Meeting the Challenge of Text
Making text ready for predictive analysis

Table of contents
Executive Summary .................................................................2
Meeting the Challenge of Text ..................................................3
Benefits of Text Mining and Predictive Text Analytics ......................5
A Primer on Text Mining ..........................................................7
Text Mining Technologies from SPSS ........................................8
Requirements for a Text Mining Solution ....................................12
Conclusion .............................................................................15
Appendix A: Text Mining Terms ...............................................17
Appendix B: Algorithms Used for Assigning Equivalence Classes ..............19
Executive Summary

The gains organizations realize through predictive analytics are becoming increasingly apparent. Predictive analytics connects data to effective action by turning information organizations already have into insights that lead to improved performance.

Now, textual data can be incorporated in the predictive analysis process, leading to a richer understanding of operational and competitive conditions, customer attitudes, preferences, and behavior. Analyzing text is also critical in specialized research domains. It enables organizations to leverage vast amounts of textual information that are currently underutilized, or even lost.

Predictive Text Analytics applies classification, clustering, and other data mining techniques to information gained through text mining, in order to predict trends in attitudes, behavior, and business conditions. The first step in making text ready for predictive analysis is to apply appropriate text mining technologies.

Text mining is the process of analyzing collections of textual materials in order to capture key concepts and uncover hidden themes, relationships, and trends.

This paper describes some of the principal benefits organizations can obtain from text mining and predictive text analytics. Benefits include:

- Improving your understanding of customer interactions
- Improving your understanding of market research data
- Shortening your R&D process
- Identifying possible security threats or illegal activity
- Obtaining greater ROI from other data mining efforts

Applying text mining to real-world challenges is still relatively new, so this paper provides an overview of text mining, with special attention to linguistics-based text mining—the most successful approach to extracting key concepts and themes from text.

Linguistics-based text mining technologies provide more accurate and useful results than other approaches to managing text, and do so cost effectively. These technologies can be applied to a group of lengthy documents, such as a body of scientific research, or to a collection of short comments, such as those found in call center logs, open-ended market research surveys, or e-mails.

The first step in text mining is the extraction of concepts from text. Once concepts have been extracted, organizations can provide this information to predictive analytics solutions. In this way, they can efficiently and reliably uncover patterns, affinities, and trends in their data, and deploy these results to both individuals and operational systems to guide customer interactions and strategic decision making.

In evaluating a linguistics-based text mining solution, decision makers should set guidelines for the following critical factors:

- Scalability—the ability to process a sufficient quantity of information in a given time period
- Accuracy—the ability to uncover all relevant information, and separate it from the irrelevant
- Customizability—the ability to adapt to the particular needs of your organization, industry, or field of research
- Discovery-orientation—the ability to expose previously unknown information or relationships
Meeting the Challenge of Text

Today, organizations like yours can give executives, managers and employees at all levels greater access to information than at any time in history. But this wealth of information presents challenges as well.

To fully understand your operations, you are likely to employ a variety of software tools to convert transactional, operational, and competitive information into “structured” form. You may capture this data in enterprise resource planning systems, or in a variety of relational and multidimensional databases. You have developed processes to create and efficiently deliver reports analyzing this data. You may also use data mining tools that discover hidden patterns and trends in such structured data.

But vast amounts of “unstructured,” or textual, information probably remain unexplored, despite the fact that this information contains important details about your products, markets, customers, and operations. This is largely because until now it has been difficult to automate the analysis of text. New technologies, however, enable you to explore textual data more effectively and gain greater insight from text.

Predictive text analytics is the latest development in analyzing unstructured information. Predictive text analytics uses the results of text mining to predict a variety of events, from a change in customer behavior to a potential security threat. Predictive text analytics delivers measurable benefits to organizations like yours in a number of specific applications, including:

- Improving your understanding of customer interactions
- Improving your understanding of market research data
- Shortening your R&D process
- Identifying possible security threats or illegal activity
- Obtaining greater ROI from other data mining efforts

While this paper includes some discussion of the benefits of text mining and predictive text analytics, it is primarily intended as an introduction to this emerging technology. This paper provides a description of the linguistic approach to text mining, SPSS’ implementation of this approach, and the factors to consider when evaluating a text mining solution.

A Growing Challenge

If you’re feeling overwhelmed by information, you’re not alone. Many studies indicate that individuals and organizations alike are generating more and more textual information, which they save and share in a variety of digital formats.

Textual information includes a wide range of materials, such as industry or technical research papers, patent applications, sales proposals, and notes embedded in sales or service databases. It also includes open-ended comments in market research surveys; customer feedback captured in call center logs, Web chat, e-mails, and other correspondence; and the content of public and private Web pages.

- Technology analyst firm IDC estimates that U.S. corporations exchange more than 8 billion e-mail messages per day. This does not include textual comments saved by inbound or outbound call center staff.1
- The commonly accepted ratio of unstructured to structured business information was 80:20. But a recent Merrill Lynch study estimated that as much as 85% of business information may be in textual format.2
Approaches to Meeting the Challenge

There are a variety of tools to help your organization conquer these mountains of textual data:

- Content and document management systems—essentially warehouses for textual data—make it possible to organize and update textual information cost effectively. Users typically “pull” information from these systems, but some systems can “push” delivery to selected users and groups.
- Portals and intranets help organizations disseminate textual information more efficiently by providing structure to the information that many users need to access.
- Search and information retrieval solutions—which can be embedded in content management solutions and in internal and public Web sites—help individuals locate specific text documents quickly.
- Text mining, categorization, and data visualization solutions support the systematic discovery of information from text and allow users to uncover unsuspected patterns and trends.
- Predictive text analytics uses text mining results to predict future behavior or conditions and support decision making.

An important difference to be aware of is that search, content management, and portals are “top down” solutions, meaning that they enable individuals to locate information they already know exists, or information that contains terms they already know. Text mining solutions and predictive text analytics, by contrast, are “bottom up.” They help you uncover information, even if you do not know it exists, and even if the exact query terms are not contained in any single document.

Using search and information retrieval tools might be compared to knowing the way to the company file room and then checking the labels on file cabinets to find what you need. But using a text mining solution and predictive text analytics is like having the relevant information pulled from the documents in the file cabinets and delivered right to your desk, organized in ways that make it easy to explore relationships and predict future events. For this reason, only text mining can be considered discovery-oriented.

A New Tool: Predictive Text Analytics

Text mining is the process of analyzing collections of textual materials in order to capture key concepts and uncover hidden themes, relationships, and trends. Predictive text analytics applies classification, clustering, and other data mining techniques to information gained through text mining, in order to predict trends in attitudes, behavior, and business conditions.

Increasingly, organizations are applying linguistics-based text mining to aid them in a broad range of endeavors. For example:

- A mobile telecommunications provider used concepts extracted from call center notes to improve the gain of existing churn models by up to 10%
- A public sector organization used text mining to improve the categorization of issues brought up in open-ended survey responses from an employee survey to better understand problems that needed to be addressed.
- A U.S.-based medical research facility combined text and data mining to study relationships between factors that affect the growth of malignant tumors.

Text mining helps alert your organization to new developments in the marketplace while there is still time to mount an effective response. For example, it informs you of customer attitudes and affinities you might not be aware of, and which might lead to lost sales opportunities or even lost customers. In short, text mining helps your company be more responsive and innovative—important success factors in a rapidly changing competitive environment.
Benefits of Text Mining and Predictive Text Analytics

There are many applications for text mining. Through predictive text analytics, which combines data mining techniques with text mining, organizations gain additional benefits. Here is more information on those mentioned on page 3.

Improving the understanding of customer interactions
Customers interact with your organization at many points. One of the most critical is through your call center. Call centers are the hub of many organizations’ sales efforts, accounting for more than $500 billion in sales annually. They also play a vital role in customer service, typically responding to millions of calls per day.

Notes taken during call center interactions can be a valuable source of information. They may point to common problems with a product, or the need for clearer instructions on its use. They may reveal unexpected applications of a product, uncover unmet needs, or opportunities to up-sell or cross-sell other products and services.

Text mining can help organizations group and analyze thousands of call center notes quickly and cost effectively. This can lead to faster identification of customer issues, which leads to the development of plans to correct unsatisfactory situations and/or successfully meet emerging customer needs.

Through predictive text analytics, the results of text mining can be used to improve “lift,” or the effectiveness, of existing data mining models. Since notes on customer contact tend to record dissatisfaction, they typically yield concepts that are effective predictors of intended customer defection. Consequently, the improved lift gained from including textual data has significant and measurable impact on problems such as customer retention.

Improving the understanding of market research data
To survive, business organizations need to understand their customers’ attitudes, preferences, and behavior. Business, government, and academic organizations spend billions annually for market research services, and spend millions more to staff their organizations to plan research efforts and interpret their findings.

Such information includes data automatically generated at point of purchase (whether online or in a physical location), internally generated marketing campaign analyses, and volumes of information from focus groups and customer surveys. Many of these surveys contain open-ended questions. These questions allow survey participants to clarify the strength of their views, or to indicate that they have preferences that were not directly addressed by other portions of the survey.

Because they are free to express their views in their own words, survey respondents often use a variety of terms and phrasings to express similar ideas. These comments are time-consuming to analyze manually. Linguistics-based text mining can create accurately group customer opinions, in effect bringing structure to participants’ unstructured responses.

By automating this process, text mining provides accurate information on people’s attitudes toward products, services, and other issues. This saves organizations from either misinterpreting information or failing to spot emerging trends—both of which can be extremely costly.
Shortening the R&D process
At many organizations, R&D represents a significant investment, both of capital and of other resources. For example, in pharmaceutical and biomedical organizations, the development and approval process for new products can take years and cost hundreds of millions of dollars. Because researchers can now delve into the relationships between pathologies, genes, proteins, drugs and other treatments in ways never before possible, they must spend significant amounts of time simply keeping abreast of developments in their field. They also must monitor the activities of other researchers and of business competitors.

Similarly, manufacturers of industrial and consumer goods, from automobiles to cell phones to video games, generate vast amounts of textual information during the design, production, and marketing of their products. Customer research, competitive intelligence, pricing studies, supply chain information: these are just some of the types of information that must be evaluated when developing and marketing products.

Linguistics-based text mining enables researchers and product development teams to identify and focus on the information that is most relevant to the task at hand. This saves hours, even weeks, of valuable time—and saves their organizations significant sums of money.

Identifying security threats and criminal activity
For some time, data mining has played a role in uncovering patterns in criminal activity, particularly in discovering patterns of fraud, waste, and abuse. As homeland security has grown in importance over the past several years, the need for tools to analyze and disseminate information found in text documents has become widely recognized.

Linguistics-based text mining is particularly effective at uncovering concepts in documents such as e-mails, phone transcripts, and investigators’ notes. It can be used to establish relationships among concepts, and to automate the routing of information to the correct investigative groups. When patterns of fraudulent or criminal activity have previously been identified through data mining, predictive models can be enhanced by adding information gained through text mining.

Obtaining greater ROI from data mining
Data mining enables organizations to see patterns and trends in structured, tabular data more clearly and easily, and create predictive models. Text mining complements data mining by allowing organizations to incorporate information from unstructured, text-based sources in these models, increasing their effectiveness.

For example, organizations in some industries can optimize their marketing efforts by sending customized offers to different customer segments. Other organizations use predictive text analytics to improve customer retention. In both situations, text mining adds depth to the understanding of customer segments, making models built upon this information more reliable and effective.

Organizations involved in biomedical research see significant benefits from adding textual data to their data mining efforts. Other applications include competitive intelligence and the protection of an organization’s intellectual property.
A Primer on Text Mining

There are several different types of text mining. First, there is the manual approach: having people read the documents, note their contents, and determine the key concepts they contain. Because people are good at understanding text, this approach is quite accurate. But it is time-consuming, labor-intensive and, with the immense volume of text now available, often in multiple languages, increasingly impractical. Another drawback to the manual approach is that while it can point individuals to information that is likely to be relevant, it cannot offer guidance in identifying relationships or trends in this information.

A different approach is to employ automated solutions based on statistics and neural networks. Using computer technology, these solutions can scan and identify key concepts more quickly than human readers can. Unfortunately, the accuracy of such solutions is fairly low. Most statistics-based systems simply count the number of times terms occur and calculate their statistical proximity to related terms. Statistics-based approaches produce many irrelevant results (also known as noise) and miss results they should have found (also known as silence).

To compensate for their limited accuracy, some automated solutions incorporate complex rules that help distinguish between relevant and irrelevant results. But these rulebooks need to be created and continually updated by analysts, adding to the total cost and complexity of the solution, and potentially creating information bottlenecks.

Linguistics-based text mining offers the speed and cost effectiveness of statistics-based systems. But it offers a far higher degree of accuracy, while requiring far less human intervention. Linguistics-based text mining is based on the field of study known as natural language processing (NLP), also known as computational linguistics.

To illustrate the difference between statistics-based and linguistics-based approaches, consider how each would respond to a query about “reproduction of documents.”

Both statistics-based and linguistics-based solutions would have to expand the term “reproduction” to include synonyms like “copy” and “duplication.” Otherwise, relevant information will be overlooked. But if a statistics-based solution attempts to do this type of synonymy—searching for other terms with the same meaning—it is likely to include the term “birth” as well, generating a number of irrelevant results.

### Chart 1

<table>
<thead>
<tr>
<th>Level</th>
<th>Examines...</th>
<th>Uncovers...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morphological</td>
<td>Words and word forms</td>
<td>Terms contained in documents</td>
</tr>
<tr>
<td>Syntactic</td>
<td>Sentence structure</td>
<td>Relationships between terms</td>
</tr>
<tr>
<td>Semantic</td>
<td>Meaning of words and sentences</td>
<td>Concepts and relationships</td>
</tr>
<tr>
<td>Pragmatic</td>
<td>Context</td>
<td>Ambiguity of meaning</td>
</tr>
<tr>
<td>Statistical</td>
<td>Co-occurrence of terms, nearness</td>
<td>Strength of relationships among concepts</td>
</tr>
</tbody>
</table>

Linguistics-based solutions analyze texts at all five different levels. In this way, linguistics-based text mining solutions can discover the meaning of a collection of texts and extract concepts from it.
Similarly, a statistics-based solution might incorporate a rule stating that if a text is about “business leadership” it will likely also use the word “manager.” When both of these terms occur, the statistics-based system would identify them as associated concepts. However, for the application of such rules to be effective, the scanned texts must be long enough to provide a reasonable sample. Even then, there is no guarantee that the expected ways of phrasing the concept will occur in the text. A researcher might overlook an interesting article on business leadership, for instance, if the author used the term “executive” but not “manager.” (For a glossary of selected text mining terms, see Appendix A.)

A linguistic approach, however, equates terms like “executive,” “manager,” and even “mgr,” if they are used in similar contexts. In addition, linguistics-based solutions can interpret the tone of text: a customer comment about “new phone” may be distinct from a related concept, “new phone ASAP,” which implies some urgency. The understanding of language cuts through the ambiguity of text, making linguistics-based text mining, by definition, the most accurate possible approach.

Text Mining Technologies from SPSS

The LexiQuest solution from SPSS Inc. is a linguistics-based text mining offering.

- LexiQuest Mine™ is a discovery tool that creates a graphical “map” of concepts and their relationships in a body of textual documents, and can be used in building taxonomies
- LexiQuest Categorize™ automates the process of placing documents or portions of documents in a taxonomy, defined as a hierarchy of categories

A complementary offering, Text Mining for Clementine®, applies classification, clustering, and other predictive techniques in order to identify patterns and combine concepts with existing structural data.

LexiQuest Mine, LexiQuest Categorize, and Text Mining for Clementine utilize a core linguistic extraction component or extractor. The extractor is designed to identify candidate terms, whether these are single words or groups of words, that are relevant to a given domain or topic. The extractor accomplishes this task without the need for any user intervention, although users can add domain knowledge to the extraction process to fine-tune results.

LexiQuest Mine and LexiQuest Categorize are available in English, French, Spanish, Italian, German and Dutch. Text Mining for Clementine is available in the same languages and in Japanese.
There are six major steps in the extraction process:
1. Document conversion and language identification
2. The identification of candidate terms
3. The identification of equivalence classes among candidate terms and the integration of synonyms
4. Type assignment
5. Indexing, using a representative term for each equivalence class
6. Pattern matching and events extraction

**Document conversion and language identification**
In this first step, the SPSS text mining software converts documents to a format that can be used for further analysis. Using built-in filters, the software converts common file types—HTML, PDF, ASCII, XML, and the major Microsoft® formats such as .doc, .xls, and .ppt—to a uniform format. Text from databases and other ODBC-compliant sources can also be converted. The tags of tag-based formats can be used to specify the text that is to be extracted. For example, the tag `<body>` in an HTML document can be used to limit extraction to the text within the body of the document.

The user can also specify the portion of the document to be used as the unit of analysis. That is, the text is broken into sections—the entire document, the paragraph, or the sentence—to enable the analysis of the proximity of words. The relationship between two words depends on the context. You might associate “poor quality” with “Brand X” if the two concepts appear in the same sentence in a document, but not if they appear pages apart. Depending on the type of data to be analyzed, determining the level of analysis is important. For example, short text in call center notes fields may require a sentence-level choice, while short newswire feeds could be analyzed at the document level.

Language identification is part of this first, data-preparation step. The LexiQuest extractor provides a language identifier that can recognize more than 80 languages in different formats. Language identification is done using n-grams, which are combinations of words or characters of a certain length (n). About 400 n-grams are used to identify each language. For example, the following are a subset of tri-grams used for recognizing French:

\{'ble', 'omm', 'b\a\b', 'bd\b', 'm\a', 'le\b', 'du\b', 'nt\b', 'ma\b', 'bet', 'te\b', 'bd\d', 'les', 'ur\b', 'ux\b', 'une', 'br\e', 'iod', 'pou', 'brp', 'ui\b', 'blp', 'bil', 'ait', 'ipa', 'pré', 'bc\e', 'it\e', 'ire', 'ée\b', 'com', 'par', 'ef\b', 'od\b', 'au\b', 'iqu', 'ref', 'b\e\t', 'oit', 'ipa', 'our', 'tio', 'air', 'eur', 'bd\d', 'es\b', 'bav', 'ns\b', 'tai'

The extractor supports English, French, Spanish, Dutch, German, Italian, and Japanese (although the extraction of Japanese concepts uses a different process not described in this document). Any documents in non-supported languages are excluded from the rest of the process.

Before moving on to the next step, the identification of candidate terms, it is important to understand the role of linguistic resources or dictionaries in linguistic extraction. First, a brief definition:

**Dictionaries are lists of words, relationships, or other information that are used to specify or tune the extraction.**

There are two kinds of dictionaries used by the extractor: compiled dictionaries that cannot be changed (“internal dictionaries”), and dictionaries that can be edited by end users (“external dictionaries”).
Internal dictionaries are core components of the extractor. The SPSS text mining solution includes several types of internal dictionaries:

- A general dictionary—a list of base forms with a part-of-speech (PoS) code—for each language. The parts of speech specified in the general dictionary are noun, verb, adjective, adverb, participle, coordination, determiner, and preposition. In a later step, words matching these parts of speech are considered as components of candidate multi-terms (multiple-word concepts or phrases).

- LexiQuest Packs—vertical- or application-specific dictionaries provided by SPSS for different domains. Currently, the Medical Subject Headings or MeSH® dictionary, the National Library of Medicine’s controlled vocabulary thesaurus for medical terms, is available as a LexiQuest Pack.

- Lists of proper names, also known as named entity dictionaries, used to assign extracted terms to the following term types: organizations, people, locations, or products. If a LexiQuest Pack such as the MeSH is used, then additional types may be added to this list.

External dictionaries are not required in order to use these solutions—the extractor dictionaries and the extraction process generally can provide good results without user intervention. However, extraction can be enhanced through user-defined dictionaries. There are several kinds of external dictionaries:

- Extraction dictionaries, which force concepts to be either excluded or included in the concept database
- Synonym dictionaries, which identify synonyms based on similar meaning or inflected form (plural, conjugation, spelling) in order to produce concepts with a higher significance
- Type dictionaries, which assign a particular category type to a word
- Keyword dictionaries, which identify products, organizations, names, terms, and locations by verifying the presence of words or patterns
- The global dictionary, which overrides type and keyword dictionaries to reconcile ambiguities between these dictionaries for specific words

Identification of candidate terms
Once the language of the text has been determined and any documents are removed from the process, the next step in the process begins: the identification of candidate terms. Candidate terms are words or groups of words that are used to identify concepts in the text.

Single words that are not in the general dictionary are considered candidate uni-terms. (You can think of the general dictionary, an internal dictionary discussed earlier, as a list of all the terms that are likely to be uninteresting or ambiguous, linguistically, as standalone concepts. These terms are excluded from extraction at this stage in the process. However, they are not entirely uninteresting, as they are important for determining parts of speech. So they may be included as candidate multi-terms.)

Candidate multi-terms contain one or more words, also called components. These multi-terms are identified using hard-coded or dynamic part-of-speech pattern extractors. For example, the multi-term “sports car,” which follows the English language “adjective noun” part-of-speech pattern, has two components. The multi-term “fast sports car,” which follows the “adjective adjective noun” part-of-speech pattern, has three components. There are typically about 30 patterns per language; the maximum pattern size is about six components, depending on the language.

Non-linguistic entities, such as phone numbers, U.S. Social Security numbers, dates, time, currency, street addresses, and e-mail addresses are extracted in the latest versions of SPSS’ text mining products. A set of rules known as “regular expressions” are used 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 56 69 1805 would be extracted and typed as a French phone number.

Finally, a special algorithm is used to handle upper-case letter-strings, such as job titles, so that these special patterns can be extracted.
Comparison of extracted candidates to identify equivalence classes and integration of synonyms

After candidate uni-terms and multi-terms are identified, the software uses a set of algorithms to compare them and identify equivalence classes. An equivalence class is a base form of a phrase, or a single form of two variants of the same phrase. The purpose of assigning phrases to equivalence classes is to ensure that, for example, “cancer of the thyroid” and “thyroid cancer” are not treated as separate concepts. (See Appendix B for a list of the algorithms applied for this purpose.) To determine which concept to use for the equivalence class, that is, whether “cancer of the thyroid” or “thyroid cancer” is used as the lead term, the extractor component applies the following rules, in order:
1. User-specified
2. The most frequent form in the full body of text
3. The shortest form in the full body of text (which usually corresponds to the base form)

Assigning of type

Next, category types are assigned to extracted concepts. Organizations are typed first, then particular persons, products, and locations. Both internal and external dictionaries are used in this step; so are keywords, global changes, and user-defined dictionaries, if available. Non-linguistic entities are covered by the “pattern matching” step described below. Additional categories can be defined by the user.

Indexing using a representative term for each equivalence class

The entire document collection is re-indexed by establishing a pointer between a text position and the representative term for each equivalence class. This assumes that all the inflected form instances of a candidate term are indexed as a candidate base form. The global frequency is calculated for each base form. A special option can also provide local, or within-document, frequency for candidate terms.

Pattern-matching for events extraction

SPSS text mining products can discover not only named entities but also relationships among them. Several algorithms, available in LexiQuest Mine and in Text Mining for Clementine, provide different syntactical points of view on the proximity of tagged entities at both the document and paragraph level.
A final step provides additional value. The latest versions of SPSS’ text mining products provide a pattern-matching technology that enables users to define relationships between entities in text processed by LexiQuest Mine. These patterns can be used to describe anything that is deemed interesting by the user: facts, events, level of customer satisfaction, etc. By combining typed terms (i.e. persons, organizations, genes, etc), linguistic dependencies, literal strings, and Boolean operators, complex patterns can be matched, and output provided in a user-defined format. (for examples of pattern-matching, see page 13.)

Requirements for a Text Mining Solution

When choosing a text mining solution, it is important to set guidelines that ensure that the solution chosen will meet your organization’s needs. Here is a brief discussion of four essential requirements for an effective linguistics-based text mining solution.

- Scalability
- Accuracy
- Customizability
- Discovery-orientation

Scalability

In a typical day, your organization’s e-mail systems may generate thousands of messages. Your call center, customer service staff, or Web site may touch hundreds of customers. A host of new Web pages may be generated. Dozens of documents, spreadsheets, and survey results may be posted to various intranets. News about competitors and economic indicators may be collected from multiple sites. And you may even file a few patent applications. So it’s important that any text mining solution you choose be able to process information in all formats and scan vast amounts of information efficiently. LexiQuest Mine, LexiQuest Categorize and Text Mining for Clementine can index one gigabyte of text per hour—approximately 250,000 pages.

- A European manufacturer uses LexiQuest Mine to scan a patents database and the Web sites of its competitors—approximately two million pages per day
- A content aggregator uses LexiQuest Mine to navigate more than 500,000 documents daily, reducing by 80% their internal costs for documentation management and delivering information to a growing number of satisfied customers
- A pharmaceutical company uses LexiQuest Mine to extract gene relationships from the entire Medline database—11 million abstracts of medical research—using a grid computing architecture

Accuracy

The challenge in analyzing text is to separate important, useful information from the masses of information that are only marginally useful, or even irrelevant.

Search engine technologies, as mentioned previously, are a “top-down” approach. When using search engines, end users must know how to structure queries to arrive at the desired information. Often, they can’t. Taxonomies, or structured categories, are another “top-down” approach. They can help users locate information in document repositories or on corporate intranets—as long as the users and the taxonomy creators think alike. Often, they don’t.

The Predictive Text Analytics™ solution from SPSS is fast. In an hour, it can scan 250,000 pages of text. It would take more than 4,000 people to scan the same amount of text in that time.
Examples of Pattern Matching

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,” the pattern matcher would match:

\[
\text{pattern}(0306')
\]
\[
\text{name} = 0306'\_\text{positive\_opinion}
\]
\[
\text{value} = \$mExtract \@(0,2) \{\text{is|are|was|has been|would|could|to} @(0,1) (a|rather|quite|pretty|very)? \$vPos \{\$SEP\}? \$vPos \text{and} \$vPos
\]
\[
\text{output} = \$1\ \$6\ \text{(positive\_opinion)}\n\$1\ \$8\ \text{(positive\_opinion)}\n\$1\ \$10\ \text{(positive\_opinion)}
\]

This leads to the understanding that:

- support services helpful \(\text{(positive\_opinion)}\)
- support services friendly \(\text{(positive\_opinion)}\)
- support services courteous \(\text{(positive\_opinion)}\)

Genomics

From the sentence, “Exogenous IL-10 inhibits NF-kappaB in monocytes,” the pattern matcher would match:

\[
\text{pattern}(300)
\]
\[
\text{name} = (300)\_G\_AV\_G\_1
\]
\[
\text{value} = (\$VarGene \$Verb \text{(the)?} \$VarGene)
\]
\[
\text{output} = \text{interleukin 10 inhibits nf-kappa b}
\]

This leads to the understanding that:

Interleukin 10 inhibits nf-kappa b.

Competitive intelligence:

From the sentence, “Atmos Energy Corporation has acquired Mississippi Valley Gas,” the pattern matcher would match:

\[
\text{pattern}(303)
\]
\[
\text{name} = 303
\]
\[
\text{value} = \$vOrg \@(0,1) \$vSupport \$vAction \text{of} \@(0,2) \$vOrg
\]
\[
\text{output} = \$1\ \$7\ \text{($4 completed)}
\]

This leads to the understanding that:

\(<\text{atmos energy corporation}\text{ completes} \text{acquisition of } \text{mississippi valley gas}\>

Sets of patterns for different domains, or tailored to your data, can be provided with SPSS LexiQuest Packs or through special consulting engagements.
Text mining, by contrast, is a “bottom up” approach. It does not require users to know particular search terms, or to think like the taxonomy creator. Text mining reveals the concepts and themes contained in the body of documents, and maps the relationships between them. This leads to a higher degree of accuracy in both precision, a measure of the relevance of the information received, and recall, a measure of the completeness of the information received.

- In an application analyzing patent applications, LexiQuest Categorize showed close to 90 percent accuracy in precision and recall
- In an application analyzing news stories, LexiQuest Mine showed between 87 and 95 percent accuracy in precision and recall

**Customizability**

Because textual information is used for such a wide range of purposes, it is important that users be able to customize a text mining solution to suit their needs. A solution must recognize numerous document formats and languages, and interpret the vocabulary, grammatical structures, and common expressions of each language.

In addition to being available in a number of languages, SPSS text mining solutions easily accommodate the creation of customized dictionaries for industry-specific terms and for names of people, places, and organizations that are of special interest to a user organization. As mentioned previously, the LexiQuest solution includes the MeSH listing of thousands of life sciences terms, and enables organizations to add terms like product names, office locations, and competitor names, to an external dictionary.

**Discovery-orientation**

The chief benefit of linguistics-based text mining solutions is that, unlike any other approach to text, they are designed for information discovery. This means exposing information or relationships that the user was not aware existed—for example, a connection between the lack of a certain nutrient and a certain medical condition, the application of an established methodology to a new field of study, or the formation of a new business alliance.

In addition, LexiQuest Mine provides additional discovery tools to expose different types of relationships between concepts. Users can select the appropriate algorithm, depending upon their needs:

- The “organize” algorithm describes how often words or terms “co-occur” in documents. It looks for regularly occurring relationships and is used to build taxonomies.
- The “discover” algorithm describes how often terms occur together and separately in documents. It enables users to uncover information in large amounts of textual information.
- The “track” algorithm is used to uncover isolated information in large amounts of text. It supports the discovery of weak signals and emerging trends.

Text mining gains even greater predictive power when paired with data mining. All the statistical techniques available through data mining can be applied to concepts, once they have been extracted from textual data.

For example, for each item of text, such as a patent application or a published article, Text Mining for Clementine first creates an index of concepts and then defines its type. Next, it identifies how often each type of concept occurs. This information is analyzed using techniques such as clustering, classification, and predictive modeling. Findings can be integrated with other data mining analyses to identify trends and anticipate future developments.

In this way, organizations can incorporate new knowledge quickly, and make informed decisions about research projects and strategic initiatives.
As discussed in this paper, the criteria to use in evaluating a text mining solution are its scalability, accuracy, customizability, and discover-orientation. The LexiQuest text mining solution from SPSS has been described with reference to these criteria. But, as with the challenge posed by other types of information, the challenge of text requires not only an analytic solution but also a predictive solution—a solution that enables organizations to anticipate and plan for future events.

Predictive Text Analytics from SPSS is an example of such a solution. It combines advanced, linguistics-based text mining with robust data mining capabilities, and enables organizations to develop predictive models from textual data through classification, clustering, and other statistical techniques.

What once seemed unimaginable—gaining predictive insight from large amounts of text—is now achievable, through Predictive Text Analytics.
Notes and Additional Resources

1 Joan Hamilton. "Like it or not, You've Got Email." Business Week, October 4, 1999.


3 American Teleservices Association and Direct Marketing Association.

4 This function is carried out by the LexiQuest TextCat, which is an implementation of the text categorization algorithm presented in Cavnar and Trenkle (1994).

Books


Articles


http://www.intelligententerprise.com/000908/feat1.shtml


http://www.intelligententerprise.com/000908/feat1.shtml


Other resources
Association for Computational Linguistics
http://www1.cs.columbia.edu/~acl/home.html

A listing of other groups conducting research in computational linguistics and natural language processing
http://dmoz.org/Computers/Artificial_Intelligence/Natural_Language/Research_Groups/
## Appendix A: Text Mining Terms

<table>
<thead>
<tr>
<th>TERM</th>
<th>EXPLANATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boolean logic/queries</td>
<td>A term from mathematical logic to indicate propositions linked by the three fundamental logical operations and, or, and not</td>
</tr>
<tr>
<td>Candidate term</td>
<td>A term representing an equivalence class and retained for purposes of cross-indexation</td>
</tr>
<tr>
<td>Categorization</td>
<td>The process of associating a document with one or more subject categories</td>
</tr>
<tr>
<td>Category</td>
<td>Any of several fundamental and distinct classes to which entities or concepts belong</td>
</tr>
<tr>
<td>Classification</td>
<td>The grouping of a set of entities sharing certain formal or external properties</td>
</tr>
<tr>
<td>Clustering</td>
<td>The process of grouping items such as documents on the basis of similarity. The goal is to divide a data set so that similar records are in the same group, and so that groups are as different from each other as possible.</td>
</tr>
<tr>
<td>Computational linguistics</td>
<td>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.</td>
</tr>
<tr>
<td>Concept</td>
<td>An abstract or generic idea generalized from particular instances</td>
</tr>
<tr>
<td>Concept class</td>
<td>A group of similar concepts that is distinct from other groups</td>
</tr>
<tr>
<td>Directory</td>
<td>An organized set of links that provide context to users on the relationship between items in a list, usually moving from the general to the more specific</td>
</tr>
<tr>
<td>Equivalence class</td>
<td>A group of inflected terms represented by one form. This form, retained for indexation, is called the candidate term. Generally, the most frequently found form of a term, or the form explicitly defined by the user.</td>
</tr>
<tr>
<td>Event extraction</td>
<td>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, a terrorist attack, a disease outbreak, a merger or acquisition, etc.</td>
</tr>
<tr>
<td>Fuzzy logic</td>
<td>A term derived from mathematics, and referring to the indeterminacy involved in the analysis of a linguistic unit or pattern</td>
</tr>
<tr>
<td>Indexing</td>
<td>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</td>
</tr>
<tr>
<td>Information retrieval</td>
<td>The process of finding documents based on key words, metadata, or other knowledge about their contents</td>
</tr>
<tr>
<td>Key words</td>
<td>The most important and discriminating words in a document set</td>
</tr>
<tr>
<td>Lift</td>
<td>A measure of increased accuracy of predictive models, as compared to a random baseline. The formula for lift is (the number of accurate hits/the total number of hits) x 100%, where “hits” is the value being predicted.</td>
</tr>
<tr>
<td>Linguistics</td>
<td>The study of the general and universal properties of language</td>
</tr>
<tr>
<td>Morphology</td>
<td>The branch of grammar that studies the structure or forms of words</td>
</tr>
<tr>
<td>Natural language processing</td>
<td>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.</td>
</tr>
<tr>
<td>TERM</td>
<td>EXPLANATION</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Pattern matching</td>
<td>The process of matching the type, form, or value of an argument against the formal argument in a definition</td>
</tr>
<tr>
<td>Polysemy</td>
<td>The condition in text of the same term having multiple meanings</td>
</tr>
<tr>
<td>Precision</td>
<td>The measure of how well information retrieval systems select documents that are relevant to a query</td>
</tr>
<tr>
<td>Recall</td>
<td>The measure of how well information retrieval systems find all the documents that are relevant to a query</td>
</tr>
<tr>
<td>Relevance</td>
<td>A measure of the success of an information system to deliver material that satisfies the needs of the user</td>
</tr>
<tr>
<td>Semantics</td>
<td>A major branch of linguistics devoted to the study of meaning in language</td>
</tr>
<tr>
<td>Statistics</td>
<td>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.</td>
</tr>
<tr>
<td>Stop word</td>
<td>Words that are ignored in full-text indexing because they occur so often that they provide no discriminating information</td>
</tr>
<tr>
<td>Synonym</td>
<td>The condition in text of having several terms with the same meaning</td>
</tr>
<tr>
<td>Syntax</td>
<td>The branch of grammar that deals with the rules governing the combination of words in sentences</td>
</tr>
<tr>
<td>Taxonomy</td>
<td>The organization of a particular set of information for a particular purpose. While in biology, species are placed in a single location within the hierarchy, in informational taxonomies, items can fit into several taxonomic categories.</td>
</tr>
<tr>
<td>Term</td>
<td>A word or expression that has a precise meaning in some uses, or is specific to a science, art, profession, or subject</td>
</tr>
<tr>
<td>Terminology extraction</td>
<td>The process of finding terms and concepts in text and establishing the types of relationships among them</td>
</tr>
<tr>
<td>Text mining</td>
<td>The process of automatically extracting information from large collections of documents</td>
</tr>
<tr>
<td>Thesaurus</td>
<td>A non-hierarchical set of related terms describing the contents of a set of documents</td>
</tr>
<tr>
<td>Weak signal</td>
<td>A signal that is difficult to detect among other signals and information. Often, an idea, trend, or relationship that is new and surprising from the signal receiver’s vantage point, and which may pose a threat or an opportunity to an organization.</td>
</tr>
</tbody>
</table>
Appendix B: Algorithms Used for Assigning Equivalence Classes

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

Inflection

\[ \text{vasopeptidase inhibitors} = \text{vasopeptidase inhibitor} \]

Synonymy

**Full-Form:** an entire extraction is equivalent to another

\[ \text{familial hyperchylomicronemia} = \text{familial lipoprotein lipase deficiency} \]

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

\[ \text{colour blindness} = \text{color blindness} \]

Omission of keywords

\[ \text{ziff-davis inc} = \text{ziff davis} \]

Geographic variant

\[ \text{tumour} = \text{tumor} \]

Lexical variant

\[ \text{geographical markets} = \text{geographic markets} \]

Omission of/variation in function words

\[ \text{ulceration of the mucosa} = \text{ulceration of mucosa} \]
\[ \text{éclipses du soleil} = \text{éclipse de soleil} \]

Variants in separators: Separators may be space, hyphen, agglutination, apostrophe’s (or apostrophe), or dot

\[ \text{zollinger-ellison syndrome} = \text{zollinger ellison syndrome} \]
\[ \text{health care} = \text{healthcare} \]
\[ \text{web-tv} = \text{web tv} \]
\[ \text{webtv} = \text{web tv} \]
\[ \text{alzheimer disease} = \text{alzheimer’s disease} \]

Inversion of components

\[ \text{generalized mytonia of Becker} = \text{Becker’s generalized mytonia} \]
\[ \text{cancer of the thyroid} = \text{thyroid cancer} \]
\[ \text{zeste râpé d’un citron} = \text{zeste de citron râpé} \]

Accented/non-accented characters. This phenomenon may be very frequent in languages such as French, Spanish, Italian, or Dutch.

\[ \text{saô Paulo} = \text{sao Paulo} \]
\[ \text{evguêni primakov} = \text{evgueni primakov} \]
\[ \text{évènements du kosovo} = \text{événements du kosovo} \]

Lower-case/upper-case characters

\[ \text{apolipoprotein A} = \text{apolipoprotein a} \]

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.

\[ \text{lipstick} = \text{cosmetics} \]
\[ \text{eyebush} = \text{cosmetics} \]

Spell checking/fuzzy matching, based on omission of vowels or double consonants, or other algorithms:

\[ \text{technical support} = \text{technical support} \]
\[ \text{technical support} = \text{technical support} \]