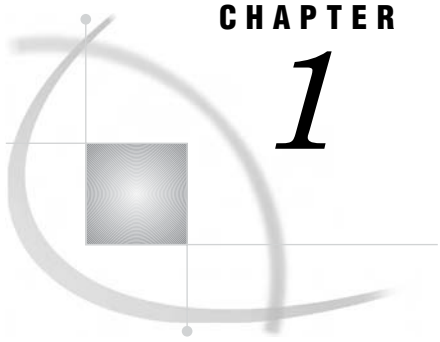

CHAPTER

1

Introduction to Text Mining and SAS Text Miner

What Is Text Mining?

Text mining helps you understand what textual documents tell you without having to read every word. Text mining uncovers the underlying themes or concepts that are contained in large document collections. 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 call centers. 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 textual 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.

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: to analyze new customers based on the data from past customers. 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 create 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 these 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 by the software. 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 into a structured form before it can be mined.

What Is SAS Text Miner?

SAS Text Miner is an add-on for the SAS Enterprise Miner environment. Enterprise Miner provides a rich set of data mining tools that facilitate the prediction aspect of text mining. The integration of SAS Text Miner within SAS Enterprise Miner combines textual data with traditional data mining variables. A Text Miner node can be embedded into a SAS Enterprise Miner process flow diagram. SAS Text Miner supports various sources of textual data: local text files, text as observations in SAS data sets or external databases, and files on the Web. The Text Miner node encompasses the parsing and exploration aspects of text mining and sets up the data for predictive mining and further exploration using other 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 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 multi-word terms
- normalization of various entities such as dates, currencies, percentages, and years
- part-of-speech tagging
- extraction of entities such as organizations, products, Social Security numbers, time, titles, and more
- support for synonyms
- language-specific analysis for English, Danish, Dutch, Finnish, French, German, Italian, Japanese, Korean, Norwegian Bokmal, Portuguese, Simplified Chinese, Spanish, Swedish, and Traditional Chinese.

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 Web pages. These documents can exist in a number of proprietary formats.

With all this functionality, SAS Text Miner becomes a very flexible tool that can solve a variety of problems. Here 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 costs, 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 Table 1.1.

Table 1.1 General Order for Text Mining

Action	Result	Tool
File preprocessing	Creates a single SAS data set from your document collection. The SAS data set is used as input for the Text Miner node and may contain the actual text or paths to the the actual text.	%TMFILTER macro—a SAS macro for extracting text from documents and creating a predefined SAS data set with a text variable
Text parsing	Decomposes textual data and generates a quantitative representation suitable for data mining purposes.	Text Miner node
Transformation (dimension reduction)	Transforms the quantitative representation into a compact and informative format.	Text Miner node
Document analysis	Performs clustering, classification, prediction, or concept linking of the document collection.	Text Miner node and/or Enterprise Miner predictive modeling nodes

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

You might not need to 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.

Accessibility Features of SAS Text Miner 3.1

SAS Text Miner 3.1 includes accessibility and compatibility features that improve usability of the product for users with disabilities. These features are related to accessibility standards for electronic information technology adopted by the U.S. Government under Section 508 of the U.S. Rehabilitation Act of 1973, as amended. SAS Text Miner 3.1 supports Section 508 standards except as noted in the following table.

Section 508 Accessibility Criteria	Support Status	Explanation
When software is designed to run on a system that has a keyboard, product functions shall be executable from a keyboard where the function itself or the result of performing a function can be discerned textually.	Supported with exceptions.	The software supports keyboard equivalents for all user actions with the exceptions noted below: □ The keyboard equivalent for exposing the system menu is not the Windows standard Alt+spacebar. The system menu can be exposed using the following shortcut keys: □ Primary window - Shirt+F10+spacebar □ Secondary window - Shift+F10+down key. □ The Explore action in the data source popup menu cannot be invoked directly from the keyboard, but there is an alternative way to invoke the data source explorer using the **View ▶ Table** menu.
Color coding shall not be used as the only means of conveying information, indicating an action, prompting a response, or distinguishing a visual element.	Supported with exception.	Node run or failure indication relies on color, but there is always the corresponding message in the bottom right panel of the main window.

If you have questions or concerns about the accessibility of SAS products, send e-mail to accessibility@sas.com.