CHAPTER 1

Introduction to Text Mining and SAS Text Miner

What Is Text Mining?

The purpose of text mining is to help you understand what the text tells you without having to read every word. Text mining applications fall into two areas: *exploring* the textual data for its content, and then *using* the information to improve the existing processes. Both are important, and can be referred to as *descriptive mining* and *predictive mining*.

Descriptive mining involves discovering the themes and concepts that exist in a textual collection. For example, many companies collect customers' comments from sources that include the Web, e-mail, and a call center. Mining the textual comments includes providing detailed information about the terms, phrases, and other entities in the textual collection, clustering the documents into meaningful groups, and reporting the concepts that are discovered in the clusters. The result enables you to better understand the collection.

Predictive mining involves classifying the documents into categories and using the information that is implicit in the text for decision making. You might want to identify the customers who ask standard questions so that they receive an automated answer. Or you might want to predict whether a customer is likely to buy again, or even if you should spend more effort in keeping him or her as a customer. In data mining terminology, this is known as *predictive modeling*. Predictive modeling involves examining past data to predict future results. You might have a data set that contains information about past buying behaviors, along with comments that the customers made. You can then build a predictive model that can be used to score new customers: that is, in the past these customers did this, so if new customers have similar comments, they are likely to do the same thing. For example, if you are a researcher for a pharmaceutical company, you know that hand-coding adverse reactions from doctors' reports in a clinical study is a laborious, error-prone job. Instead, you could train a model by using all your historical textual data, noting which doctors' reports correspond to which adverse reactions. When the model is constructed, processing the textual data can be done automatically by scoring new records that come in. You would just have to examine the "hard-to-classify" examples, and let the computer handle all the rest.

Both of the above aspects of text mining share some of the same requirements. Namely, text documents that human beings can easily understand must first be represented in a form that can be mined. The raw documents need processing before the patterns and relationships that they contain can be discovered. Although the human mind comprehends chapters, paragraphs, and sentences, computers require structured (quantitative or qualitative) data. As a result, an unstructured document must be converted to a structured form before it can be mined.

What Is SAS Text Miner?

SAS Text Miner contains a sophisticated Text Miner node that can be embedded into a SAS Enterprise Miner process flow diagram. The node analyzes text that exists in a SAS data set, that is in an external database through SAS/ACCESS, or as files in a file system. The Text Miner node encompasses the parsing and exploration aspect of text mining, and sets up the data for predictive mining and further exploration using the rest of the Enterprise Miner nodes. This enables you to analyze the new structured information that you have acquired from the text however you want, combining it with other structured data as desired. The node is highly customizable and allows a variety of parsing options. It is possible to parse documents for detailed information about the terms, phrases, and other entities in the collection. You can also cluster the documents into meaningful groups and report the concepts that you discover in the clusters. All of this is done in an environment that enables you to interact with the collection. Sorting, searching, filtering (subsetting), and finding similar terms or documents all enhance the exploration process.

The Text Miner node's extensive parsing capabilities include

- stemming
- automatic recognition of multiple-word terms
- normalization of various entities such as dates, currency, percent, and year
- part-of-speech tagging
- extraction of entities such as organizations, products, social security numbers, time, titles, and more
- support for synonyms.

A secondary tool that Text Miner uses is a SAS macro that is called %tmfilter. This macro accomplishes a text preprocessing step and allows SAS data sets to be created from documents that reside in your file system or on the Web pages. These documents can exist in a number of proprietary formats.

SAS Text Miner is part of Enterprise Miner. Enterprise Miner provides a rich set of data mining tools that facilitate the prediction aspect of text mining. The integration of Text Miner within Enterprise Miner enables the combining of textual data with traditional data-mining variables.

With all of this functionality, SAS Text Miner becomes a very flexible tool that can be used to solve a variety of problems. Below are some examples of tasks that can be accomplished.

- filtering e-mail
- grouping documents by topic into predefined categories
- routing news items
- clustering analysis of research papers in a database
- clustering analysis of survey data
- clustering analysis of customer complaints and comments
- predicting stock market prices from business news announcements
- predicting customer satisfaction from customer comments
- predicting cost, based on call center logs.

The Text Mining Process

Whether you intend to use textual data for descriptive purposes, predictive purposes, or both, the same processing steps take place, as shown in the following table.

Table 1.1 The General Order for Text Mining

Action	Result
File preprocessing	Creates a single SAS data set from your document collection. The SAS data set will be used as input for the Text Mining node. (This is an optional step. Do this if the text is not already in a SAS data set or external database.)
Text parsing	Decomposes textual data and generates a quantitative representation suitable for data mining purposes.
Transformation (dimension reduction)	Transforms the quantitative representation into a compact and informative format.
Document analysis	Performs clustering or classification of the document collection.

Finally, the rules for clustering or predictions can be used to score a new collection of documents at any time.

You might or might not include all of these steps in your analysis, and it might be necessary to try a different combination of text parsing options before you are satisfied with the results.

Tips for Text Mining

Using the Text Miner node to process a very large collection of documents can require a lot of computing time and resources. If you have limited resources, it might be necessary to take one or more of the following actions:

□ use a sample of the document collection.

□ deselecting some options in the Text Miner Settings window, such as stemming and entity extraction, and the search for words that occur in a single document.

□ reduce the number of SVD dimensions or roll-up terms. If you have memory problems when you use the SVD approach, you can roll up a certain number of terms, and drop the remaining terms. If you do that and perform SVD at the same time, only the rolled up terms are used in the calculation of SVD. This way you can reduce the size of the problem.

□ limit parsing to high-information words by deselecting the parts of speech other than nouns, proper nouns, noun groups, and verbs.