About SPSS Modeler Text Analytics

IBM SPSS Modeler Text Analytics offers powerful text analytic capabilities, which use advanced linguistic technologies and Natural Language Processing (NLP) to rapidly process a large variety of unstructured text data and, from this text, extract and organize the key concepts. Furthermore, IBM SPSS Modeler Text Analytics can group these concepts into categories.

Around 80% of data held within an organization is in the form of text documents—for example, reports, Web pages, e-mails, and call center notes. Text is a key factor in enabling an organization to gain a better understanding of their customers' behavior. A system that incorporates NLP can intelligently extract concepts, including compound phrases. Moreover, knowledge of the underlying language allows classification of terms into related groups, such as products, organizations, or people, using meaning and context. As a result, you can quickly determine the relevance of the information to your needs. These extracted concepts and categories can be combined with existing structured data, such as demographics, and applied to modeling in IBM SPSS Modeler's full suite of data mining tools to yield better and more-focused decisions.

Linguistic systems are knowledge sensitive—the more information contained in their dictionaries, the higher the quality of the results. IBM SPSS Modeler Text Analytics is delivered with a set of linguistic resources, such as dictionaries for terms and synonyms, libraries, and templates. This product further allows you to develop and refine these linguistic resources to your context. Fine-tuning of the linguistic resources is often an iterative process and is necessary for accurate concept retrieval and categorization. Custom templates, libraries, and dictionaries for specific domains, such as CRM and genomics, are also included.

Deployment. You can deploy text mining streams using the IBM SPSS Modeler Solution Publisher for real-time scoring of unstructured data. The ability to deploy these streams ensures successful, closed-loop text mining implementations. For example, your organization can now analyze scratch-pad notes from inbound or outbound callers by applying your predictive models to increase the accuracy of your marketing message in real time.

Note: To run IBM SPSS Modeler Text Analytics with IBM SPSS Modeler Solution Publisher, add the directory <install_directory>/ext/bin/spss.TMWBServer to the $LD_LIBRARY_PATH environment variable.

Automated translation of supported languages. IBM SPSS Modeler Text Analytics, in conjunction with SDL's Software as a Service (SaaS), enables you to translate text from a list of supported languages, including Arabic, Chinese, and Persian, into English. You can then perform your text analysis on translated text and deploy these results to people who could not have understood the contents of the source languages. Since the text mining results are automatically linked back to the corresponding foreign-language text, your organization can then focus the much-needed native speaker resources on only the most significant results of the analysis. SDL offers automatic language translation using statistical translation algorithms that resulted from 20 person-years of advanced translation research.

About Text Mining

Today an increasing amount of information is being held in unstructured and semistructured formats, such as customer e-mails, call center notes, open-ended survey responses, news feeds, Web forms, etc. This abundance of information poses a problem to many organizations that ask themselves, "How can we collect, explore, and leverage this information?"

Text mining is the process of analyzing collections of textual materials in order to capture key concepts and themes and uncover hidden relationships and trends without requiring that you know the precise words or terms that authors have used to express those concepts. Although they are quite different, text mining is sometimes confused with information retrieval. While the accurate retrieval and storage of information is an enormous challenge, the extraction and management of quality content, terminology, and relationships contained within the information are crucial and critical processes.

Text Mining and Data Mining

For each article of text, linguistic-based text mining returns an index of concepts, as well as information about those concepts. This distilled, structured information can be combined with other data sources to address questions such as:

- Which concepts occur together?
- What else are they linked to?
- What higher level categories can be made from extracted information?
- What do the concepts or categories predict?
- How do the concepts or categories predict behavior?

Combining text mining with data mining offers greater insight than is available from either structured or unstructured data alone. This process typically includes the following steps:

1. **Identify the text to be mined.** Prepare the text for mining. If the text exists in multiple files, save the files to a single location. For databases, determine the field containing the text.
2. **Mine the text and extract structured data.** Apply the text mining algorithms to the source text.
3. **Build concept and category models.** Identify the key concepts and/or create categories. The number of concepts returned from the unstructured data is typically very large. Identify the best concepts and categories for scoring.
4. **Analyze the structured data.** Employ traditional data mining techniques, such as clustering, classification, and predictive modeling, to discover relationships between the concepts. Merge the extracted concepts with other structured data to predict future behavior based on the concepts.

Text Analysis and Categorization

Text analysis, a form of qualitative analysis, is the extraction of useful information from text so that the key ideas or concepts contained within this text can be grouped into an appropriate number of categories. Text analysis can be performed on all types and lengths of text, although the approach to the analysis will vary somewhat.

Shorter records or documents are most easily categorized, since they are not as complex and usually contain fewer ambiguous words and responses. For example, with short, open-ended survey questions, if we ask people to name their three favorite vacation activities, we might expect to see many short answers, such as *going to the beach*, *visiting national parks*, or *doing nothing*. Longer, open-ended responses, on the other hand, can be quite complex and very lengthy, especially if respondents are educated, motivated, and have enough time to complete a questionnaire. If we ask people to tell us about their political beliefs in a survey or have a blog feed about politics, we might expect some lengthy comments

about all sorts of issues and positions.

The ability to extract key concepts and create insightful categories from these longer text sources in a very short period of time is a key advantage of using IBM SPSS Modeler Text Analytics. This advantage is obtained through the combination of automated linguistic and statistical techniques to yield the most reliable results for each stage of the text analysis process.

Linguistic Processing and NLP

The primary problem with the management of all of this unstructured text data is that there are no standard rules for writing text so that a computer can understand it. The language, and consequently the meaning, varies for every document and every piece of text. The only way to accurately retrieve and organize such unstructured data is to analyze the language and thus uncover its meaning. There are several different automated approaches to the extraction of concepts from unstructured information. These approaches can be broken down into two kinds, linguistic and nonlinguistic.

Some organizations have tried to employ automated nonlinguistic solutions based on statistics and neural networks. Using computer technology, these solutions can scan and categorize 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 words occur and calculate their statistical proximity to related concepts. They produce many irrelevant results, or noise, and miss results they should have found, referred to as silence.

To compensate for their limited accuracy, some solutions incorporate complex nonlinguistic rules that help to distinguish between relevant and irrelevant results. This is referred to as *rule-based text mining*.

Linguistics-based text mining, on the other hand, applies the principles of natural language processing (NLP)—the computer-assisted analysis of human languages—to the analysis of words, phrases, and syntax, or structure, of text. A system that incorporates NLP can intelligently extract concepts, including compound phrases. Moreover, knowledge of the underlying language allows classification of concepts into related groups, such as products, organizations, or people, using meaning and context.

Linguistics-based text mining finds meaning in text much as people do—by recognizing a variety of word forms as having similar meanings and by analyzing sentence structure to provide a framework for understanding the text. This approach 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.

To illustrate the difference between statistics-based and linguistics-based approaches during the extraction process with all language texts except Japanese, consider how each would respond to a query about reproduction of documents. Both statistics-based and linguistics-based solutions would have to expand the word reproduction to include synonyms, such as 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. The understanding of language cuts through the ambiguity of text, making linguistics-based text mining, by definition, the more reliable approach.

The use of linguistic-based techniques through the Sentiment analyzer makes it possible to extract more meaningful expressions. The analysis and capture of emotions cuts through the ambiguity of text, and makes linguistics-based text mining, by definition, the more reliable approach.

Linguistic systems are knowledge sensitive—the more information contained in their dictionaries, the higher the quality of the results. Modification of the dictionary content, such as synonym definitions, can simplify the resulting information. This is often an iterative process and is necessary for accurate concept retrieval. NLP is a core element of IBM SPSS Modeler Text Analytics.

IBM SPSS Modeler Text Analytics Nodes

Along with the many standard nodes delivered with IBM SPSS Modeler, you can also work with text mining nodes to incorporate the power of text analysis into your streams. IBM SPSS Modeler Text Analytics offers you several text mining nodes to do just that. These nodes are stored in the IBM SPSS Modeler Text Analytics tab of the node palette.

The following nodes are included:

- The **File List source node** generates a list of document names as input to the text mining process. This is useful when the text resides in external documents rather than in a database or other structured file. The node outputs a single field with one record for each document or folder listed, which can be selected as input in a subsequent Text Mining node.
- The **Web Feed source node** makes it possible to read in text from Web feeds, such as blogs or news feeds in RSS or HTML formats, and use this data in the text mining process. The node outputs one or more fields for each record found in the feeds, which can be selected as input in a subsequent Text Mining node.
- The **Text Mining node** uses linguistic methods to extract key concepts from the text, allows you to create categories with these concepts and other data, and offers the ability to identify relationships and associations between concepts based on known patterns (called text link analysis). The node can be used to explore the text data contents or to produce either a concept model or category model. The concepts and categories can be combined with existing structured data, such as demographics, and applied to modeling.
- The **Text Link Analysis node** extracts concepts and also identifies relationships between concepts based on known patterns within the text. Pattern extraction can be used to discover relationships between your concepts, as well as any opinions or qualifiers attached to these concepts. The Text Link Analysis node offers a more direct way to identify and extract patterns from your text and then add the pattern results to the dataset in the stream. But you can also perform TLA using an interactive workbench session in the Text Mining modeling node.
- The **Translate node** can be used to translate text from supported languages, such as Arabic, Chinese, and Persian, into English or other languages for purposes of modeling. This makes it possible to mine documents in double-byte languages that would not otherwise be supported and allows analysts to extract concepts from these documents even if they are unable to speak the language in question. The same functionality can be invoked from any of the text modeling nodes, but use of a separate Translate node makes it possible to cache and reuse a translation in multiple nodes.
- When mining text from external documents, the **Text Mining Output node** can be used to generate an HTML page that contains links to the documents from which concepts were extracted