Technical report

Mastering New Challenges in Text Analytics

Making unstructured data ready for predictive analytics

Table of contents

Introduction

It's no secret that the world has seen an explosion of information in the past 15 years, an explosion that experts predict will continue as the millions of people who use online resources continue to expand their usage, and the millions of people who do not yet have access to such resources gain it. Similarly, information stored as text in both business and government organizations has grown exponentially.

To name just a few examples:

- Opinion surveys are increasingly conducted online and results shared in real time
- The boom in software applications supporting sales, customer service, or call center operations has led to massive amounts of text stored electronically in these applications' notes fields
- Technology analysts at IDC estimate that 62 billion e-mails are sent every day
- Searchable Web sites generate enough information every day to fill millions of books
- Web logs (blogs) and wikis, created by individuals and groups for personal and professional purposes are increasing exponentially: as of this writing, there may be more than 100 million blogs, with a new one created every second

Such a vast expansion of the scale of global information exchange would have been almost unimaginable 40 years ago, when most business and government communications, as well as news reports and advertising, were paper-based.

Yet it was 40 years ago that visionary researchers began to seek ways to enrich the knowledge of those working in medicine and other sciences, in government agencies, and in business by making it possible to uncover previously unknown connections in large collections of textual documents by using computer technologies. They created the discipline known as *computational linguistics*, which is now practiced at numerous universities and public and private research centers worldwide.

Computational linguists initially focused their efforts on finding ways to categorize and explore concepts found in books, scholarly journals, legal briefs, patent applications, newspapers, reports, and other paper-based records that could be converted to digital formats. More recently, their efforts have expanded to include ways to "mine" the vast amount of textual information that is published digitally—online editions of newspapers, academic journals, and conference proceeding, for example. In addition, there is a wealth of content that originates in digital form—such as Web sites, *blogs, wikis*, e-mails, instant messaging (IM), as well as text embedded in forms, surveys, and in scientific, government, or corporate databases.

There is a growing recognition that analyzing text has become essential in various types of scientific research, and that it adds significant value to other forms of data analysis, particularly when used to predict how people may act in certain situations. For example, in obtaining a well–rounded view of customer behavior, text analytics is critical because it provides insight into the nuances of attitudes and opinions that influence behavior. With the exponential growth of text in online formats, ways must be found to structure this information and make it available to researchers and decision makers.

This paper briefly defines text analytics, describes various approaches to text analytics, and then focuses on the natural language processing techniques used by SPSS Inc.'s text analytics solutions. It concludes with descriptions of SPSS solutions for text analytics and their role in predictive analytics.

What is text analytics and how is it used?

First, it may be helpful to clarify what we mean by the terms text analytics and predictive analytics.

To clear up one misconception, text analytics is not the same as search. Search engines are a "top down" approach to finding information in textual material. This means that end users must know how to structure queries to arrive at exactly the desired information. Text analytics, by contrast, is a "bottom up" approach. It does not require users to know particular search terms. Instead, text analytics reveals the concepts and themes contained in a body of documents, and then maps the relationships between them.

To provide a more formal definition: Text analytics is a method for extracting usable knowledge from unstructured text data through identification of core concepts, sentiments, and trends, and then using this knowledge to support decision making. A "document" might be a scholarly journal article, free text responses to a market research survey, records from a database—such as call center notes or customer e-mails—contents of a news feed, or even a crime scene report.

Text analytics discovers connections and relationships not within a single document but across a large collection or "corpus" of documents. These connections and relationships can then be organized in ways that permit analysis either alone or in combination with other types of data. Practitioners of text analytics may use algorithms to describe clusters of concepts, or associations between certain concepts or named entities. Text analytics results can then be incorporated in models used for predictive analytics.

Predictive analytics informs and directs decision making by applying a combination of advanced analytics and decision optimization to data, with the objective of improving business processes to meet specific organizational goals. Including textual or "unstructured" data along with the "structured" data found in databases or transaction records adds depth to the insights gained through data mining. Textual data often reveals attitudes and sentiments that, when combined with demographic or behavioral data, enable analysts to more reliably predict events, behaviors, or actions that individuals or groups are likely to engage in.

Text analytics has been shown to deliver measurable benefits to organizations in a wide range of applications. For commercial organizations, these include:

- Supporting improved customer relationship management (CRM) by providing a more well-rounded view of customers, their wishes and preferences, leading to more effective marketing, reduced churn, and improved customer loyalty and lifetime value
- Catching the "voice of the customer" through surveys or data from Web 2.0 interactions to improve customer loyalty and brand monitoring
- Accelerating cycle times in the development and refinement of products, and early detection of product issues through warranty analysis
- Achieving a clearer view of the competitive landscape

Text analytics also has applications in the public sector; for instance in:

- Uncovering patterns that suggest fraudulent behavior may be occurring
- Detecting connections among groups of criminals
- Identifying possible security threats or illegal activity

In addition, text analytics can be invaluable in scientific and medical research; for example by:

- Speeding the exploration of secondary research materials, such as patent reports and journal articles
- Identifying previously unknown associations among people, research projects, or products
- Minimizing the time spent in the drug discovery process

These are just some examples of how text analytics is being used, and how it can enhance predictive analytics. More applications are being implemented every day. Organizations simply cannot afford to ignore this wealth of textual information.

Approaches to understanding text

There are several approaches that an organization might take when performing text analytics. In the past, the tradeoff has been between accuracy and speed; between the cost of human labor and the cost of computer technologies. Today, organizations are reaping the benefits of increased accuracy and reduced cost in applying computer technologies to text analytics; but there will always be a need to incorporate human knowledge into the process.

One approach to understanding text is simply having people read the documents, note their contents, and determine into which categories they should be placed. Market researchers, for example, often categorize or "code" free-text responses in surveys. Because people are good at understanding text, this approach is quite accurate; but it is time-consuming and expensive. In addition, a manual approach cannot offer guidance in identifying relationships or trends in the information analyzed. With the immense volume of text now available, often in multiple languages, other approaches are needed.

A second approach is to employ automated solutions based on statistics. Some of these, however, simply count the number of times terms occur and calculate their proximity to related terms. Because they cannot factor in the ambiguities in human languages, relevant relationships may be hidden in masses of irrelevant findings—or missed altogether. Some of these statistics-based solutions compensate by providing ways for analysts to create rule books that help suppress irrelevant results. But these rulebooks need to be created and continually updated by analysts, which adds cost and complexity.

Other statistics-based solutions rely on self-learning tools such as Bayesian networks, neural networks, support vector machines (SVM), and/or latent semantic analysis (LSA). While these solutions can be more effective than other statistical approaches, they have the drawback of being "black boxes"—that is, using hidden mechanisms that cannot be adjusted except by highly skilled statisticians or programmers.

Linguistics-based text analytics offers the speed and cost effectiveness of statistics-based systems, but it offers a far higher degree of accuracy. Linguistics-based text analytics is based on the field of study known as *natural language processing* (NLP). (For a glossary of selected text analytics terms, see Appendix A.) The understanding of language that is possible with the NLP approach cuts through the ambiguity of text, making linguistics-based text analytics the most accurate possible approach.

Initially, linguistics-based solutions may require some human intervention—in developing dictionaries for a particular industry or field of study, for example. But the benefit obtained from these efforts is significant: results are more accurate and the techniques involved are more transparent, meaning that they can be modified by users to further increase the accuracy of results.

The SPSS text analytics process

Like data mining, text analytics is an iterative process, and is most effective when it follows a proven methodology. This maximizes analyst productivity, supports comparability of results, allows findings from one analysis to be used to inform or guide others, and facilitates data-driven decision making.

In data mining, the industry-standard methodology—used by thousands of organizations worldwide—is the CRoss-Industry Standard Practice for Data Mining (CRISP-DM). This same methodology supports text analytics.

This paper describes the linguistic processes involved in text analytics, which follow the broad outlines of the CRISP-DM methodology in that once data is understood, prepared, and modeled, the resulting models are evaluated—whether they involve only text analytics results or are combined with other types of data. Finally, results are deployed, either as reports or as scores driving automated systems such as recommendation engines. As with data mining, the two main steps in text analytics are data preparation and data understanding.

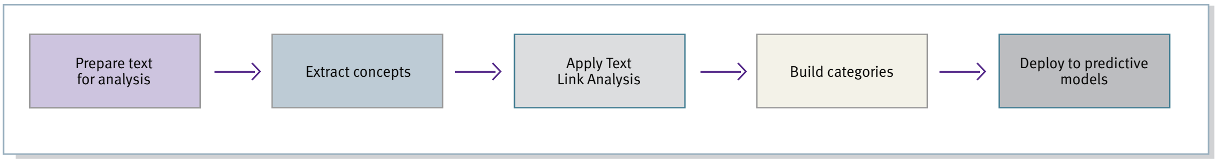
The next sections describe how analysts would use SPSS' text analytics products to engage in text analytics.

There are seven major steps in the text analytics process:

1. Preparing text for analysis
2. Extracting *concepts*
3. Uncovering opinions, relationships, facts, and events through Text Link Analysis
4. Building *categories*
5. Building text analytics models
6. Merging text analytics models with other data models
7. Deploying results to predictive models

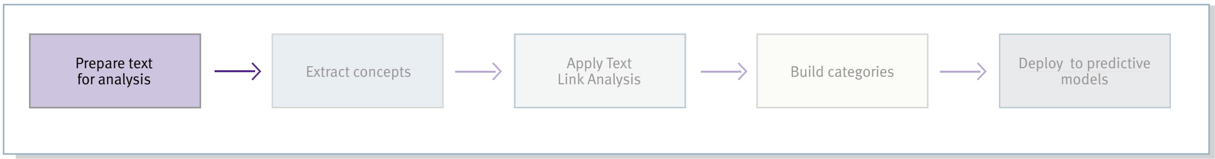
Because this paper focuses on the linguistic capabilities built into SPSS' text analytics products, it will cover the first four steps in this process, with some discussion of deployment to predictive models.

Workflow



The workflow is similar, whether the goal is to analyze journal articles, internal documents, Web pages, verbatim responses to surveys, call center notes, or other sources of text data.

Workflow



Preparing text for analysis

In order to conduct text analytics, a body of documents, or “corpus,” is needed. A corpus can range from a small sample to tens of millions of documents. The documents may be written in multiple languages and represent a variety of file types: HTML, PDF, ASCII, e-mail, and common Microsoft® Office formats.

SPSS text analytics solutions can process text in all these formats. In addition, they can process survey text saved in SPSS’ Dimensions™ format, as well as text from RSS feeds (including blogs and news feeds), databases, and other ODBC-compliant sources.

SPSS text analytics solutions use powerful, linguistics-based capabilities to prepare text documents for analysis. The three steps in the preparation of documents are:

- Language identification
- Document conversion
- Segmentation

Although these steps take place “behind the scenes,” it is valuable to understand what occurs during this phase of the text analytics process.

Language identification

For corpora that use multiple languages, language identification is the first step in the extraction process. (For single-language corpora, this step is not necessary.)

The SPSS text analytics extractor can recognize more than 80 languages in different formats, based on patterns known as “n-grams” that are specific to each language. About 400 n-grams are used to identify each language. Below is a subset of tri-grams used for recognizing French (some are combinations of letters, others are combinations of letters and spaces):

“ le,” “omm,” “ à,” “mma,” “le ,” “du ,” “nt ,” “ma ,” “ et,” “té ,” “ dé,” “les,” “ur ,” “ux ,” “une,” “ ré,” “iod,” “pou,” “rp,” “ui ,” “ait,” “rpa,” “pré,” “ ce,” “ité,” “ire,” “ée ”, “com,” “par,” “ef ,” “od ,” “au ,” “iqu,” “ref,” “ ét,” “oit,” “lpa,” “our,” “tio,” “air,” “eur,” “ du,” “és” “.av,” “ns ,” “tai”

SPSS text analytics solutions are available for seven native language extractors: English, French, Spanish, Dutch, German, Italian, and Portuguese. (SPSS text analytics products also support the extraction of Japanese concepts; Japanese extraction uses a different process not described in this document.)

Additionally, through the use of Language Weaver software, the English language extractor supports translations from the following 14 languages: Arabic, Chinese, Dutch, French, German, Hindi, Italian, Persian, Portuguese, Romanian, Russian, Somali, Spanish, Swedish. Language Weaver continues to add new translation capabilities, which SPSS products will continue to support.

Document conversion

Once the language has been identified, the SPSS text analytics solution converts documents to a format that can be used for further analysis. Using built-in filters, the software converts common file types to a plain text format. Text from databases and other ODBC-compliant sources can also be converted. For example, in an XML-based document, the tags can be used to specify the text that is to be extracted, including page titles, metadata, and document tags, if desired. The text analytics solution also removes non-textual elements such as graphic files, which are unusable for text analytics.

Segmentation

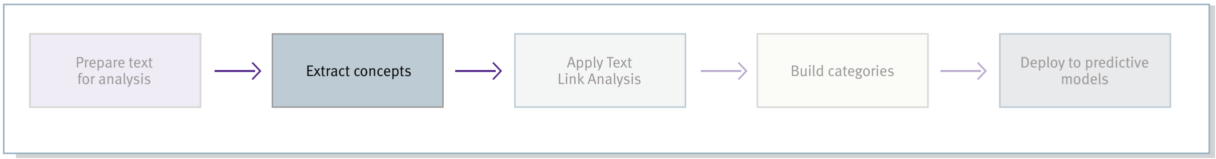
After the documents are converted to plain text, the text analytics solution segments the text into individual elements from which concepts will be extracted. SPSS text analytics software identifies markers for the ends of sentences, paragraphs, and documents. It also removes certain special characters or character sequences or replaces them with spaces.

During this step, the software automatically corrects or prepares text so that it's optimal for mining. For example, the software identifies character strings from the input text, based on delimiters. Delimiters include spaces, tabs, carriage returns, and punctuation marks. In SPSS text analytics technologies, any word that contains a punctuation mark that is not preceded or followed by a space, will, in the next steps of the process, be treated as part of a term. For example:

- U.S.
- xalpha(s) protein
- x.k-atpase beta-m subunit

SPSS text analytics solutions can also accommodate poor punctuation in the text, such as improper use of periods, commas, forward-slashes, and other forms of punctuation.

Workflow



Extracting concepts

The processes involved in concept extraction enable analysts to discover concepts they would not necessarily have known existed in a particular set of documents, and to find instances of these concepts wherever they occur, even in a vast set of text documents.

The five major steps in the concept extraction process are:

- Managing linguistic resources
- Term extraction
- Type assignment
- Creation of *equivalence classes*
- *Indexing*

Managing linguistic resources

Although in many situations it is not necessary to modify linguistic resources shipped with the SPSS text analytics solution, it is good to understand what resources are available.

Linguistic resources are arranged in a hierarchy. At the highest level, there are specialized resource templates, each of which comprises a set of libraries, compiled resources, and some advanced resources. Libraries, in turn, contain several types of dictionaries.

For all supported languages, a Basic Resources template is included. In addition, for English, there are specialized templates for a variety of specific application areas, such as CRM, market intelligence, gene ontology, genomics, Medical Subject Headings or MeSH®, IT, opinions, and security intelligence. Specialized templates are available for several other languages as well.

Each template may contain several libraries. The budget library, for example, is used to extract terms referring to the cost of something. The opinions library contains thousands of words that represent attitudes, qualifiers, or preferences that indicate an opinion about something. It is available for English, French, Spanish, Dutch, German, and Japanese. A core library is available in all languages.

Each library contains several dictionaries, which are lists of words, relationships, or other information used to specify or tune the extraction. There are two kinds of dictionaries in the SPSS text analytics solutions: compiled dictionaries, which end users cannot modify, and other dictionaries, which they can.

The SPSS text analytics solution includes two types of compiled dictionaries:

- An extraction dictionary—a list of base forms with part-of-speech (PoS) codes—for each language. The parts of speech specified in the extraction dictionary for English, for example, include *noun, verb, adjective, adverb, participle, coordination, determiner, and preposition*.
- Lists of proper names, also known as named entity dictionaries, used to assign extracted terms to types. Types include organizations, people, locations, products.

Users don't need to customize dictionaries in order to obtain satisfactory results from SPSS text analytics solutions. Text miners can, however, enhance extraction efficiency through user-defined dictionaries. These dictionaries may include:

- Type dictionaries, which assign a particular category type to a word. For example, you can create types that commonly occur in your industry, and include your company's product names—so that, for example, an organization reviewing warranty claims related to automobile sales could correctly type the various car parts referred to in the documents.
- Exclusion dictionaries, which force concepts to be excluded in the concept database
- Synonym dictionaries, which identify terms with similar meanings in order to produce concepts with a higher degree of significance. These dictionaries are also used to define acronyms.
- Keyword dictionaries, which identify products, organizations, names, terms, and locations by verifying the presence of words
- The global dictionary, which overrides type and keyword dictionaries to reconcile ambiguities between these dictionaries for specific words (in specific domains)

SPSS' text analytics products include a built-in Resource Editor. The Resource Editor enables users to edit existing dictionaries, create and edit custom dictionaries, and create specialized rules, such as those governing the Text Link Analysis

step (described on pages 11-13). Through the Resource Editor, users can create custom type assignments. For example, a company reviewing documents related to automobile industry might want to define a type assignment for specific car models.

Also, with the Resource Editor, linguistic resources developed for one application can be shared by different applications and users. Analysts can easily import existing dictionaries, as well as export and share templates and libraries of user-defined dictionaries, set up rules, and define priorities to be considered during the term extraction process.

Term extraction

The SPSS text analytics solution begins the concept extraction process by identifying *candidate terms*, which are then subject to further analysis. Candidate terms are words or groups of words that are used to identify concepts in the text.

To facilitate this, SPSS solutions have specific built-in techniques for identifying linguistic and non-linguistic entities.

Identification of non-linguistic entities

SPSS' text analytics solutions allow for extraction of entities within text that are not considered words. These non-linguistic entities include: URLs, e-mail and IP addresses, phone numbers, Social Security numbers, currency, times and dates, weights, and measures.

The text analytics solution uses a set of rules known as "regular expressions" to extract known patterns for these non-linguistic entities. For example, a number with the format *999-99-9999* would be extracted and typed as a U.S. Social Security number. Similarly, a number such as +33.1.55.55.5555 would be extracted and typed as a French phone number. For broader applicability, users can define their own rules for identifying non-linguistic identities.

Identification of linguistic entities

After named and non-linguistic entities have been identified, SPSS text analytics software uses linguistic extraction techniques to identify relevant words and groups of words from the input text. A one-word term is known as a *uni-term*; terms made up of more than one word are called *multi-terms*.

Single words that are not in the extraction dictionary are considered uni-terms. Specific treatment is applied to uni-terms, based on the value they provide to the analysis.

Candidate multi-terms are often grammatically/linguistically structured as noun phrases. These multi-terms are identified using part-of-speech pattern extractors. For example, the multi-term sports car, which follows the English language "noun noun" part-of-speech pattern, has two components. The multi-term *fast sports car*, which follows the "adjective noun noun" part-of-speech pattern, has three components. There are typically about 15-20 patterns per language; the maximum pattern size is about seven components, depending on the language.

SPSS solutions are shipped with standard term patterns, which are mainly noun phrases. However, users can easily write their own patterns by using the software's Resource Editor (described on pages 8-9)

Type assignment

Once terms have been extracted, they are assigned a type. Type assignment makes it easier to understand the content of a text document.

One step in this process is the identification of named entities. Named entities include persons, companies, product names, and locations. Many times, the lists of named entities play an important role in defining categories or in discovering relationships that, in turn, shed light on certain conditions or behavior patterns.

Named entity and internal dictionaries are used to verify the presence of words or patterns and categorize a term as a named entity. These dictionaries feature an exhaustive list of first names by language; if identified, the text analytics solution treats first names as candidate terms. Additionally, the text analytics solution uses a special algorithm to handle upper-case letter strings, such as job titles, so that these special patterns can be extracted. Scientific terms, such as genes, amino acids, and proteins, can also be identified, using extensions of the linguistic rules embedded in the application.

SPSS text analytics solutions also employ compiled and user-defined dictionaries, to assign a semantic type to other extracted terms. The solution reviews the extracted list of terms using a system of priorities. The compiled dictionaries enforce a specific order for typing organizations, individuals, products, and locations. User-defined dictionaries are applied according to the order in which they are defined in the Resource Editor.

Creation of equivalence classes

An equivalence class is a single form of several variants of the same word or phrase.

The SPSS text analytics solution uses a set of synonym files and built-in algorithms to compare candidate terms and identify equivalence classes. This function ensures that, for example, *cancer of the thyroid* and *thyroid cancer* are treated as the same term. Additionally, it ensures consistency among extracted terms over several extraction runs.

Users can also force the substitution of a term by another one. For example, replace:

- *mgr* by *manager*
- *d/k* by *don't know*

The text analytics solution will always apply substitutions, even if the substituting term is not in the documents themselves.

In addition, the text analytics solution uses *fuzzy logic* to group similar terms without requiring any user-defined resources. It identifies spelling variants by removing vowels and double or triple consonants and then performing a comparison. For example:

- *techinical support = technical support*
- *furnature = furniture*
- *addidas = adidas*

These functions are extremely useful when the text quality is poor, as may be the case for open-ended survey responses, e-mail, and CRM data.

SPSS text analytics software also uses built-in algorithms to detect and “correct” the following: (For more information on these algorithms, please see Appendix B on page 21.)

- Removal of inflection suffixes. For example, *american consumer* = *american consumers*.
- Removal of function words. So that *production of the pentium* = *production of pentium*.
- Variants in separators. Stress free = stress-free. Additionally, *stressfree* = *stress-free*.
- Permuted components.
  For example, officials of the *companies* = *company officials*.
- Accented/non-accented characters. Therefore, *evguéni primakov* = *evgueni primakov*.

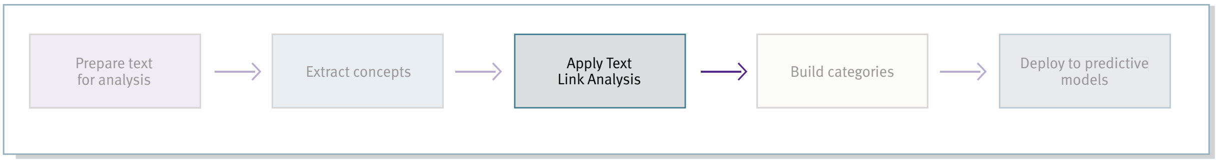
To determine which concept to use for the equivalence class lead term, the extractor component applies the following rules, in order:

- User-specified synonym
- The most frequent form of the term in the corpus
- The shortest form of the term (which usually corresponds to the base form of the term)
- The first one that occurs in the list of extracted terms

Indexing

At the conclusion of the extraction process, the text analytics solution presents a list of extracted terms, grouped and typed. Indices show how often a term occurs in each document and in the corpus as a whole. Indices are presented for each document in the corpus.

Workflow



Uncovering opinions, relationships, facts, and events through Text Link Analysis

Once the extraction process is complete, analysts have the option of using Text Link Analysis to describe relationships between concepts at the sentence level, as well as any opinions or qualifiers attached to these concepts.

Also used to describe facts and events, Text Link Analysis enables analysts to identify and segregate positive and negative concepts in text responses. In addition to simple positive/negative statements, SPSS text analytics solutions provide insight into positive or negative attitudes by “reading” contextual cues, such as sentence structure. In this way, sentiments like those in the sentences below would be grouped correctly, despite the fact that one opinion is positive, one is negative, and one is mixed:

> *The hotel manager was very courteous.*
> *The hotel manager was really rude.*
> *The hotel staff was courteous but the room was too small.*

By combining typed terms (i.e., persons, organizations, genes, etc), linguistic dependencies, literal strings, and Boolean operators, Text Link Analysis allows complex linkages to be discovered and output provided in a user-defined format. Text expressions can be transformed into data that can be quantified and combined with other quantifiable results.

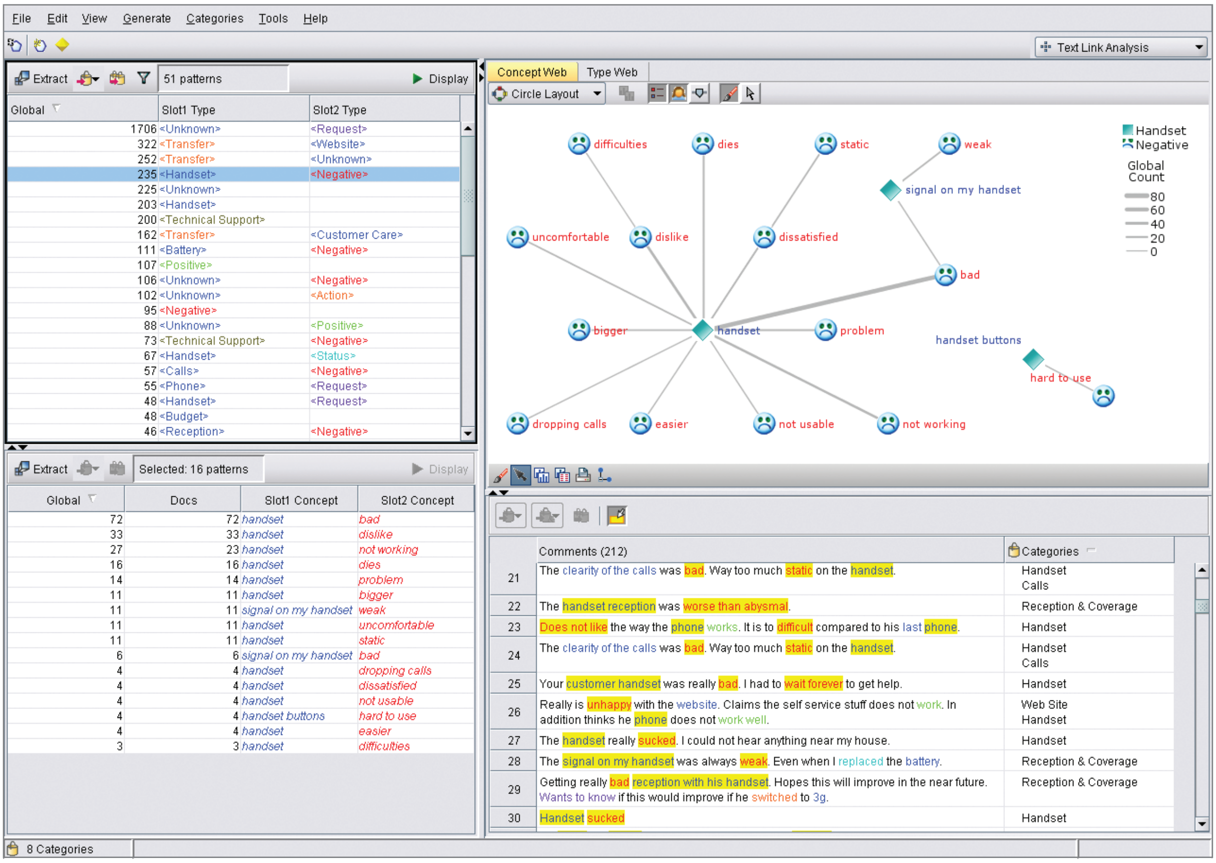
Organizations might use such data to predict, for example, which types of customers (by demographic, by value, by industry) are likely to care most about certain product or service features.

Organizations can also use Text Link Analysis' ability to uncover connections between facts and events to support initiatives as diverse as market intelligence, fraud detection, and life sciences research. NLP-based text analytics would determine that the following three phrases all have the same meaning:

> *Company A was acquired by Company B*
> *Company B acquired Company A*
> *Company B's acquisition of Company A is complete*

And, if a text document should read "Company B failed to acquire company A," TLA would correctly identify that the transaction did not occur.

Rules governing how Text Link Analysis works in SPSS text analytics products reside in the Resource Editor. For examples of Text Link Analysis, please see Appendix C on p. 22.



Global	Slot1 Type	Slot2 Type
1706	<Unknown>	<Request>
322	<Transfer>	<Website>
252	<Transfer>	<Unknown>
235	<Handset>	<Negative>
225	<Unknown>	
203	<Handset>	
200	<Technical Support>	
162	<Transfer>	<Customer Care>
111	<Battery>	<Negative>
107	<Positive>	
106	<Unknown>	<Negative>
102	<Unknown>	<Action>
95	<Negative>	
88	<Unknown>	<Positive>
73	<Technical Support>	<Negative>
67	<Handset>	<Status>
57	<Calls>	<Negative>
55	<Phone>	<Request>
48	<Handset>	<Request>
48	<Budget>	
46	<Reception>	<Negative>

Global	Docs	Slot1 Concept	Slot2 Concept
72	72	handset	bad
33	33	handset	dislike
27	23	handset	not working
16	16	handset	dies
14	14	handset	problem
11	11	handset	bigger
11	11	signal on my handset	weak
11	11	handset	uncomfortable
11	11	handset	static
6	6	signal on my handset	bad
4	4	handset	dropping calls
4	4	handset	dissatisfied
4	4	handset	not usable
4	4	handset buttons	hard to use
4	4	handset	easier
3	3	handset	difficulties

	Comments (212)	Categories
21	The clearity of the calls was bad. Way too much static on the handset.	Handset Calls
22	The handset reception was worse than abysmal.	Reception & Coverage
23	Does not like the way the phone works. It is to difficult compared to his last phone.	Handset
24	The clearity of the calls was bad. Way too much static on the handset.	Handset Calls
25	Your customer handset was really bad. I had to wait forever to get help.	Handset
26	Really is unhappy with the website. Claims the self service stuff does not work. In addition thinks he phone does not work well.	Web Site Handset
27	The handset really sucked. I could not hear anything near my house.	Handset
28	The signal on my handset was always weak. Even when I replaced the battery.	Reception & Coverage
29	Getting really bad reception with his handset. Hopes this will improve in the near future. Wants to know if this would improve if he switched to 3g.	Reception & Coverage
30	Handset sucked	Handset

SPSS text analytics solutions allow you to see a list of extracted concepts and opinions together with visualizations such as the web graph at the upper right. The graph indicates concepts or opinions that are found together in surveys or other documents, with thicker lines indicating those that are found together more frequently.

Elements of the Text Link Analysis module

The Text Link Analysis module comprises three sections: variables, macros, and rules.

A variable can be seen as a “semantic class”: that is, it corresponds to the types assigned by the extractor engine during the type assignment step. All the extracted terms grouped in the same type will, therefore, be grouped as the same variable. A variable definition consists of the following syntax:

- A unique variable name
- A type

For instance, where Person is the name of the variable as used in macros and rules, and P is the internal type code assigned by the extractor component:

```
[variable] name=Person
value=P
```

A macro is used within a pattern to group variables or lists of words and to simplify pattern rules. A macro definition consists of the following syntax:

- A unique macro name
- A definition—that is, the list of variables, words, and/or macros

Let’s suppose the three variables Positive, Negative, and Contextual, and the macro mOpinion:

```
[macro]
name=mOpinion
value=($Positive|$Negative|$Contextual)
```

Instead of writing a rule with ($Positive|$Negative|$Contextual), you can use the macro $mOpinion instead, because the two are equivalent.

A pattern is a Boolean query that is used to perform a match on a sentence. Patterns contain one or more of the following elements: variables, macros, or literal strings. The syntax for patterns is as follows:

- A unique pattern ID#
- A pattern name (need not be unique)
- The value (the pattern syntax to match)
- The output (the format to be created when the pattern is matched). There may be several outputs for a single rule, on a single sentence or part of sentence (especially in the case of coordination).

For example, let’s suppose the following rule, where #@# John Doe is the director of ABCD Inc. in France.

```
[pattern(201)]
name = 1_201
value = $Person ($SEP|$mDet|$mSupport|as|then){1,2} @{0,1} $Function (of|with|for|in|to|at) @{0,1}
$Organization @{0,2} $Location
output(1) =$1\t#1\t$4\t#4\t$7\t#7\t$9\t#9
```

The extractor component will read each sentence and will try to match the following sequence:

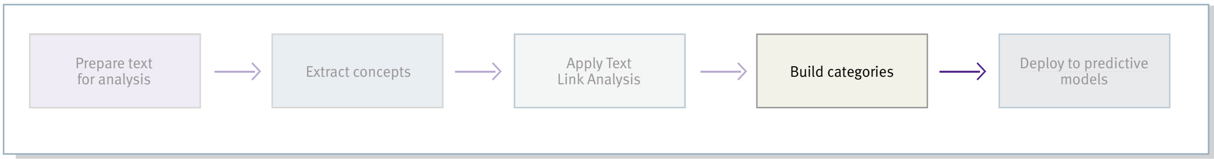
- Any person name, followed by
- One or two commas ($SEP), determiner ($mDet), auxiliary verb ($mSupport), the strings “then” or “as,” followed by:
- 0 or 1 word (@{0,1}, followed by
- A function ($Function), followed by
- One of the following strings: “of”, “with”, “for”, “in”, “to”, or “at”, followed by
- 0 or 1 word (@{0,1}, followed by
- An organization name, followed by
- 0, 1, or 2 words (@{0,2}, followed by
- A location name ($Location)

This will match sentences like:

- John Doe, the director of ABCD Inc. in France
- John Doe is the director of ABCD Inc. in France
- Company C appointed Jane Doe
  as the CEO of DFG Ltd for the United States

Patterns are compiled, based not on their order of appearance but on their ID#. Because the first rule that matches a pattern “wins”—keeps other rules from matching—it is important that the most specific patterns be declared first, and then the more general ones.

Workflow



Building categories

Building categories and categorizing documents are the next steps in analyzing text documents.

Because every dataset is unique, the choice of techniques and the order in which a researcher would apply them are likely to vary from one project to another. In all cases, however, the classification process is iterative: a researcher applies certain techniques, evaluates the results, makes changes either to the technique chosen or to the resultant categories, and refines the results.

Both automated and manual classification techniques are available with SPSS solutions.
The automated, linguistics-based techniques available include:

- Concept derivation
- Concept inclusion
- Semantic networks
- Co-occurrence rules

These techniques can be used both on noun terms and on qualifiers or adjective terms. They classify terms by identifying those that are likely to have the same meaning (also called *synonyms*) or are more specific than the category represented by a term (also called *hyponyms*). For cleaner results, these linguistic techniques exclude adjective terms and other qualifiers.

Concept derivation is a technique that classifies a concept by finding others that are related to it. This is done by analyzing whether any of the concept's components are morphologically related. For example, the concept "opportunities to advance" would be grouped with the concepts "opportunity for advancement" and "advancement opportunity." This technique works well with text data of varying lengths and generates a small number of compact groups.

Concept inclusion groups concepts by looking for concepts that are included in other concepts. For example, the terms "relational database" and "multidimensional database" would be grouped with the term "database." A concept series based on inclusion often corresponds to a taxonomic hierarchy (a semantic "ISA" relationship).

This technique begins by identifying uni- or multi-terms that are included in other multi-terms (and positioned as suffix, prefix, or optional elements) and then groups them together. When determining inclusion, the algorithm ignores word order and the presence of function words such as "in" or "of." This technique works with text of survey response data of varying lengths and typically generates a large number of compact groups.

Semantic networks group terms based on known word relationships contained in a built-in network. This technique begins by identifying the possible senses of each concept. Concepts that are synonyms or hyponyms are then grouped together. This technique can produce very good results when the concepts are known to the semantic network and are not too ambiguous. It is less helpful when text contains a large amount of specialized domain-specific terminology unknown to the network. In the early stages of classifying terms, users may want to use this technique by itself to see what sort of categories it produces.

Co-occurrence rules based on *co-word analysis* are used to group terms based on how frequently the terms co-occur within the set of documents. Terms strongly co-occur if they appear frequently in the same documents, survey responses, or other text and if they occur separately only rarely. This technique can produce good results, especially with larger datasets.

Co-occurrence rules enable you to discover and group concepts that are strongly related within the set of documents or records. When using this approach, analysts can limit the number of co-occurring concepts that can be grouped together into a rule. They can also speed up the categorization process by limiting the number of documents or records to be used when creating categories.

Modifications available for improved effectiveness

When building categories using linguistics-based techniques, users may select specific techniques and then modify parameters such as the number of categories to be created, or the number in which a single term can appear.

For example, if a semantic network has been selected as one of the techniques, the analyst might select the profile to define the behavior of the underlying algorithms—either "wide" or "narrow." A wide profile handles ambiguous terms effectively. It creates more categories but may group terms into categories that are not closely linked to the context of your data. A narrow profile excludes very ambiguous terms and focuses on the clearest relationships between terms. It will tend to create fewer, smaller categories.

Additionally, users can define the minimum proximity score required for grouping terms. The lower the score, the more results will be shown; however, these results may be more ambiguous. By selecting a higher score, an analyst may obtain fewer results, but these results are more likely to be significantly linked or associated with each other.

Another way to influence category creation is to set a minimum number of times that a concept must co-occur in the text in order for it to be extracted. For example a value of two limits the extraction to those concepts that occur at least twice in the set of records or documents.

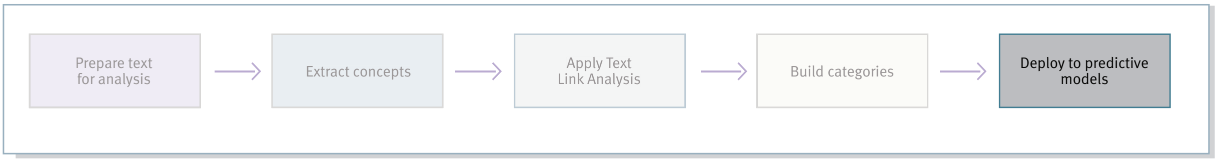
In combination with the automated linguistic techniques described above, manual techniques enable users of SPSS text mining solutions to include terms in groups (or specifically exclude them), using drag-and-drop functionality. In addition, users can apply their own code frames or import categories that have previously been exported from SPSS Text analytics for Surveys™. Yet another option is to copy, paste, and edit category codes and code frames by using the Code Frame Manager feature.

As categories are developed, users have a number of options for refining them. For example, an analyst might:

- Add concepts or opinions to a category definition
- Edit a category definition
- Merge categories
- Move categories from one "branch" of the tree diagram to another
- Delete categories
- Create visual graphs that show how categories work together, and then make adjustments
- Make some changes to the linguistic resources used, and then re-extract categories

Once categories have been created, organizations can assign identifiers to incoming comments, e-mails, or documents based on the likelihood that certain types of terms will occur in the text. This document categorization enables organizations to efficiently deliver comments or documents to appropriate individuals, groups, or systems.

Workflow



Deploying results to predictive models

Deployment of text analytics results to predictive models is the step that links text analytics to decision making. In early implementations of text analytics, deployment consisted of creating visualizations of concept relationships and perhaps embedding these in reports. Reports then had to be interpreted by managers before strategic or tactical plans could be developed. More recently, organizations began to employ batch scoring—often conducted at off-peak hours—to more efficiently incorporate updated predictions based on text analytics models in their databases.

Currently, computer processing efficiencies and other technological advances make it possible to analyze massive amounts of textual data in just a few hours. Specialized reports can be created and routed based on an individual's role or their membership in a specific work group. Alternatively, models can be integrated with systems to generate sales offers automatically, identify creditworthy customers immediately, highlight extremely positive or negative customer or citizen feedback, or suggest patterns of possibly criminal behavior—to name just a few examples.

As organizations develop a large number of models, these models and the processes associated with them require the same type of management attention that any other valuable asset demands. Recognizing this, SPSS offers a solution to store such assets centrally and securely, and ensure that correct models are used in analysis—and that there is an auditable record of who has accessed, changed, or applied each model.

Applying text analytics at the enterprise level

Organizations that understand the value of text analytics typically begin by addressing a specific business problem.

- A college or university may conduct surveys, hoping to increase the level of satisfaction students feel about their institution's course offerings and, by doing so, identify changes that could improve student retention
- A company experiencing a high rate of customer defection or "churn" may analyze customer feedback contained in call center notes and compare patterns discovered in this text to specific customer behaviors, so that customer contact staff can recognize when a customer is at risk of leaving and take appropriate action to minimize the likelihood of this occurring
- A pharmaceutical company might evaluate the effectiveness of a certain treatment regimen by incorporating text comments from study participants describing how they felt before, during, and after the treatment
- A business that relies on evaluating massive amounts of textual information may use text analytics to identify trends or patterns, enabling staff members to focus their attention on the most relevant documents for increased productivity
- A market research agency or a large company with international operations may have to field surveys in 20 different countries. They obtain verbatim responses in 15 different languages and don't have the resources to analyze all those languages natively. Through advanced translation, they can perform sentiment analysis and centralize all results in English.
- An intelligence agency may need to review documents, phone transcripts, or e-mails in several different languages to uncover relationships among terrorist cells

Once organizations achieve some success using text analytics, they often want to employ text analytics in other departments or geographic regions or to address other business problems. Conducting text analytics at the enterprise level can significantly increase the return on an organization's investment in text analytics and related technologies. It does present several challenges; however, with the right text analytics solution and the right guidance, these challenges can be met successfully.

Conclusion

The challenge of text is part of the larger information challenge organizations face today.

As the amount of available information has vastly increased in the last decades, so has the importance of being able to locate information quickly, distinguish the relevant from the irrelevant reliably, and share insight with others to support tactical responsiveness and strategic planning. Mastering the new challenges represented by this information explosion can mean a significant competitive advantage for businesses, and a marked increase in effectiveness for researchers and public service organizations.

Organizations that adopted SPSS linguistics-based text analytics technologies—with the goal of using all of their data more effectively and strategically—have experienced measurable benefits. Their experiences were described in a recent report by independent analyst firm Nucleus Research in *Guidebook: SPSS Text Mining*. Examples of benefits cited in the report include the following:

- By using insights gained from customer comments, one telecommunications company has seen 51 percent of its dissatisfied customers become company promoters (very satisfied customers) after only two months
- Organizations such as insurers and financial institutions can leverage call center data and combine it with other information to identify better programs to retain profitable customers
- In some companies, analysts were able to increase their productivity by up to 50 percent
- A company in the technology sector uses text analytics to quickly provide senior managers with feedback on a particular product line

As one user interviewed for the report stated: “Before, we’d have to choose: do we want qualitative research or quantitative? Now we are not having to make that decision because we can offer the combination.”

SPSS text analytics solutions have the depth of techniques, as well as the scalability and customizability to support any organization’s text analytics challenges. In addition, they support the proven CRISP-DM methodology and have an open architecture that makes the findings gained through text analytics available to other organizational systems and processes.

By improving the relevancy and accuracy of predictive models, SPSS text analytics solutions help organizations reap significant, measurable benefits from textual data, and gain an advantage in meeting the new challenges—and opportunities—posed by current and future waves of textual information.

SPSS products for text analytics

SPSS offers several text analytics solutions:

- **Text Mining for Clementine®** provides “best of breed” text analytics capabilities through user-friendly interfaces. It is an add-on to SPSS’ Clementine data mining workbench (both client and server versions). It employs SPSS’ linguistics-based text analytics technologies to discover concepts and relationships in text and then perform classification, clustering, and other statistical techniques on those concepts. Text Mining for Clementine can process text natively in English, French, Dutch, German, Spanish, Italian, Portuguese, and Japanese, and by translation from more than a dozen other languages. Text Mining for Clementine supports the creation of models that combine text or unstructured data with tabular or structured data to more reliably predict conditions, actions, or behaviors. In addition, models developed with Text Mining for Clementine can be embedded in operational systems through SPSS’ predictive applications.
- **SPSS Text analytics for Surveys™** is a desktop tool that uses SPSS’ text analytics technologies to quantify free-form text responses found in surveys so that opinions and sentiments can be analyzed along with other survey data. SPSS Text analytics for Surveys automates the classification and categorization of text concepts while still allowing users to intervene manually to refine results. Results can be exported either as tables of records or as dichotomies for further analysis, using SPSS statistical software.
- **SPSS Predictive Enterprise Services™** provides a central repository for text analytics results and offers automation, authoring, and versioning capabilities to applications using SPSS text analytics products.

About SPSS Inc.

SPSS Inc. (NASDAQ: SPSS) is a leading global provider of predictive analytics software and solutions. The company’s predictive analytics technology improves business processes by giving organizations consistent control over decisions made every day. By incorporating predictive analytics into their daily operations, organizations become Predictive Enterprises—able to direct and automate decisions to meet business goals and achieve measurable competitive advantage.

More than 250,000 public sector, academic, and commercial customers rely on SPSS technology to help increase revenue, reduce costs, and detect and prevent fraud. Founded in 1968, SPSS is headquartered in Chicago, Illinois. For additional information, please visit **www.spss.com**.

Appendix A: An explanation of some text analytics terms

Term	Explanation
Blog	A Web site that provides commentary on a specific topic. Readers can add comments in an interactive format. Entries are usually displayed in reverse chronological order. Also known as a “Web log.”
Bayesian network	A probabilistic graphical model that represents a set of variables and their probabilistic independencies. For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network could compute the probabilities of the presence of various diseases.
Boolean logic/queries	A term from mathematical logic to indicate propositions linked by the three fundamental logical operations: and, or, and not
Candidate term	A term representing an equivalence class and retained for purposes of cross-indexation
Category	Any of several fundamental and distinct classes to which entities or concepts belong
Classification	The grouping of a set of entities sharing certain formal or external properties
Clustering	The process of grouping items such as documents on the basis of similarity. The goal is to divide a dataset so that similar records are in the same group, and so that groups are as different from each other as possible.
Computational linguistics	A branch of linguistics that uses computers to model language systems. It encompasses automatic parsing, machine processing, and computer simulation of grammatical models for the generation and parsing of sentences. Its goal is the modeling of human language as a cognitive system.
Concept	An abstract or generic idea generalized from particular instances
Concept class	A group of similar concepts that is distinct from other groups
Equivalence class	A group of inflected terms represented by one form. This form, retained for indexation, is called the candidate term. Generally, it is the most frequently found form of a term, or the form explicitly defined by the user.
Event extraction	The process of finding the occurrence of concepts and relationships through an understanding of the sense of a body of text. Events may include a person’s job, or an occurrence in the real world, such as a merger or acquisition, a disease outbreak, a terrorist attack, etc.
Fuzzy logic	A term derived from mathematics, and referring to the indeterminacy involved in the analysis of a linguistic unit or pattern
Indexing	The process of finding key concepts within a set of documents and developing a map from the concepts to the documents in which they are found
Key words	The most important and discriminating words in a document set
Latent semantic analysis	A patented mathematical or statistical technique for extracting and representing the similarity of meaning of words and passages of text by analyzing large amounts of text using a general form of factor analysis
Linguistics	The study of the general and universal properties of language
Morphology	The branch of grammar that studies the structure or forms of words
Natural language processing	Computer analysis and generation of natural language text. The goal is to enable natural languages to serve either as the medium through which users interact with computer systems, or as the object that a system processes into some more useful form.
Precision	The measure of how well information retrieval systems select documents that are relevant to a query

Recall	The measure of how well information retrieval systems find all the documents that are relevant to a query
Relevence	A measure of the success of an information system to deliver material that satisfies the needs of the user
Semantics	A major branch of linguistics devoted to the study of meaning in language
Statistics	A set of methods used to derive general information from specific data. The term is also used to describe the computed values derived from these methods.
Stop word	A commonly used word (such as “a” or “the”) that an NLP program has been programmed to ignore, both when extracting concepts from documents and when developing indices.
Support vector machines	A set of related supervised learning methods used for classification and regression belonging to a family of generalized linear classifiers. A special property of support vector machines (SVMs) is that they simultaneously minimize the empirical classification error and maximize the geometric margin; hence, they are also known as maximum margin classifiers.
Synonymy	The condition in text of having several terms with the same meaning
Syntax	The branch of grammar that deals with the rules governing the combination of words in sentences
Taxonomy	The practice and science of classification. Taxonomies, or taxonomic schemes, are composed of units known as taxa that are arranged in a hierarchical structure, typically related by subtype-supertype relationships.
Term	A word or expression that has a precise meaning in some uses, or is specific to a science, art, profession, or subject
Text analytics	The process of automatically extracting information from large collections of documents
Text Link Analysis	A technique for identifying and segregating positive and negative concepts, as well as facts and events, in a body of text
Thesaurus	A collection of synonyms and antonyms. Thesaurus databases, created by international standards, are generally arranged hierarchically by themes and topics. By placing each term in context, such a thesaurus allows users to distinguish between similar terms with different meanings. Often used as the basis for indexing online material. Also referred to as an ontology.
Wiki	Software that enables users to easily create, edit, and link Web pages. Wikis are often used to create collaborative websites for knowledge management.

Appendix B: Algorithms used for assigning equivalence classes

In the SPSS text analytics solution, the following algorithms are applied to assign concepts to equivalence classes.

Inflection

vasopeptidase inhibitors = vasopeptidase inhibitor

Synonymy

Full-Form: an entire extraction is equivalent to another

familial hyperchylomicronemia = familial lipoprotein lipase deficiency

Component: two distinct extractions are equivalent, modulo variation in components

colour blindness = color blindness

Omission of keywords

ziff-davis inc = ziff davis

Geographic variant

tumour = tumor

Lexical variant

geographical markets = geographic markets

Lower-case/upper-case characters

apolipoprotein A = apolipoprotein a

Omission of/variation in function words

ulceration of the mucosa = ulceration of mucosa éclipses du soleil = éclipse de soleil

Variants in separators

Separators may be space, hyphen, agglutination, apostrophe *s* (or apostrophe), or dot

zollinger-ellison syndrome = zollinger ellison syndrome

health care = healthcare

web-tv = web tv

webtv = web tv

alzheimer disease = alzheimer's disease

Inversion of components

generalized myotonia of Becker = Becker's generalized myotonia

cancer of the thyroid = thyroid cancer

zeste râpé d'un citron = zeste de citron râpé

Accented/non-accented characters

This phenomenon may be very frequent in languages such as French, Spanish, Italian, or Dutch

são Paulo = sao Paulo

evguéni primakov = evgueni primakov

évènements du kosovo = événements du kosovo

Generic-specific

Grouping extracts under a normalized term can be seen as finding the “best descriptor.” In some applications, specific terms could be mapped to generic terms.

lipstick = cosmetics

eyebrush = cosmetics

Spell checking/fuzzy matching

Based on omission of vowels or double consonants, or other algorithms

technichal support = technical support

techinical support = technical support

Appendix C: Examples of Text Link Analysis

Open-ended surveys, call center data, and data from other CRM systems:

From the sentence, “I have found support services to be very helpful, friendly, and courteous,” Text Link Analysis would match:

```
[pattern(0306)]
name = 0306 _positive_opinion
value = $mExtract @{0,2} ($mSupport|would|could|to) @{0,1} (a|rather|quite|pretty|very)? $mOpinion $SEP? $mOpinion ($SEP|$mCoord){1,2} $mOpinion
output(1) = $1\t#1\t$6\tPositive
output(2) = $1\t#1\t$8\tPositive
output(3) = $1\t#1\t$10\tPositive
```

This leads to the understanding that:

support services ‹Unknown› helpful ‹Positive›

support services ‹Unknown› friendly ‹Positive›

support services ‹Unknown› courteous ‹Positive›

From the sentence, “My problem has not been solved,” Text Link Analysis would match:

```
pattern(011)]
name = 011
value = ($mTopic|$Negative) @{0,1} $mAdvNeg @{0,1} $Positive
output= $1\t#1\tnot $5\tNegative
```

problem ‹Negative› not resolved ‹Negative›

Genomics:

From the sentence, “studies with the protein kinase C inhibitor, Calphostin C,...,“ Text Link Analysis would match:

```
[pattern(003)]
name = (003)
value = $Gene $Agent $SEP? $Gene
output(1) = $4\t#4\t$2\t#2\t$1\t#1
```

This leads to the understanding that: calphostin C ‹Gene› inhibits ‹Action› protein kinase C ‹Gene›.

Market intelligence:

From the sentence, “SPSS Inc. completes acquisition of LexiQuest,” Text Link Analysis would match:

```
[pattern(303)]
name = 303
value = $Org @{0,1} $mSupport $Action of @{0,2} $Org
output = $1\t#1\t$4\t#4\t$7\t#7\tcompleted)\tStatus
```

This leads to the understanding that:

spss inc. ‹Organization› acquires ‹Action› lexiquest ‹Organization› completed ‹Status›

Additional reading on text analytics

Books

Andersson, Birger, Maria Bergholtz, and Paul Johannesson (Eds.) *Natural Language Processing and Information Systems: 6th International Conference on Applications of Natural Language to Information Systems*. NLDB 2002, Stockholm, Sweden, June 27-28, 2002: Revised Papers. (Lecture Notes in Computer Science, 2553, Heidelberg: Springer-Verlag, 2002)

Berry, Michael W. and Malu Castellanos (Eds.). *Survey of Text Mining II: Clustering, Classification, and Retrieval.* London: Springer-Verlag London Ltd, 2008

Feldman, R. and J. Sanger. *The Text analytics Handbook*. Cambridge, England: Cambridge University Press, 2007.

Jackson, Peter and Isabelle Moulinier. *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*. Amsterdam: John Benjamins Publishing Company, 2002.

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

Manning, Christopher D. and Hinrich Schutze. *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press, 2001.

Sullivan, Dan. *Document Warehousing and Text analytics: Techniques for Improving Business Operations, Marketing and Sales*. New York: Wiley Computer Publishing, 2001.

Articles and Papers

Anderson Analytics, LLC. *Leverage the Voice of Your Customers*. Stamford, Conn. June 2007.

Grimes, Seth. "A Brief History of Text Analytics," b-eye-network, October 20, 2007.
http://www.b-eye-network.com/view/6311

Hearst, Marti A. "Untangling Text Data Mining." *Proceedings of the ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics*. College Park: University of Maryland, June 1999.

Jouve, O. et al. "Two measures for identifying the perception of risk associated with the introduction of transgenic plants." *Scientometrics,* 1999, Vol 44, No. 3, pp. 401-426.

— "Leximappe is dead: long live co-word analysis! Application to identify the main actors within the field of risk assessment through the introduction of transgenic plants." 1998: International Conference on Science and Technology Indicators: Use of ST indicators for science policy and decision-making. Hinxton (Great Britain).

Martin, E., E. Bremer, MC. Guerin, C. DeSesa, and O. Jouve. "Analysis of Protein-Protein Interactions through Biomedical Literature: Text Mining of Abstracts vs. Text Mining of Full Text Articles." Knowledge Exploration in Life Science Informatics, International Symposium, KELSI 2004, Milan, Italy, November 25-26, 2004, Proceedings.

Nucleus Research. *Guidebook: SPSS Text analytics*. Document H99. Wellesley, Mass. December 2007.

Other resources

— Association for Computational Linguistics www.aclweb.org
— Information about other groups conducting research in computational linguistics and natural language processing www.dmoz.org/Computers/Artificial_Intelligence/Natural_Language
— Text Analytics Summit http://www.textanalyticsnews.com/

To learn more, please visit www.spss.com. For SPSS office locations and telephone numbers, go to www.spss.com/worldwide.