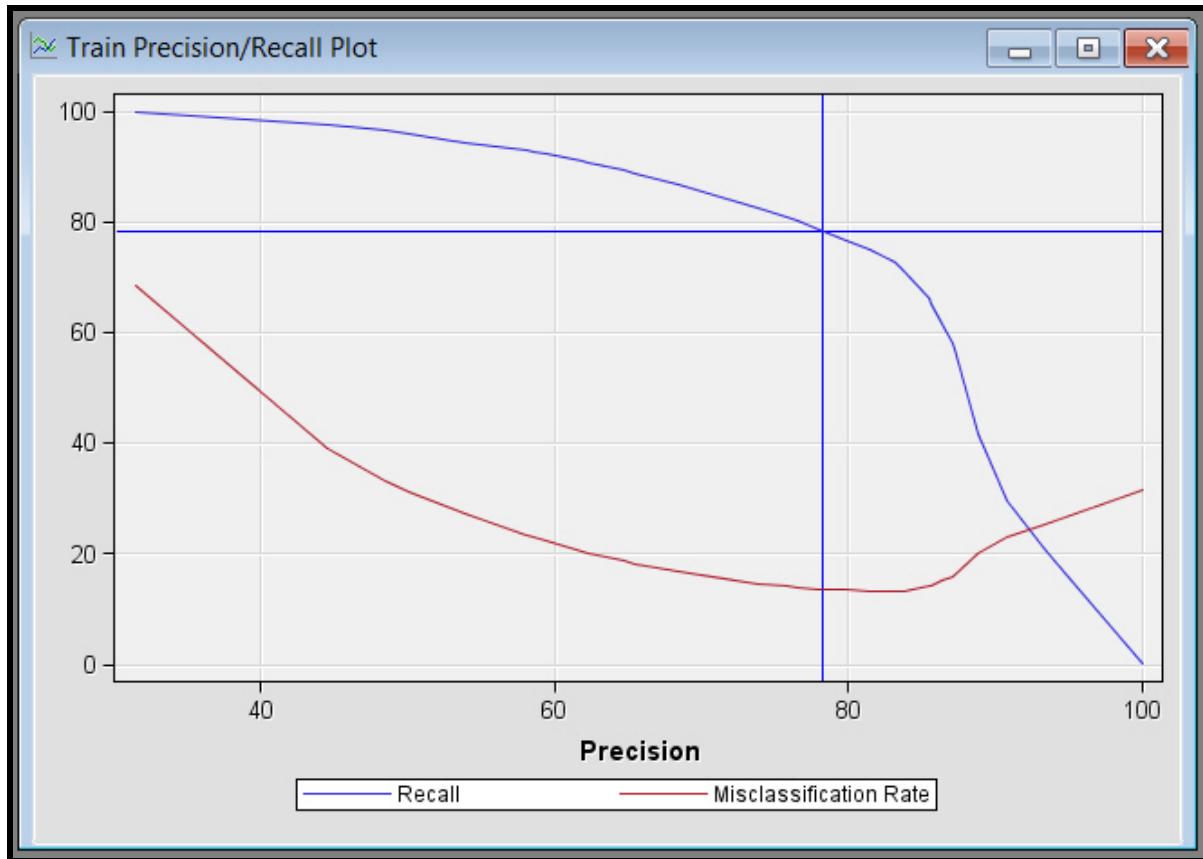
Fit Statistics

Target	Fit Statistics	Statistics Label	Train	Validation	Test
Target19	_NOBS_	Sum of Frequencies	16139	5380	
Target19	_MISC_	Misclassification Rate	0.13204	0.150743	
Target19	_MAX_	Maximum Absolute Error	0.975649	0.975649	
Target19	_SSE_	Sum of Squared Errors	3250.704	1213.841	
Target19	_ASE_	Average Squared Error	0.10071	0.112811	
Target19	_RASE_	Root Average Squared Error	0.317348	0.335873	
Target19	_DIV_	Divisor for ASE	32278	10760	
Target19	_DFT_	Total Degrees of Freedom	16139		

The Variable Importance table appears below.

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
TextCluster3_SVD3		5	1	1	1
TextTopic3_raw45	circuitbreaker,panel,check,maintain,reset	3	0.383128	0.375102	0.979051
TextCluster3_SVD9		3	0.218034	0.244681	1.122213
TextTopic3_raw16	gear,gear,land gear,retract,land	1	0.165196	0.156067	0.944738
TextTopic3_raw2	taxiway,runway,+hold,taxi,hold	2	0.151631	0.122541	0.80815
TextTopic3_raw8	flightattendant,passenger,number,cabin,cockpit	3	0.145931	0.128705	0.881956
TextTopic3_raw6	maintain,+find,+foot,zzz,traffic	1	0.141256	0.142067	1.005741
TextTopic3_raw17	cabin,pressurize,oxygenmask,emergency,cabi...	1	0.114895	0.083167	0.723848
TextTopic3_8	_1_0_flightattendant,passenger,number,cabin,...	1	0.111307	0.06302	0.56618
TextTopic3_raw25	system,maintain,problem,normal,operate	1	0.095324	0	0
TextTopic3_raw10	zzz,xxx,+find,restriction,restrict	1	0.088511	0.075132	0.848835
TextTopic3_raw13	airspace,classb,mile,operate,+foot	1	0.086125	0.07404	0.85968
TextTopic3_raw35	fail,problem,number,cause,maintain	2	0.084521	0.087558	1.035928
TextTopic3_raw5	traffic,trafficalertandcollisionavoidancesystem,...	1	0.080136	0.019854	0.247751
TextCluster3_SVD11		1	0.075872	0.030255	0.39876
TextTopic3_raw28	turn,head,degree,turn,degree head	1	0.075569	0.030411	0.402822
TextCluster3_SVD32		2	0.073659	0.032063	0.435291
TextCluster3_SVD5		1	0.06286	0.039656	0.63087
TextCluster3_SVD23		1	0.061739	0.027192	0.440441
TextCluster3_SVD15		1	0.057736	0.026064	0.451431
TextCluster3_SVD26		1	0.053677	0.049551	0.923141
TextTopic3_raw12	smoke,cabin,+evacuate,smell,fire	1	0.052045	0.038863	0.746723
TextTopic3_raw48	indicate,check,indication,fire,panel	1	0.051173	0	0
TextTopic3_raw31	temperature,engine,exhaustgas,maintain,exha...	1	0.050522	0.045803	0.906589
TextTopic3_raw50	xyz,xxx,abc,maintain,gate	1	0.044212	0.039087	0.884083
TextTopic3_raw43	perform,check,power,pound,operate	1	0.037364	0.032949	0.881838

12. Attach a SAS Code node to the Text Topic node. Select the **Code Editor** property. Position the cursor inside the Training Code panel (white area), right-click, and select **Open**. Navigate to the course SAS source folder and select the program **SCN_PrecisionAndRecall.sas**. Save the program and exit the code editor. Run the SAS Code node. View the results. Following is the precision/recall plot for the training data.



The break-even table follows.

Break Even Precision/Recall	Break Even Percentile	Break Even Misclassification	Cutoff	Data Roll
78.24199	30.87902	13.68759	0.416521	Train
76.04393	32.0504	15.06981	0.395238	Validate

For a cutoff of 0.4165, the tree achieves precision and recall values of approximately 78.24%. The misclassification rate is 13.69%. These results are very good and could provide a solution for automatically categorizing reports into the **Target19** category. However, given that sufficient time is provided for in the project, you always strive to improve results. As indicated above, there are many strategies for improving prediction accuracy. One of the easiest strategies when using SAS Enterprise Miner is to add different predictive modeling nodes to the process flow. A Regression node and a Neural Network node will be added to illustrate this strategy.

13. Attach a Regression node to the Metadata node. Change the Selection Model property to **Stepwise**, and change the Selection Criterion property to **Validation Error**. The Validation Error criterion minimizes the error defined to be the negative of the log-likelihood evaluated using the maximum likelihood estimates (derived for the training data) applied to the validation data. Run the Regression node. Examine the Regression node results. The Fit Statistics table follows.

Target	Fit Statistics	Statistics Label	Train	Validation	Test
Target19	AIC	Akaike's Information Criterion	9736.769	.	.
Target19	ASE	Average Squared Error	0.087817	0.090212	.
Target19	AVERRR	Average Error Function	0.298493	0.302426	.
Target19	DFE	Degrees of Freedom for Error	16088	.	.
Target19	DFM	Model Degrees of Freedom	51	.	.
Target19	DFT	Total Degrees of Freedom	16139	.	.
Target19	DIV	Divisor for ASE	32278	10760	.
Target19	ERR	Error Function	9634.769	3254.108	.
Target19	FPE	Final Prediction Error	0.088374	.	.
Target19	MAX	Maximum Absolute Error	0.998624	0.995109	.
Target19	MSE	Mean Square Error	0.088095	0.090212	.
Target19	NOBS	Sum of Frequencies	16139	5380	.
Target19	NW	Number of Estimate Weights	51	.	.
Target19	RASE	Root Average Sum of Squares	0.296339	0.300353	.
Target19	RFPE	Root Final Prediction Error	0.297277	.	.
Target19	RMSE	Root Mean Squared Error	0.296808	0.300353	.
Target19	SBC	Schwarz's Bayesian Criterion	10128.91	.	.
Target19	SSE	Sum of Squared Errors	2834.55	970.6802	.
Target19	SUMW	Sum of Case Weights Times Freq	32278	10760	.
Target19	MISC	Misclassification Rate	0.114505	0.116914	.

Comparing ASE to the tree model, the regression value of 0.0902 is smaller (better) than the tree ASE value of 0.1128. Misclassification uses a somewhat arbitrary cutoff value, so you should not pick a model based on the overall misclassification rate given in the Fit Statistics table. The stepwise selection method produces output that exceeds the output line limit, resulting in the following note:

NOTE: File view has been truncated.
Refer to C:\EM_Projects\DMTXT51\Workspaces\EMWS1\Reg4\EMOUTPUT.out
on this server for entire file contents.

Examining the file **EMOUTPUT.out** reveals a summary of the model selected using the stepwise selection method.

```

Summary of Stepwise Selection

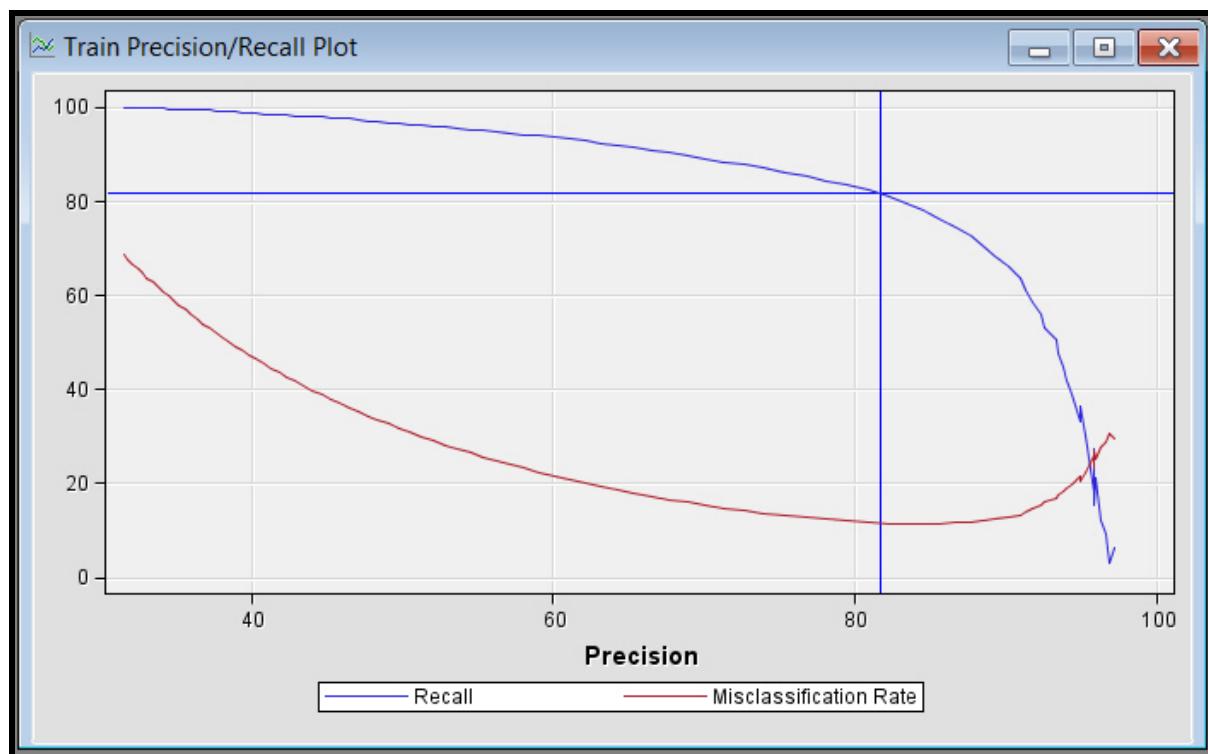
The selected model, based on the error rate for the validation data, is the model trained in Step
64. It consists of the following effects:

Intercept TextCluster3_SVD1 TextCluster3_SVD10 TextCluster3_SVD11 TextCluster3_SVD13
TextCluster3_SVD15 TextCluster3_SVD16 TextCluster3_SVD17 TextCluster3_SVD19 TextCluster3_SVD2
TextCluster3_SVD21 TextCluster3_SVD25 TextCluster3_SVD26
TextCluster3_SVD27 TextCluster3_SVD3 TextCluster3_SVD30 TextCluster3_SVD31 TextCluster3_SVD34
TextCluster3_SVD4 TextCluster3_SVD5 TextCluster3_SVD9 TextTopic3_10 TextTopic3_3
TextTopic3_raw11 TextTopic3_raw13 TextTopic3_raw14 TextTopic3_raw17
TextTopic3_raw18 TextTopic3_raw19 TextTopic3_raw2 TextTopic3_raw20 TextTopic3_raw23
TextTopic3_raw24 TextTopic3_raw25 TextTopic3_raw27 TextTopic3_raw28 TextTopic3_raw33
TextTopic3_raw39 TextTopic3_raw4 TextTopic3_raw40 TextTopic3_raw41
TextTopic3_raw42 TextTopic3_raw44 TextTopic3_raw45 TextTopic3_raw46 TextTopic3_raw47
TextTopic3_raw49 TextTopic3_raw50 TextTopic3_raw6 TextTopic3_raw8 TextTopic3_raw9

```

The model includes 50 variables, including SVD variables, rotated SVD variables, and binary topic variables.

14. Attach a SAS code node and open the **SCN_PrecisionAndRecall.sas** program as was done for the decision tree. The precision/recall plot for the training data appears below.



The break-even table follows.

Break Even Precision/Recall	Break Even Percentile	Break Even Misclassification	Cutoff	Data Roll
81.64807	31.45152	11.54604	0.444232	Train
81.25611	31.4559	11.79286	0.447408	Validate

The results appear to be superior to those obtained from the tree model.

15. Attach a decision tree node to the Metadata node. Change the Method option in the Subtree section to **Largest**. Change the Number of Surrogate Rules property in the Node section to **1**. You will be using the Decision Tree node as a variable selection node to choose inputs for the Neural Network node. Run the Decision Tree node. The Variable Importance table follows. (Note that the last two variables, **TextTopic3_raw41** and **TextTopic3_raw23**, are omitted because they exceeded the display capture size.) You can rename this node the Tree Variable Selection node.

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
TextCluster3_SVD3		6	1	1	1	1
TextTopic3_raw12	smoke,cabin,+evacuate,smell,...	2	2	0.871283	0.870582	0.999195
TextTopic3_raw45	circuitbreaker,panel,check,mai...	3	1	0.385042	0.374849	0.973528
TextTopic3_raw40	+replace,maintain,remove,act...	1	1	0.323061	0.31509	0.975327
TextCluster3_SVD9		3	1	0.224378	0.244516	1.08975
TextTopic3_raw1	engine,number,declare,emerg...	0	3	0.222739	0.231411	1.038937
TextCluster3_SVD10		0	3	0.20011	0.22132	1.10599
TextTopic3_raw8	flightattendant,passenger,num...	3	2	0.189977	0.148182	0.779998
TextTopic3_raw16	gear,gear,land gear,retract,land	3	0	0.186049	0.155962	0.838285
TextTopic3_raw2	taxiway,runway,+hold taxi,hold	2	3	0.183971	0.124707	0.677859
TextTopic3_raw21	flap,retract,maintain,flap,degree	0	2	0.1646	0.144004	0.874874
TextTopic3_raw25	system,maintain,problem,nor...	2	3	0.147	0.096952	0.659536
TextTopic3_raw6	maintain,+find,+foot,zzz,traffic	1	0	0.140874	0.141971	1.007792
TextCluster3_SVD6		0	3	0.139585	0.122971	0.880971
TextTopic3_6	_1_0_maintain,+find,+foot,zzz...	0	1	0.13159	0.132616	1.007792
TextTopic3_raw17	cabin,pressurize,oxygenmask,...	2	0	0.130574	0.083111	0.6365
TextCluster3_SVD11		2	2	0.127572	0.080132	0.628129
TextTopic3_raw11	fuel,tank,pound,engine,gallon	0	2	0.124958	0.103338	0.826977
TextTopic3_raw7	runway,approach,land,tower,...	0	1	0.123429	0.109939	0.890706
TextTopic3_8	_1_0_flightattendant,passenge...	1	0	0.111006	0.062977	0.567334
TextTopic3_raw20	+foot,+cross,descend,altitude,...	0	2	0.108062	0.035877	0.332007
TextTopic3_raw9	install,+replace,inspect,mainta...	0	1	0.107161	0.078724	0.734632
TextCluster3_SVD5		1	1	0.100098	0.049009	0.489608
TextTopic3_raw35	fail,problem,number,cause,ma...	2	1	0.093531	0.087499	0.935507
TextTopic3_raw10	zzz,xxx,+find,restriction,restrict	0	0	0.088272	0.075081	0.850566
TextTopic3_raw13	airspace,classb,mile,operate,...	1	0	0.085891	0.07399	0.861434
TextCluster3_SVD13		1	3	0.0845	0.043256	0.511901
TextTopic3_2	_1_0_taxiway,runway,+hold,ta...	1	1	0.083547	0.038126	0.456335
TextTopic3_raw5	traffic,trafficalertandcollisionav...	1	0	0.079919	0.01984	0.248257
TextTopic3_raw44	inoperative,equipmentlist,main...	1	2	0.077652	0.042211	0.543589
TextTopic3_raw24	direction,veryhighfrequencycom...	0	1	0.07699	0.066322	0.861434
TextTopic3_raw28	turn,head,degree,turn,degree ...	1	0	0.075364	0.03042	0.403643
TextCluster3_SVD32		2	0	0.07346	0.032042	0.436179
TextTopic3_raw27	problem,cause,troubleshoot,m...	0	2	0.071063	0.024375	0.343011
TextTopic3_raw32	manual,operate,maintain,man...	0	2	0.065353	0	0
TextCluster3_SVD7		0	1	0.065281	0.026085	0.399573
TextCluster3_SVD23		1	0	0.061572	0.027174	0.441339
TextCluster3_SVD26		2	0	0.060026	0.065856	1.097122
TextCluster3_SVD33		1	1	0.059847	0	0
TextTopic3_raw14	inspect,maintain,perform,+find...	1	0	0.058225	0	0
TextCluster3_SVD15		1	0	0.05758	0.026046	0.452351
TextTopic3_raw43	perform,check,power,pound,o...	1	1	0.0507086	0.032927	0.576793
TextTopic3_raw33	normal,normal,operate,indicati...	1	1	0.054461	0.029321	0.538381
TextCluster3_SVD14		0	1	0.053368	0	0
TextCluster3_SVD28		0	1	0.052621	0.048676	0.925024
TextTopic3_raw48	indicate,check,indication,fire,p...	1	0	0.051034	0	0
TextTopic3_raw31	temperature,engine,exhaustga...	1	0	0.050385	0.045772	0.908438
TextCluster3_SVD18		0	1	0.045949	0.022665	0.493263
TextCluster3_SVD27		1	0	0.044976	0	0
TextTopic3_raw50	xyz,xxx,abc,maintain,gate	1	0	0.044092	0.039061	0.885887
TextCluster3_SVD30		1	0	0.043637	0	0
TextCluster3_SVD16		0	1	0.037072	0.032758	0.883637

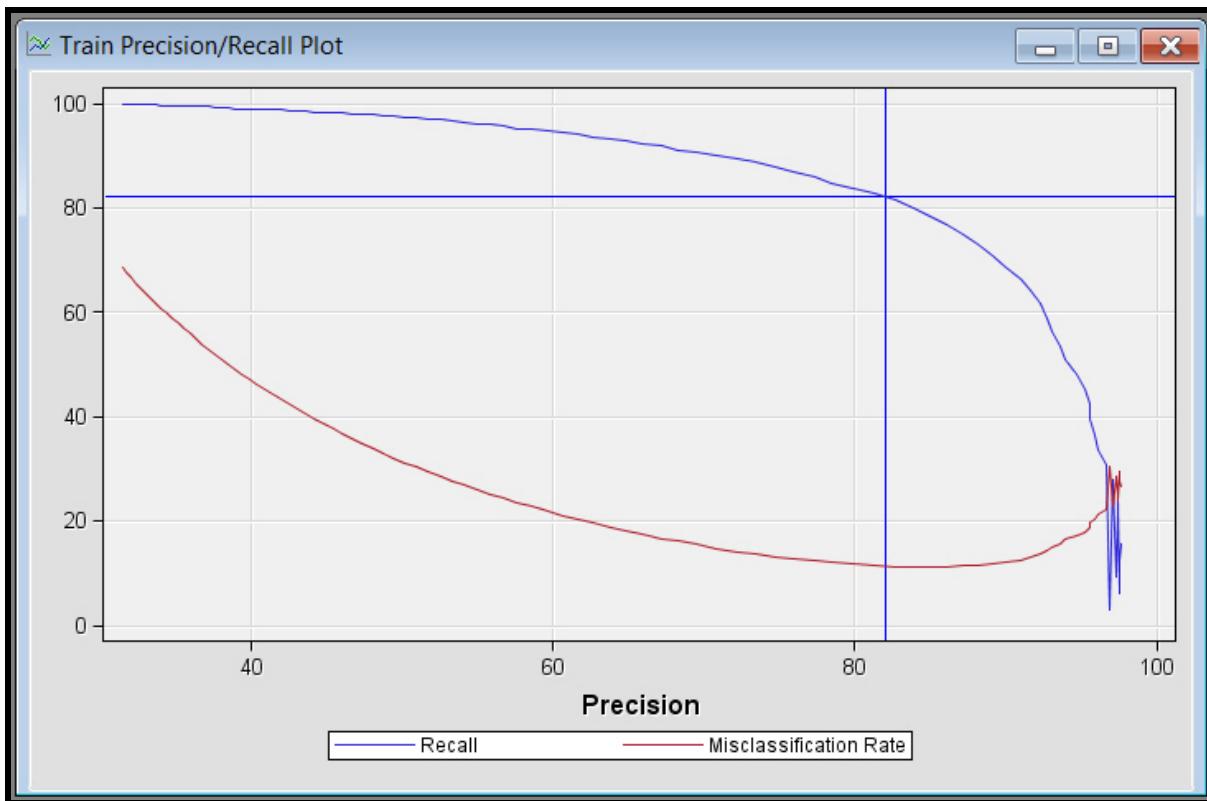
By choosing more liberal stopping and pruning rules, a leafier tree results in more variables being selected than in the tree that was previously selected.

16. Attach a Neural Network node to the Decision Tree node from the previous step. Run the Neural Network node. The Fit Statistics table follows.

Target	Fit Statistics	Statistics Label	Train	Validation	Test
Target19	_DFT_	Total Degrees of Freedom	16139		
Target19	_DFE_	Degrees of Freedom for Error	15985		
Target19	_DFM_	Model Degrees of Freedom	154		
Target19	_NW_	Number of Estimated Weights	154		
Target19	_AIC_	Akaike's Information Criterion	9423.025		
Target19	_SBC_	Schwarz's Bayesian Criterion	10607.13		
Target19	_ASE_	Average Squared Error	0.08377	0.096703	
Target19	_MAX_	Maximum Absolute Error	0.994556	0.993259	
Target19	_DIV_	Divisor for ASE	32278	10760	
Target19	_NOBS_	Sum of Frequencies	16139	5380	
Target19	_RASE_	Root Average Squared Error	0.28943	0.310971	
Target19	_SSE_	Sum of Squared Errors	2703.924	1040.523	
Target19	_SUMW_	Sum of Case Weights Times Freq	32278	10760	
Target19	_FPE_	Final Prediction Error	0.085384		
Target19	_MSE_	Mean Squared Error	0.084577	0.096703	
Target19	_RFPE_	Root Final Prediction Error	0.292205		
Target19	_RMSE_	Root Mean Squared Error	0.290821	0.310971	
Target19	_AVERR_	Average Error Function	0.282391	0.320264	
Target19	_ERR_	Error Function	9115.025	3446.043	
Target19	_MISC_	Misclassification Rate	0.111593	0.129554	
Target19	_WRONG_	Number of Wrong Classifications	1801	697	

The ASE value of 0.0967 is similar to the value obtained using a logistic regression model.

17. Attach a SAS Code node to the Neural Network node. Open the **SCN_PrecisionAndRecall.sas** program as before. Run the SAS Code node. The precision/recall plot for the training data appears below.



The break-even table follows.

Break Even Precision/Recall	Break Even Percentile	Break Even Misclassification	Cutoff	Data Roll
82.02042	31.45142	11.31176	0.398651	Train
79.38545	31.45524	12.96884	0.388647	Validate

The results are similar to those obtained using a logistic regression model.

Trying different models did not produce dramatic improvements over the original decision tree model. Precision and recall values of 80% are probably adequate for an automated text categorization system. Using a cutoff of 0.3987 for the derived neural network model should extract about 80% of all **Target19** documents.

5.02 Multiple Answer Poll

Select all of the choices that represent text mining derived values that can be used as inputs for a predictive model.

- a. Text Cluster node SVD variables
- b. Cluster ID values
- c. Text topic binary variables
- d. Text topic rotated SVD variables



Exercises

1. Predictive Modeling

- a. Continue with the Workers' Compensation Recovery Potential diagram, using the SAS data set **LWDMDXT.WORKCOMP**. Add a neural network model and imitate the variable selection strategy used in the ASRS Text Categorization PM diagram. That is, attach a Decision Tree node to the Metadata node and specify properties to make it a variable selection tree. Then attach a Neural Network node to the Decision Tree node and use average error as a model selection criterion. Attach this node to the Model Comparison node and run the node.
- b. What is the ASE for the new neural network model?
- c. Propose additional strategies for reducing the ASE statistic.

5.4 Chapter Summary

Text mining inputs can enhance the accuracy of predictive models. The SAS Text Miner nodes add four sources of input variables:

- Text Cluster node SVD variables
- Cluster ID as a nominal input variable
- Text Topic node binary variables as binary input variables
- Text Topic node raw SVD variables

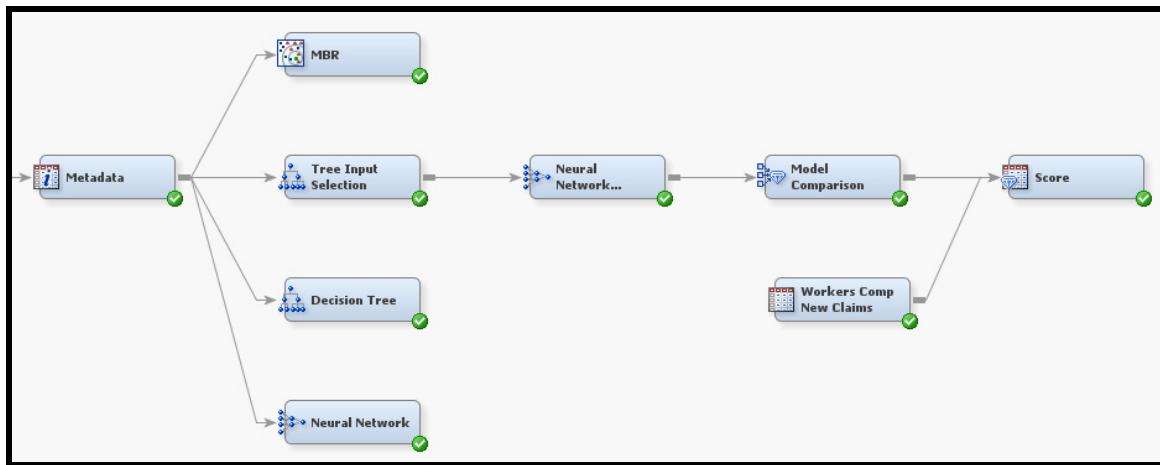
Text categorization is a predictive modeling problem that predicts categories that have been assigned to each document. While predictive modeling in general will use fit statistics like average square error and average profit, text categorization problems favor statistics such as precision and recall to evaluate the effectiveness of the model.

5.5 Solutions

Solutions to Exercises

1. Predictive Modeling

- a. The portion of the process flow should look like this:



- b. The ASE for the second neural network model is 0.1555, which is slightly worse than the neural network that uses all inputs.
- c. There are many predictive modeling nodes in SAS Enterprise Miner. You could try another model node and see whether accuracy can be improved. You could also manipulate text mining options, such as (1) adding synonyms, (2) adding entities, (3) modifying the stop list, (4) choosing different term or frequency weights, (5) modifying text topics to create custom topics, or (6) choosing more dimensions for Text Cluster node SVD columns or Text Topic node raw columns.

Solutions to Student Activities (Polls/Quizzes)

5.02 Multiple Answer Poll – Correct Answers

Select all of the choices that represent text mining derived values that can be used as inputs for a predictive model.

- a. Text Cluster node SVD variables
- b. Cluster ID values **(as a nominal input)**
- c. Text topic binary variables **(as binary inputs)**
- d. Text topic rotated SVD variables

Appendix A Index

%

%TMFILTER macro, 2-6

A

Advanced Advisor, 1-74
algorithms, 1-35, 3-32, 4-32
analysis element organization, 1-64
anomaly detection, 1-11
association discovery, 4-54
attributes, 1-106

B

Boolean search, 1-56–1-59, 2-24

C

centered terms, 3-11
chi-square weights, 3-21–3-22
clustering, 4-32
concept linked terms, 3-11
concept linking
 conditional counts, 3-11–3-12
cosine distance, 3-7–3-9
Custom Entities property, 1-84

D

Data Import wizard, 2-4
data mining, 1-18
 signal versus noise, 1-21–1-23
data sources
 creating, 1-72–1-75
 text mining, 1-69–1-72
Decision Tree node, 5-13
defining a data source, 1-65
diagram workspace
 SAS Enterprise Miner, 1-80
dictionaries, 1-7
dimensionality reduction, 3-32–3-33, 5-12
distance measures, 3-7–3-9
document categorization, 1-10
documents, 4-54
 categorizing, 4-25–4-52
 characteristics, 3-4
 comparing, 3-6–3-9

E

entities, 1-105
entropy weights, 3-16, 3-19, 3-22
Euclidean distance, 3-7–3-9

F

factor analysis, 3-28
filters, 1-14
forensic linguistics, 1-11, 2-14–2-17
frequency weights, 3-15

G

global weights, 3-19

H

Help panel
 SAS Enterprise Miner, 1-79

I

information retrieval (IR), 1-10, 2-21–2-24
 filtering and querying, 2-22–2-23
Interactive Filter Viewer, 2-22, 3-11
Interactive Topic Viewer, 1-32
inverse document frequency (IDF) weights,
 3-16, 3-19, 3-22

K

keywords, 1-31, 1-72, 3-15

L

language keywords, 1-72
latent semantic analysis (LSA), 1-56, 1-91,
 3-25
latent semantic indexing (LSI), 1-56, 1-91,
 2-24
linear algebra, 1-56–1-59, 3-12, 3-25
local weights, 3-19, See also frequency
 weights

M

market basket analysis, 4-54

- MBR node. See Memory-Based Reasoning node
- measurement levels
- interval, 1-67
 - nominal, 1-67
 - ordered categorical, 1-67
 - ordinal, 1-67
 - ratio, 1-67
 - unary, 1-67
- Memory-Based Reasoning node, 2-17, 2-20
- menu bar
- SAS Enterprise Miner, 1-78
- Model Comparison node, 5-15
- Model tab
- SAS Enterprise Miner, 5-4
- models
- assessing, 5-14–5-16
- N**
- neural network model, 1-16
- noun groups, 1-104
- O**
- outlier detection, 4-32
- P**
- parts of speech, 1-104
- pattern discovery, 1-6, 1-20
- text mining, 1-30–1-32
- Perl regular expressions, 2-4
- phi coefficient, 3-8
- predicting fraud, 4-25–4-32
- predictive modeling, 1-6, 5-17
- SAS Enterprise Miner, 5-4–5-16
 - text categorization, 5-25–5-39
 - text-based inputs, 5-17
- Princomp node, 2-20
- process flow
- SAS Enterprise Miner, 1-80
- Project panel
- SAS Enterprise Miner, 1-78
- Properties panel
- SAS Enterprise Miner, 1-79
- Q**
- query operators, 2-23
- R**
- Regression node, 5-13
- relative frequency, 3-4
- S**
- SAS
- data access features, 2-4
 - selected functionality, 2-5
- SAS Code node, 1-69, 2-6, 2-7
- SVD vectors, 3-30
- SAS Content Categorization, 1-84
- SAS Enterprise Guide, 2-4
- SAS Enterprise Miner
- analysis element organization, 1-64
 - defining a data source, 1-65
 - diagram workspace, 1-80
 - dimensionality reduction, 5-12
 - Help panel, 1-79
 - input selection, 5-12
 - menu bar, 1-78
 - Model tab, 5-4
 - predictions data, 5-8–5-9
 - predictive modeling, 5-4–5-16
 - process flow, 1-80
 - Program Editor, 1-69
 - Project panel, 1-78
 - Properties panel, 1-79
 - SEMMA tools palette, 1-82
 - source data, 5-7–5-8
- SAS macros, 2-7
- SAS programs
- running in SAS Enterprise Miner, 2-6
- SAS Sentiment Analysis, 3-4
- SAS Text Miner, 1-7
- association discovery, 4-54–4-57
 - attributes, 1-106
 - entities, 1-105
 - noun groups, 1-104
 - parts of speech, 1-104
 - predicting recovery potential, 5-20–5-24
 - text categorization, 4-25–4-52
 - transaction data set, 4-55
 - types of input variables, 5-11
 - untransformed view, 3-15
 - warranty analysis, 4-3–4-24
- SAS/ACCESS, 2-4
- Score node, 1-68
- SEMMA tools palette
- SAS Enterprise Miner, 1-82
- sentiment analysis, 1-5
- separators, 3-3
- sequence discovery, 4-54

singular value decomposition (SVD), 1-91, 3-25–3-33
 dimensionality reduction, 3-32–3-33
 start list, 1-7
 stylometry, 1-10, 1-33–1-37, 2-13
 forensic linguistics, 2-14
 support vector machine (SVM), 1-35

T

table roles, 1-68
 term weights
 guidelines, 3-21–3-22
 simulation study, 3-22
 terms, 3-3, 4-54
 centered, 3-11
 comparing, 3-10
 concept linked, 3-11
 quantifying, 3-14
 text
 categorizing, 4-25–4-52
 text analytics, 1-3–1-4, 1-6
 association discovery, 4-54
 market basket analysis, 4-54
 sequence discovery, 4-54
 warranty analysis, 4-3–4-24
 text categorization, 2-34–2-38
 predicting fraud, 4-25–4-32
 predictive modeling, 5-25–5-39
 supervised, 2-38
 unsupervised, 2-38
 Text Cluster node, 1-90–1-91
 text extraction
 types, 3-4

Text Filter node, 1-14, 1-87–1-88
 Interactive Filter Viewer, 3-11
 query operators, 2-23
 Text Import node, 2-7–2-8, 2-10–2-11
 Text Miner node, 1-91
 SVD vectors, 3-30
 text mining, 1-6–1-11, 1-60–1-61
 assessing results, 1-27–1-28
 data sources, 1-69–1-72
 pattern discovery, 1-30–1-32
 signal versus noise, 1-25–1-27
 text categorization, 4-25–4-52
 warranty analysis, 4-3–4-24
 Text Parsing node, 1-83–1-86
 Custom Entities property, 1-84
 Text Topic node, 1-91–1-93, 2-35, 3-28
 distance measures, 3-9
 SVD columns, 3-29
 tokens, 3-3

V

variable roles, 1-66, 1-68
 vector space model (VSM), 1-56, 1-91
 vocabulary, 1-7

W

warranty analysis, 4-3–4-24
 weighted term-document frequency matrix, 3-16, 3-20

Z

Zipf's Law, 1-7, 3-5–3-6

