THE POWER TO KNOW.

Getting Started with SAS® Text Miner 13.2

SAS® Documentation

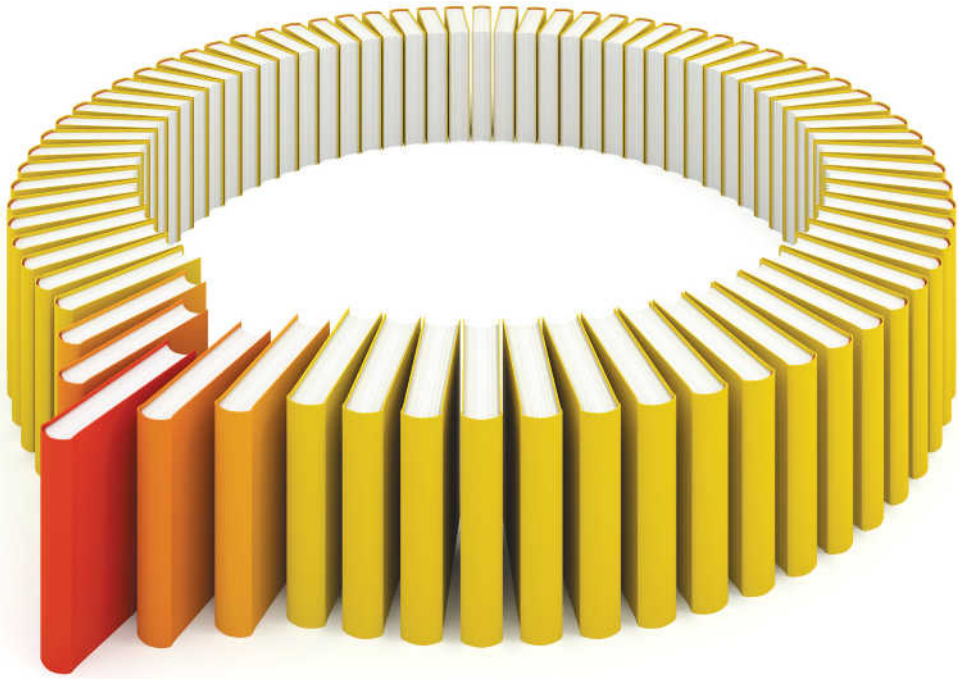
The correct bibliographic citation for this manual is as follows: SAS Institute Inc 2014. *Getting Started with SAS® Text Miner 13.2*. Cary, NC: SAS Institute Inc.

Getting Started with SAS® Text Miner 13.2

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

Electronic book 1, August 2014

SAS® Publishing provides a complete selection of books and electronic products to help customers use SAS software to its fullest potential. For more information about our e-books, e-learning products, CDs, and hard-copy books, visit the SAS Publishing Web site at **support.sas.com/publishing** or call 1-800-727-3228.



Gain Greater Insight into Your SAS® Software with SAS Books.

Discover all that you need on your journey to knowledge and empowerment.

support.sas.com/bookstore
for additional books and resources.

Contents

Overview

Audience

This book is intended for users who are new to SAS Text Miner. The first seven chapters illustrate how you can use SAS Text Miner nodes in the context of a hypothetical text mining analysis. After completing these chapters, you should be able to create projects, process flow diagrams, understand how you can set properties for SAS Text Miner nodes, run them, and explore the results.

Chapters 8 to 13 provide additional examples about the following SAS Text Miner nodes.

- The Text Import Node
- The Text Parsing Node
- The Text Filter Node
- The Text Cluster Node
- The Text Topic Node
- The Text Rule Builder Node
- The Text Profile Node

The final two chapters provide some text mining tips and a quick look at additional features.

See the SAS Text Miner help for more information about the text mining process or any of the SAS Text Miner nodes.

Recommended Reading

- *Many of the concepts and topics that are discussed in additional product documentation for SAS Text Miner 13.2 (http://support.sas.com/documentation/onlinedoc/txtminer) and SAS Enterprise Miner 13.2 (http://support.sas.com/documentation/onlinedoc/miner) might also help you use SAS Text Miner 13.2.*

For a complete list of SAS books, go to support.sas.com/bookstore. If you have questions about which titles you need, please contact a SAS Book Sales Representative:

SAS Books
SAS Campus Drive
Cary, NC 27513-2414
Phone: 1-800-727-3228
Fax: 1-919-677-8166
E-mail: sasbook@sas.com
Web address: support.sas.com/bookstore

Chapter 1

Introduction to Text Mining and SAS Text Miner 13.2

What Is Text Mining?

Text mining uncovers the underlying themes or concepts that are contained in large document collections. Text mining applications have two phases: exploring the textual data for its content and then using discovered 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 contact 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. Results from descriptive mining enable 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. For example, you might want to identify the customers who ask standard questions so that they receive an automated answer. In addition, you might want to predict whether a customer is likely to buy again, or even if you should spend more effort to keep the customer.

Predictive modeling involves examining past data to predict results. Consider that you have a customer data set that contains information about past buying behaviors, along with customer comments. You could build a predictive model that can be used to score new customers—that is, 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 the rest.

Both of these aspects of text mining share some of the same requirements. Namely, textual 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 13.2?

SAS Text Miner is a plug-in for the SAS Enterprise Miner environment. SAS 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. Text mining nodes 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.

SAS Text Miner 13.2 includes the following nodes that you can use in your text mining analysis:

- **Text Import** node
- **Text Parsing** node
- **Text Filter** node
- **Text Topic** node
- **Text Cluster** node
- **Text Rule Builder** node
- **Text Profile** node

For more information about the SAS Text Miner nodes, see the corresponding chapter in this book, or the SAS Text Miner Help.

Together, the Text Miner nodes encompass the parsing and exploration aspects of text mining and the preparation of data for predictive mining and further exploration when you use other SAS Enterprise Miner nodes. You can analyze structured text information, and combine the structured output of the Text Miner nodes with other structured data as desired. The Text Miner nodes are highly customizable and enable you to choose among a variety of options. For example, the **Text Parsing** node enables you to parse documents for detailed information about the terms, phrases, and other entities in the collection. The **Text Cluster** node enables you to cluster documents into meaningful groups and to report concepts that you discover in the clusters. Sorting, searching, filtering (subsetting), and finding similar terms or documents all enhance the exploration process.

SAS Text Miner also enables you to use a SAS macro that is called %TMFILTER. This macro accomplishes a text preprocessing step and enables 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.

SAS Text Miner is a flexible tool that can solve a variety of problems. Here are some examples of tasks that can be accomplished using SAS Text Miner:

- 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 the following table:

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 Parsing** node, and might contain the actual text or paths to the actual text.	**Text Import** node %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 Parsing** node
Transformation (dimension reduction)	Transforms the quantitative representation into a compact and informative format.	**Text Filter** node
Document analysis	Performs classification, prediction, or concept linking of the document collection. Creates clusters, topics, or rules from the data.	**Text Cluster** node **Text Topic** node **Text Rule Builder** node **Text Profile** node SAS Enterprise Miner predictive modeling nodes

Note: The **Text Miner** node is not available from the **Text Mining** tab in SAS Text Miner 13.2. The Text Miner node has now been replaced by the functionality in other SAS Text Miner nodes. You can import diagrams from a previous release of SAS Text Miner that had a **Text Miner** node in the process flow diagram. However, new **Text Miner** nodes can no longer be created, and property values cannot be changed in imported **Text Miner** nodes. For more information, see the Converting

SAS Text Miner Diagrams from a Previous Version topic in the SAS Text Miner Help.

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 options before you are satisfied with the results.

Accessibility Features of SAS Text Miner 13.2

SAS Text Miner 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 supports Section 508 standards except as noted in the following table.

Section 508 Accessibility Criterion	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: (1) Primary window — Shift + F10 + spacebar, or (2) Secondary window — Shift + F10 + down key. The Explore action in the data source pop-up menu cannot be invoked directly from the keyboard. There is an alternative way to invoke the data source explorer using the **View** —> **Explorer** 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. There is also a corresponding pop-up message in a dialog box that indicates node success or failure.

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

Chapter 2

Learning by Example: Using SAS Text Miner 13.2

About the Scenario in This Book

The first seven chapters describe an extended example that is intended to familiarize you with SAS Text Miner. Each topic builds on the previous topic, so you must work through these chapters in sequence. Several key components of the SAS Text Miner process flow diagram are covered. In this step-by-step example, you learn to do basic tasks in SAS Text Miner, such as creating a project and building a process flow diagram. In your diagram, you perform tasks such as accessing data, preparing the data, building multiple predictive models using text variables, and comparing the models. The extended example in this book is designed to be used in conjunction with SAS Text Miner software. The remaining chapters focus on each of the SAS Text Miner nodes, and provide additional information that you might find useful for your text mining analysis.

The Vaccine Adverse Event Reporting System (VAERS) data is publicly available from the U.S. Department of Health and Human Services (HHS). Anyone can download this data in comma-separated value (CSV) format from **http://vaers.hhs.gov**. There are separate CSV files for every year since the U.S. started collecting the data in 1990. This data is collected from myriad sources, but most reports come from vaccine manufacturers and health care providers. Providers are required to report any contraindicated events for a vaccine or any very serious complications. In the context of a vaccine, a contraindication event would be a condition or a factor that increases the risk of using the vaccine.

See the following in the **Getting Started Examples** zip file:

- ReportableEventsTable.pdf for a complete list of reportable events for each vaccine
- VAERS README file for a data dictionary and list of abbreviations used

Note: See "Prerequisites for This Scenario" on page 7 for information about where to download the **Getting Started Examples** zip file.

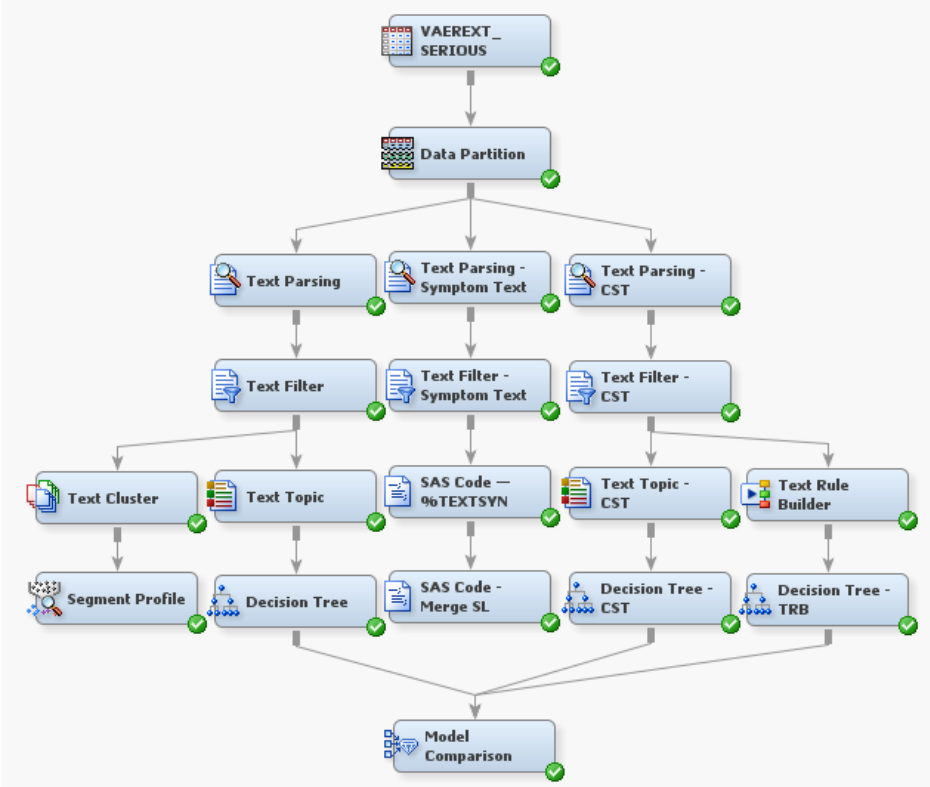
The following figure shows the first 8 columns in the first 10 rows in a table of VAERS data. Included is a unique identifier, the state of residence, and the recipient's age. Additional columns (not in the following figure) include an unstructured text string

SYMPTOM_TEXT that contains the reported problem, specific symptoms, and a symptom counter.

	VAERS_ID	RECVDATE	STATE	AGE_YRS	CAGE_YR	CAGE_MO	SEX	RPT_DATE
1	179605.0	Jan 2, 2002	FL	64.0	64.0	.	F	Dec 26, 2001
2	179606.0	Jan 2, 2002		29.0	29.0	.	F	Dec 26, 2001
3	179612.0	Jan 2, 2002	NJ	40.0	0.0	0.3	F	Dec 23, 2001
4	179613.0	Jan 2, 2002	NY	40.0	1.0	0.6	F	Feb 28, 1998
5	179614.0	Jan 2, 2002	TX	4.0	4.0	.	M	Dec 28, 2001
6	179615.0	Jan 2, 2002	WI	38.0	38.0	.	F	Dec 21, 2001
7	179616.0	Jan 2, 2002	KY	69.0	69.0	.	F	Dec 26, 2001
8	179617.0	Jan 2, 2002	FL	77.0	77.0	.	M	Dec 21, 2001
9	179618.0	Jan 2, 2002		7.0	7.0	.	M	Nov 23, 2001
10	179619.0	Jan 2, 2002		50.0	.	.	F	Dec 20, 2001

As you go through this example, imagine you are a researcher trying to discover what information is contained within this data set. You also want to know how you can use it to better understand the adverse reactions that children and adults are experiencing from their vaccination shots. These adverse reactions might be caused by one or more of the vaccinations that they are given, or they might be induced by an improper procedure from the administering lab (for example, a non-sanitized needle). Some of them will be totally unrelated. For example, perhaps someone happened to get a cold just after receiving a flu vaccine and reported it. You might want to investigate serious reactions that required a hospital stay or caused a lifetime disability or death.

When you are finished with this example, your process flow diagram should resemble the one shown here:



Prerequisites for This Scenario

Before you can perform the tasks in this book, administrators at your site must have installed and configured all necessary components of SAS Text Miner 13.2. You must also perform the following:

1. Download the Example Data for Getting Started with SAS Text Miner 13.2 zip file under the SAS Text Miner 13.2 heading from the following URL:

   `http://support.sas.com/documentation/onlinedoc/txtminer`

2. Unzip this file into any folder in your file system.
3. Create a folder called **`vaersdata`** on your C:\ drive.
4. Copy the following files into **C:\vaersdata**:
   - vaerext.sas7bdat
   - vaer_abbrev.sas7bdat
   - engdict.sas7bdat

   Note: The preceding list of files might or might not be capitalized depending on which environment you are viewing them in.

How to Get Help for SAS Text Miner 13.2

Select **Help** ⇨ **Contents** from the main SAS Enterprise Miner menu bar to get help for SAS Text Miner.

Chapter 3

Setting Up Your Project

About the Tasks That You Will Perform

To set up your project, perform the following main tasks:

1. Create a new project where you will store all your work.
2. Create a library to store your data sources.
3. Create a new diagram in your project that you will use to interact with nodes.
4. Explore and modify the VAERS data.
5. Create the SAS Enterprise Miner data source VAEREXT_SERIOUS.

Create a Project

To create a project:

1. Open SAS Enterprise Miner.
2. Click **New Project** in the SAS Enterprise Miner window.

   The Select SAS Server page appears.
3. Click **Next**.

   The Specify Project Name and Server Directory page appears.
4. Enter a name for the project, such as *Vaccine Adverse Events*, in the **Project Name** field.

5. Enter the path to the location on the server where you want to store data for your project in the **SAS Server Directory** field. Alternatively, browse to a folder to use for your project, or accept the default directory path that appears.

   Note: The project path depends on whether you are running SAS Enterprise Miner as a complete client on your local machine or as a client and server application. If you are running SAS Enterprise Miner as a complete client, your local machine acts as its own server. Your SAS Enterprise Miner projects are stored on your local machine in a location that you specify, such as **C:\EM_Projects**. If you are running SAS Enterprise Miner as a client/server application, all projects are stored on the SAS Enterprise Miner server. If you see a default path in the SAS Server Directory box, you can accept the default project path, or you can specify your own project path.

6. Click **Next**.

   The Register the Project page appears.

7. Click **Next**.

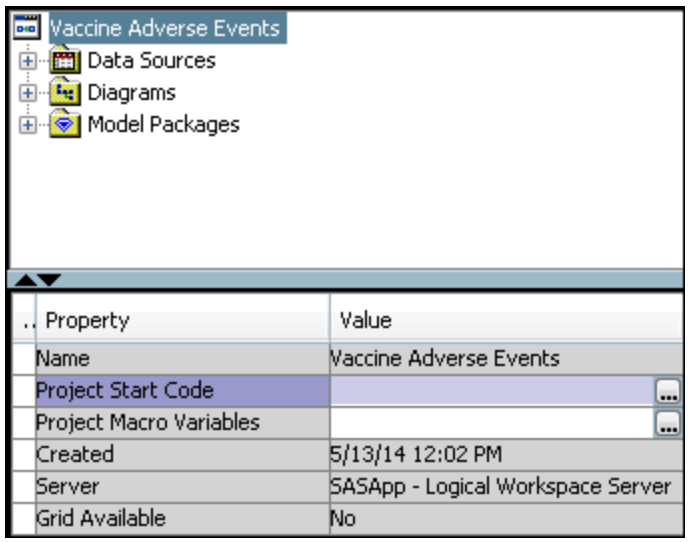
   The New Project Information page appears.

8. Click **Finish** to create your project.

Create a Library

To create a library:

1. Select the project name **Vaccine Adverse Events** to display the project Properties Panel.



2. Click ... for the **Project Start Code** property.

   The Project Start Code dialog box appears.

3. Enter the following code on the **Code** tab:

```
libname mylib "c:\vaersdata";
```

   Note: The location depends on where you have stored the data for this tutorial on your system.

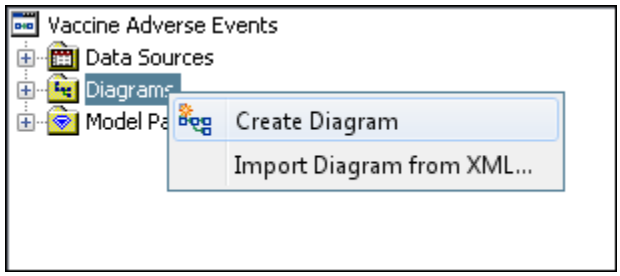
4. Click **Run Now**.
5. Click **OK** to close the Project Start Code dialog box.

Note: An alternate way to create a library is to use the library wizard. To use the library wizard, select **File** ⇨ **New** ⇨ **Library** from the main menu.

Create a Diagram

To create a diagram, complete the following steps:

1. Right-click the Diagrams folder in the Project Panel and select **Create Diagram**.



   The Create New Diagram dialog box appears.
2. Type *VAERS Example* in the **Diagram Name** field.
3. Click **OK**.

   The empty VAERS Example diagram opens in the diagram workspace.

Explore and Modify the VAERS Data

After you create a library, you can view the data before creating a data source to use in SAS Enterprise Miner. For example, consider the situation where you want to create another variable to summarize the values of multiple variables.

Perform the following steps to view available SAS data files and create a new SAS data file that you will use as a data source.

1. Select **View** ⇨ **Explorer** from the main menu.

   The Explorer window appears.
2. Select **Mylib** in the SAS Libraries tree.

   The contents of the Mylib library appear, and include the three files **`Engdict`**, **`Vaerext`**, and **`Vaer_abbrev`**.
3. Double-click **`Vaerext`**.

   The contents of the **`Vaerext`** file appears in a new window.
4. Scroll to the right to view the available variable names that appear as column headings.

   Notice the following variables:

- **DISABLE** — a binary variable that has a value of 'Y' if there was a disability
- **DIED** — a binary variable that has a value of 'Y' if there was a death
- **ER_VISIT** — a binary variable that has a value of 'Y' if there was an emergency room visit
- **HOSPITAL** — a binary variable that has a value of 'Y' if there was a hospitalization

Consider that we want to create a new data set, **vaerext_serious**, that includes a binary variable **serious** that has a value of 'Y' if there was disability, death, emergency room visit, or hospitalization, and a value of 'N' otherwise.

5. Close the **MYLIB.VAEREXT** window and the Explorer window.
6. Select the **Utility** tab, and drag a **SAS Code** node into the diagram workspace
7. Select the **SAS Code** node.
8. Click  for the **Code Editor** property.

   A window appears.
9. Enter the following code on the **Training Code** pane:

```
data mylib.vaerext_serious;
   set mylib.vaerext;
   if DISABLE='Y' or DIED='Y' or ER_VISIT='Y' or HOSPITAL='Y' then serious='Y';
   else serious='N';
run;
```

   This code creates a new SAS file, **vaerext_serious** from the **vaerext** file in the **mylib** library, adds a variable **serious**, and assigns it a value of **Y** or **N**, depending on the value of the **DISABLE**, **DIED**, **ER_VISIT**, and **HOSPITAL** variables.
10. Click  , and then close the code editor window.
11. Right-click **SAS Code** node in the diagram workspace, and select **Run**.
12. Click **Yes** in the Confirmation dialog box.
13. Click **OK** in the Run Status dialog box when the node has finished running.
14. Select **View** ⇨ **Explorer** from the main menu.

    The Explorer window appears.
15. Select **Mylib** in the SAS Libraries tree.

    Notice that the Mylib library now contains a new entry for the file **Vaerext_serious**.

    Note: You might need to refresh your view of the Explorer window to see the file **Vaerext_serious**.
16. Double-click **Vaerext_serious**.

    The contents of the **Vaerext_serious** file appears in a new window.
17. Scroll all the way to the right to see the new column **serious**.
18. Close the **MYLIB.VAEREXT_SERIOUS** window.
19. Close the Explorer window.

Create a Data Source

To create a data source:

1. Right-click the Data Sources folder in the Project Panel and select **Create Data Source**.

   The Data Source wizard appears.

2. Select **SAS Table** in the Source drop-down menu.
3. Click **Next**.

   The Select a SAS Table window appears.

4. Click **Browse**.
5. Click the SAS library named **Mylib** in the library tree.

   The Mylib library folder contents are displayed on the Select a SAS Table dialog box.

   Note: If you do not see SAS data files in the Mylib folder, click **Refresh**.

6. Select the **Vaerext_serious** table.
7. Click **OK**.

   The two-level name **`MYLIB.VAEREXT_SERIOUS`** is displayed in the **Table** field.

8. Click **Next**.

   The Table Information page appears.

9. Click **Next**.

   The Metadata Advisor Options page appears.

10. Select **Advanced**.
11. Click **Next**.

    The Column Metadata page appears.

12. Select the following variable roles by clicking the role value for each variable value and selecting the indicated value from the drop-down list.

    - Set the role for **V_ADMINBY** to **Input**.
    - Set the role for **V_FUNDBY** to **Input**.
    - Set the role for **serious** to **Target**.

13. Click **Next**.

    The Decision Configuration page appears.

14. Click **Next**.

    The Create Sample page appears.

15. Click **Next**.

    The Data Source Attributes page appears.

16. Click **Next**.

    The Summary page appears.

17. Click **Finish**.

    The VAEREXT_SERIOUS table is added to the Data Sources folder in the Project Panel.

Chapter 4

Analyzing the SYMPTOM_TEXT Variable

About the Tasks That You Will Perform

The SYMPTOM_TEXT variable contains the text of an adverse event as it was reported. This chapter explains how you can analyze the SYMPTOM_TEXT variable by performing the following tasks:

1. Identify the VAERS_SERIOUS data source with an **Input Data** node.
2. Partition the input data using the **Data Partition** node.
3. Parse the document collection using the **Text Parsing** node.
4. Reduce the total number of parsed terms using the **Text Filter** node.
5. Cluster documents using the **Text Cluster** node.
6. View the results.
7. Examine data segments using the **Segment Profile** node.

Identify Input Data

To identify input data:

1. Select the **VAEREXT_SERIOUS** data source from the Data Sources folder in the Project Panel.

2. Drag **VAEREXT_SERIOUS** into the diagram workspace to create an **Input Data** node.

Partition Input Data

The **Data Partition** node enables you to partition your input data into one of the following data sets:

- **Train** — used for preliminary model fitting. The analyst attempts to find the best model weights by using this data set.
- **Validation** — used to assess the adequacy of the model in the Model Comparison node. The validation data set is also used for model fine-tuning in the **Decision Tree** model node to create the best subtree.
- **Test** — used to obtain a final, unbiased estimate of the generalization error of the model.

For more information about the **Data Partition** node, see the SAS Enterprise Miner Help.

Perform the following steps to add a **Data Partition** node to the analysis:

1. Select the **Sample** tab on the node toolbar and drag a **Data Partition** node into the diagram workspace.
2. Connect the **VAEREXT_SERIOUS** input data node to the **Data Partition** node.

   Note: To connect one node to another node in the default horizontal view, position the mouse pointer at the right edge of a node. A pencil icon appears. Hold the left mouse button down, and drag the line to the left edge of the node that you want to connect to, and then release the left mouse button. To change your view of connected nodes to a vertical view, right-click in the diagram workspace, and select **Layout** ⇨ **Vertically** in the menu that appears.

3. Select the **Data Partition** node to view its properties.

   Details about the node appear in the Properties Panel.
4. Set the Data Set Allocations properties as follows:
   - Set the **Training** property to **60.0**.
   - Set the **Validation** property to **20.0**.
   - Set the **Test** property to **20.0**.

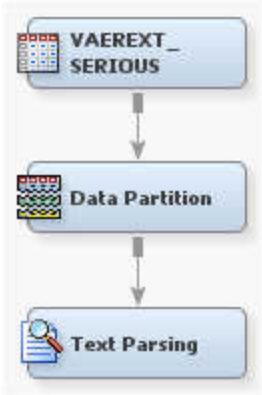
   These data partition settings ensure adequate data when you build prediction models with the VAEREXT_SERIOUS data.

Parse Data

The **Text Parsing** node enables you to parse a document collection in order to quantify information about the terms that are contained therein. You can use the **Text Parsing** node with volumes of textual data such as e-mail messages, news articles, Web pages, research papers, and surveys. For more information about the **Text Parsing** node, see the SAS Text Miner Help.

Perform the following steps to add a **Text Parsing** node to the analysis:

1. Select the **Text Mining** tab on the node toolbar, and drag a **Text Parsing** node into the diagram workspace.
2. Connect the **Data Partition** node to the **Text Parsing** node.



3. Select the **Text Parsing** node.

   The properties for the **Text Parsing** node appear in the Properties Panel.
4. Set the **Different Parts of Speech** property value to `No`.

   For the VAERS data, this setting offers a more compact set of terms.
5. Click the ... for the **Synonyms** property.

   A dialog box appears.
6. Click **Replace Table**.

   The Select a SAS Table dialog box appears.
7. Select **No data set to be specified**.
8. Click **OK** to exit the Select a SAS Table dialog box.
9. Click **Yes** in the Confirmation dialog box.
10. Click **OK** to exit the Synonyms dialog box.
11. Click the ... for the **Ignore Parts of Speech** property.

    The Ignore Parts of Speech dialog box appears.
12. Select the following items, which represent parts of speech:
    - Aux
    - Conj

- Det
- Interj
- Part
- Prep
- Pron
- Num

Note: Hold down the CTRL key to select more than one.

Any terms with the parts of speech that you select in the Ignore Parts of Speech dialog box are ignored during parsing. The selections indicated here ensure that the analysis ignores low-content words such as prepositions and determiners.

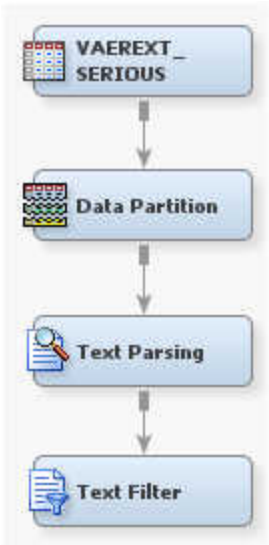
13. Click **OK**.

Filter Data

The **Text Filter** node can be used to reduce the total number of parsed terms or documents that are analyzed. Therefore, you can eliminate extraneous information so that only the most valuable and relevant information is considered. For example, the **Text Filter** node can be used to remove unwanted terms and to keep only documents that discuss a particular issue. This reduced data set can be orders of magnitude smaller than the one that represents the original collection, which might contain hundreds of thousands of documents and hundreds of thousands of distinct terms. For more information about the **Text Filter** node, see the SAS Text Miner help.

To filter the data:

1. Select the **Text Mining** tab on the node toolbar, and drag a **Text Filter** node into the diagram workspace.
2. Connect the **Text Parsing** node to the **Text Filter** node.



3. Select the **Text Filter** node.
4. Set the value of the **Term Weight** property to `Mutual Information`.

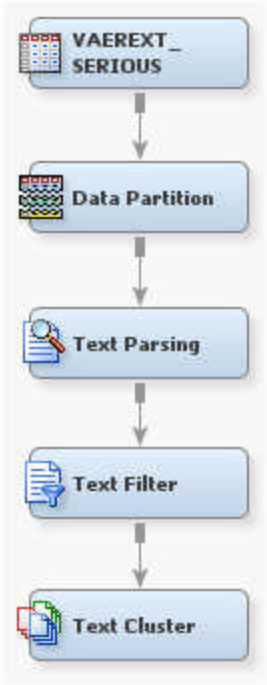
This causes the terms to be differentially weighted when they correspond to serious reactions.

Cluster Data

The **Text Cluster** node clusters documents into disjointed sets of documents and reports on the descriptive terms for those clusters. Two algorithms are available. The Expectation Maximization algorithm clusters documents with a flat representation, and the Hierarchical clustering algorithm groups clusters into a tree hierarchy. Both approaches rely on the singular value decomposition (SVD) to transform the original weighted, term-document frequency matrix into a dense but low dimensional representation. For more information about the **Text Cluster** node, see the SAS Text Miner help.

To cluster the data:

1. Select the **Text Mining** tab on the node toolbar, and drag a **Text Cluster** node into the diagram workspace.
2. Connect the **Text Filter** node to the **Text Cluster** node.



3. Select the **Text Cluster** node.
4. Set the **Descriptive Terms** to `12` to ease cluster labeling.
5. Right-click the **Text Cluster** node in the diagram workspace, and select **Run**.
6. Click **Yes** in the Confirmation dialog box when you are asked whether you want to run the path.
7. Click **OK** in the Run Status dialog box that appears after the **Text Cluster** node has finished running.

View Results

After the process flow diagram has completed running, you can view the results that were obtained by each node.

1. Select the **Text Parsing** node.

   The Properties for the **Text Parsing** node appear in the Properties Panel.

   Notice that the **Text Parsing** node's **Parse Variable** property has been populated with the SYMPTOM_TEXT variable. This is because the SYMPTOM_TEXT variable was the longest variable with a role of **`Text`** in the **VAEREXT_SERIOUS** input data source.

2. Right-click the **Text Parsing** node and select **Results**.

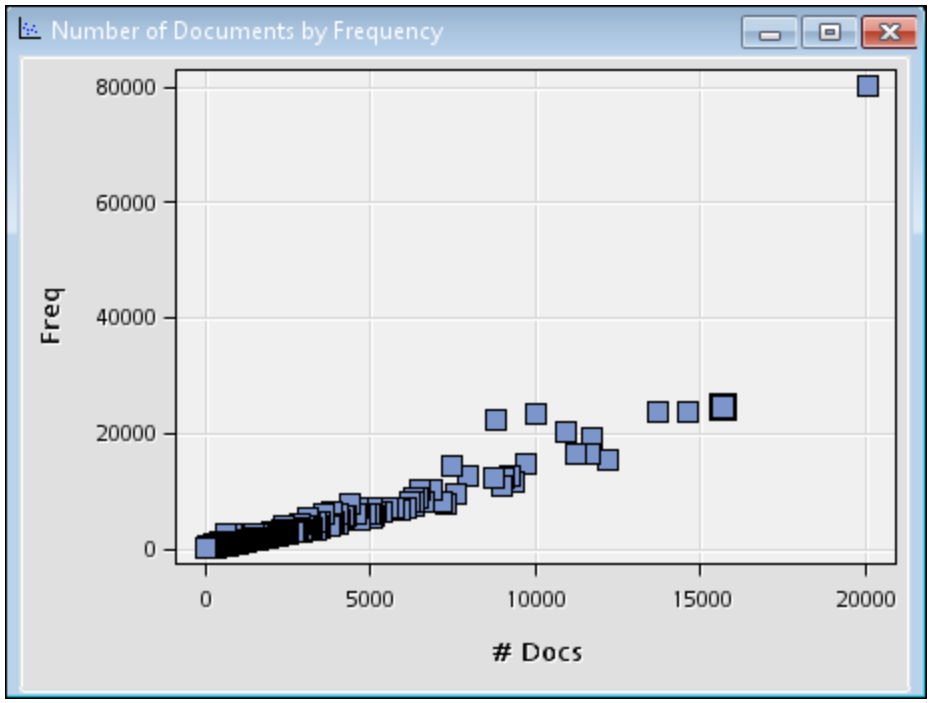
   The Results window appears for the **Text Parsing** node.

3. Select the Terms window.

4. Click the **Freq** column heading to sort the terms by frequency.

   Scroll through the list of terms. Notice that for each term, the Terms window provides the number of documents the term appeared in, the frequency of the term, and whether the term was kept.

5. Select a term. Notice that the point corresponding to this term is selected in the ZIPF Plot and the Number of Documents by Frequency plot.



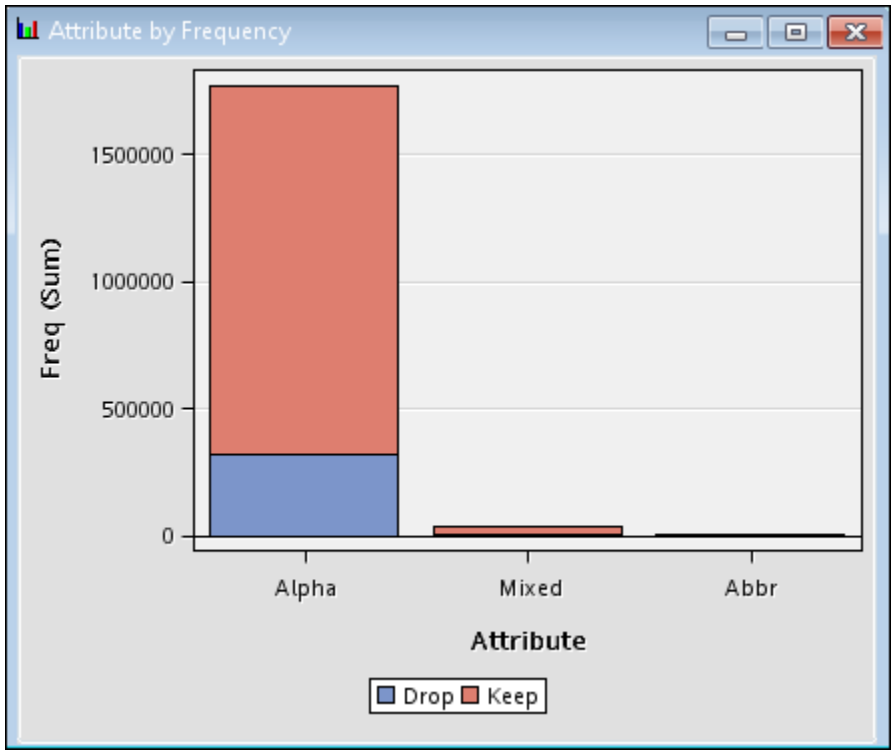
6. Close the Results window.

7. Select the **Text Filter** node.

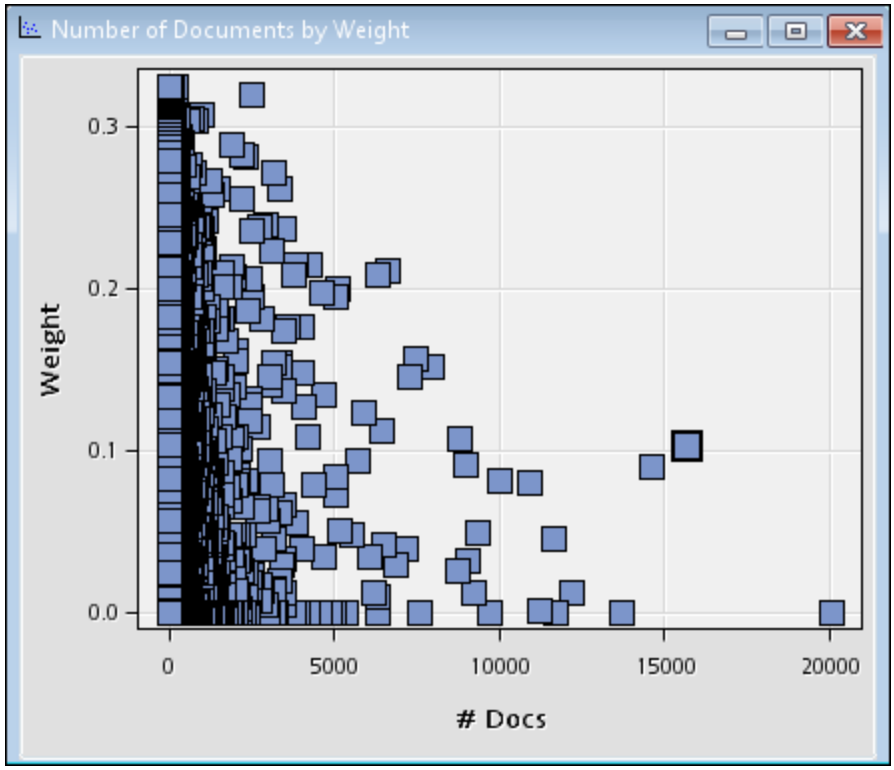
8. Right-click the **Text Filter** node and select **Results**.

   The Results window appears for the **Text Filter** node.

Notice that the Attribute by Frequency window and the Role by Freq window now show the number of terms in each category that were dropped or kept.



The Number of Documents by Weight plot shows the number of documents in which each term appears relative to each term's weight.



9. Close the Results window.
10. Click the ... for the **Filter Viewer** property.

The Interactive Filter Viewer window appears.

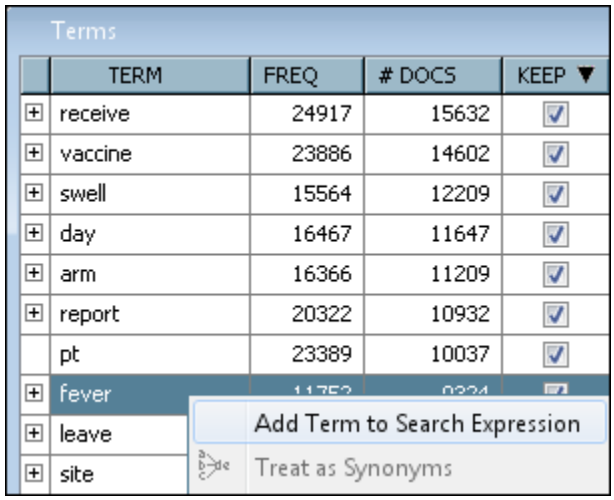
11. View the terms in the Terms window. The terms are sorted first by their keep status and then by the number of documents that they appear in.

    Note: You can change the sorted order by clicking a column heading.

12. View the documents in the Documents window.

13. Right-click a cell in the **SYMPTOM_TEXT** column, and then select **Toggle Show Full Text** to see the full text contained in SYMPTOM_TEXT.

14. Select a term that is related to an adverse reaction that you want to investigate further. For example, select **fever** under the TERM column of the Terms window. Right-click on the term and select **Add Term to Search Expression**.



15. Click **Apply**.

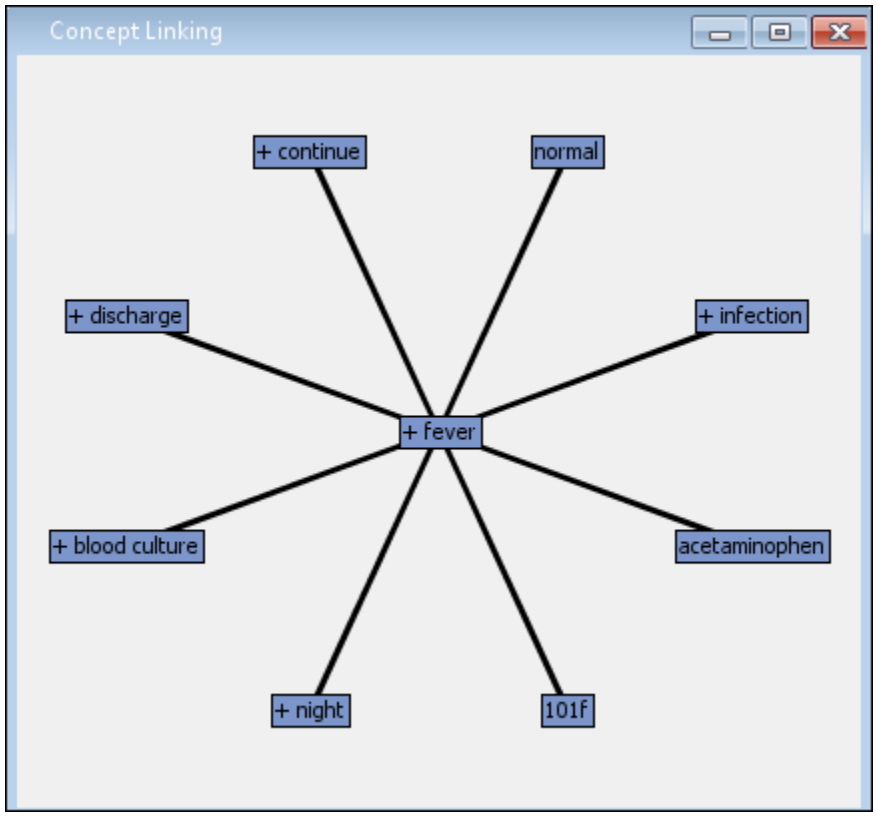
    Notice that the Documents window updates to only include entries that contain the term **fever**.

16. Click **Clear**, and then click **Apply**.

    The terms in the Terms window resets.

17. Select the term **fever** in the Terms window, and then right-click it and select **View Concept Links**.

    The Concept Linking window appears. Concept linking is a way to find and display the terms that are highly associated with the selected term in the Terms table. The selected term is surrounded by the terms that correlate the strongest with it. The Concept Linking window shows a hyperbolic tree graph with **fever** in the center of the tree structure. It shows you the other terms that are strongly associated with the term **fever**.



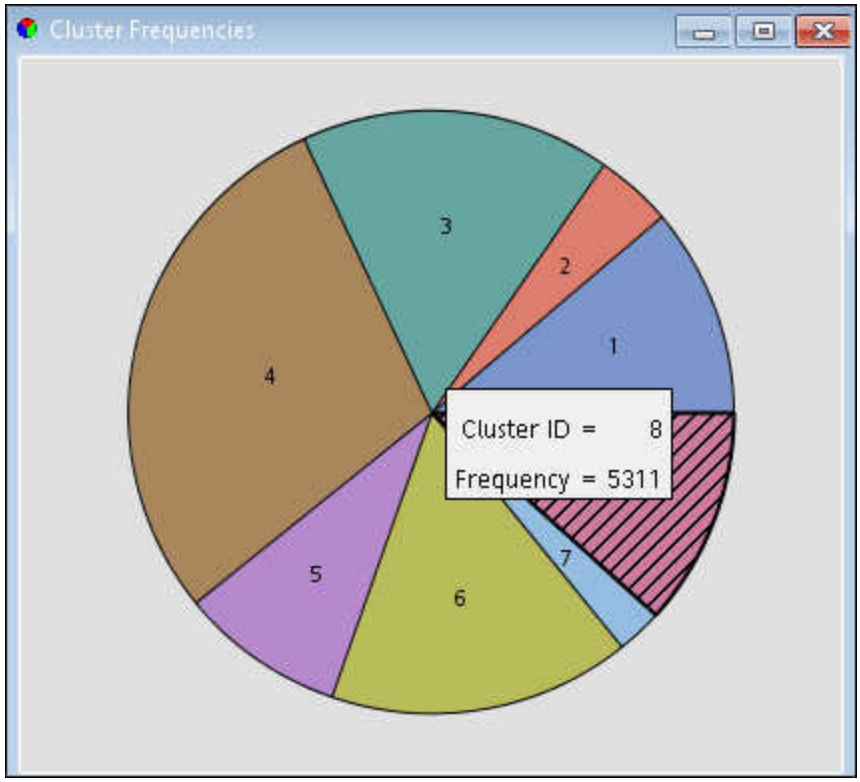
To expand the Concept Linking view, select a term that is not in the center of the graph, right-click it, and select **Expand Links**.

18. Close the Results window.
19. Select the **Text Cluster** node.
20. Right-click the **Text Cluster** node and select **Results**.

    The Results window appears.

21. View the clusters in the Clusters window. Select a cluster.

    Notice how the corresponding cluster is selected in the Cluster Frequencies chart, the Cluster Frequency by RMS plot, and the Distance Between Clusters plot.



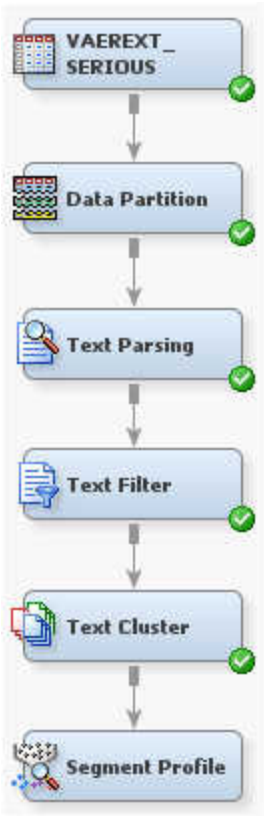
22. Close the Results window.

Examine Data Segments

In this section, you will examine segmented or clustered data using the **Segment Profile** node. A segment is a cluster number that you derive analytically by using SAS Text Miner clustering techniques. The **Segment Profile** node enables you to get a better idea of what makes each segment unique or at least different from the population. The node generates various reports that aid in exploring and comparing the distribution of these factors within the segments and population. For more information about the **Segment Profile** node, see the SAS Enterprise Miner Help.

To examine data segments, complete the following steps:

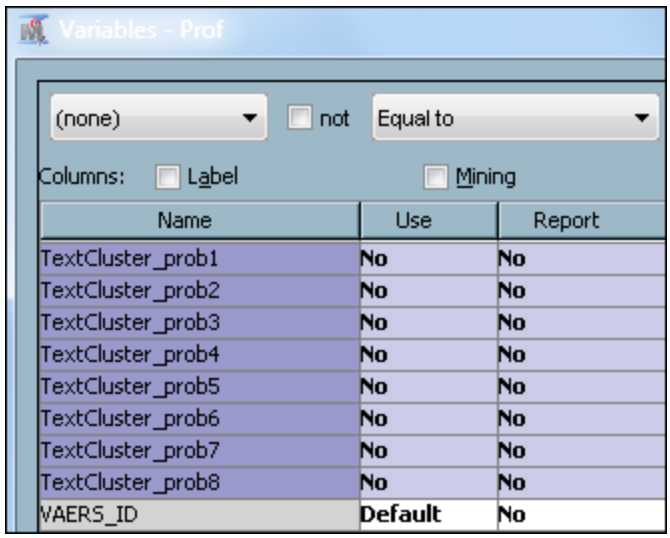
1. Select the **Assess** tab on the node toolbar, and drag a **Segment Profile** node into the diagram workspace.
2. Connect the **Text Cluster** node to the **Segment Profile** node.



3. Select the **Segment Profile** node.
4. Click the ... for the **Variables** property.

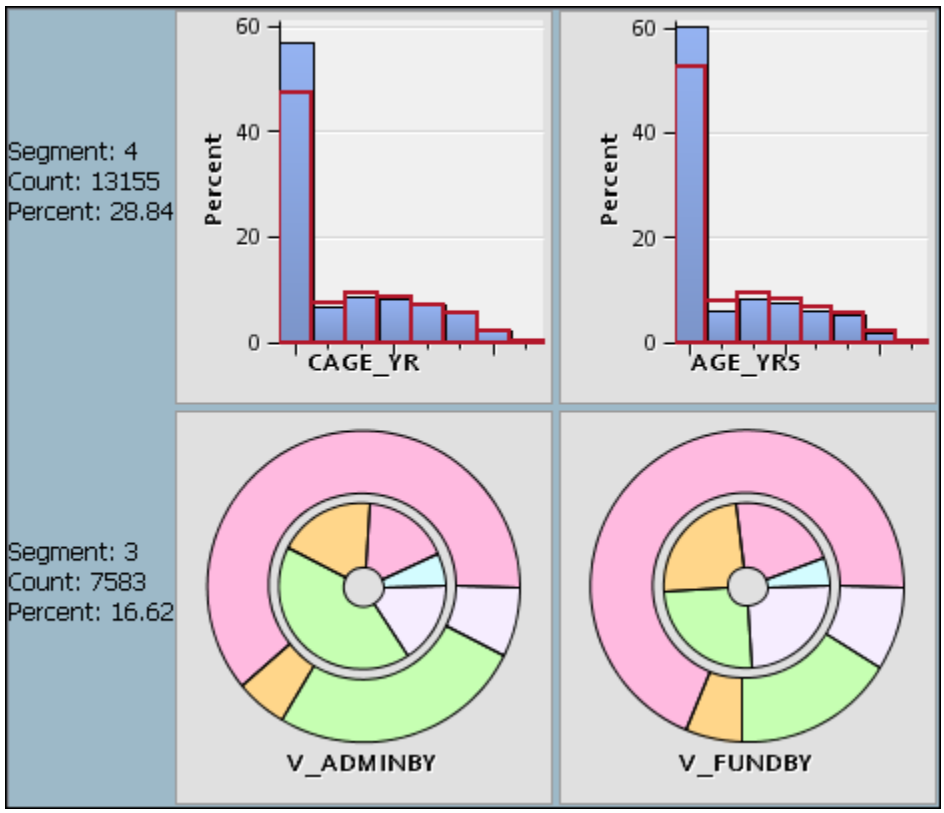
   The Variables window appears.
5. Select all the "_prob" variables and set their Use value to **No**.

   Note: You can hold down Shift and select all the "_prob" variables by clicking on the first "_prob" variable and dragging the pointer to select all "_prob" variables. After all "_prob" variables are selected, you can change the Use value of each selected "_prob" variable by changing the Use value of one of the "_prob" variables. This will change the other "_prob" Use values to the selected value as well.



6. Select all the "_SVD" variables and set their **Use** value to **No**.
7. Click **OK**.
8. Select the **Segment Profile** node in the diagram workspace.
9. Enter *0.0010* as the value for the **Minimum Worth** property.
10. Right-click the **Segment Profile** node, and select **Run**.
11. Click **Yes** in the Confirmation dialog box when you are asked whether you want to run the path.
12. After the node finishes running, click **Results** in the Run Status dialog box.
13. Maximize the Profile window.

    The following shows a portion of this window.



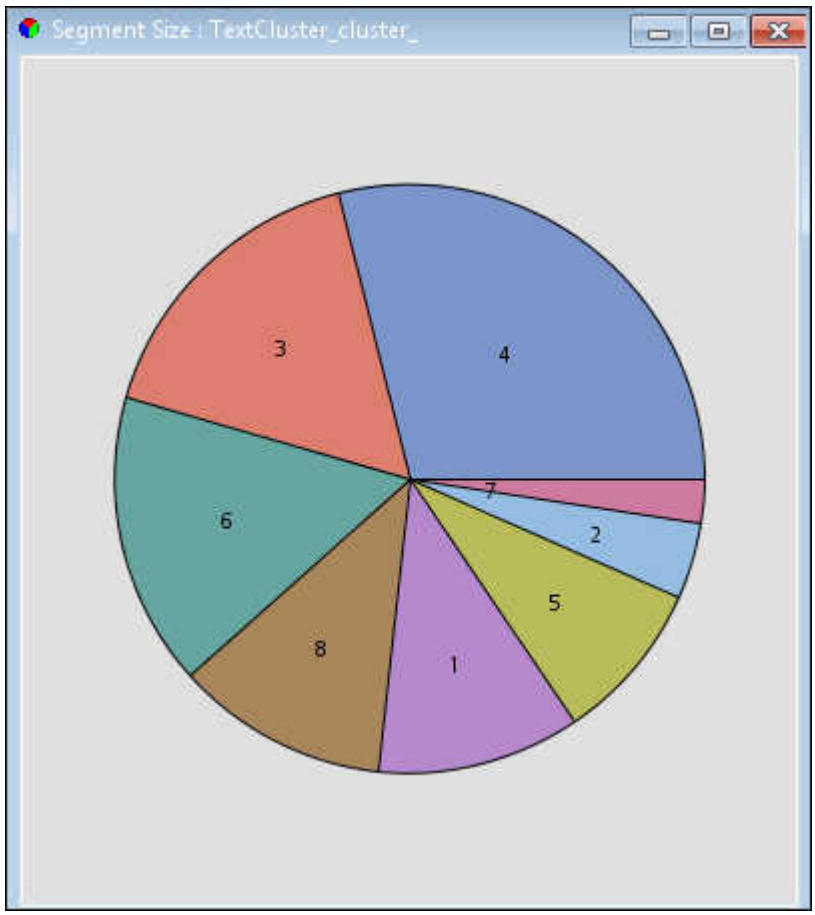
The Profile window displays a lattice, or grid, of plots that compare the distribution for the identified and report variables for both the segment and the population. The graphs shown in this window illustrate variables that have been identified as factors that distinguish the segment from the population that it represents. Each row represents a single segment. The far-left margin identifies the segment, its count, and the percentage of the total population.

The columns are organized from left to right according to their ability to discriminate that segment from the population. Report variables, if specified, appear on the right in alphabetical order after the selected inputs. The lattice graph has the following features:

- Class variable — is displayed as two nested pie charts that consist of two concentric rings. The inner ring represents the distribution of the total population. The outer ring represents the distribution for the given segment.

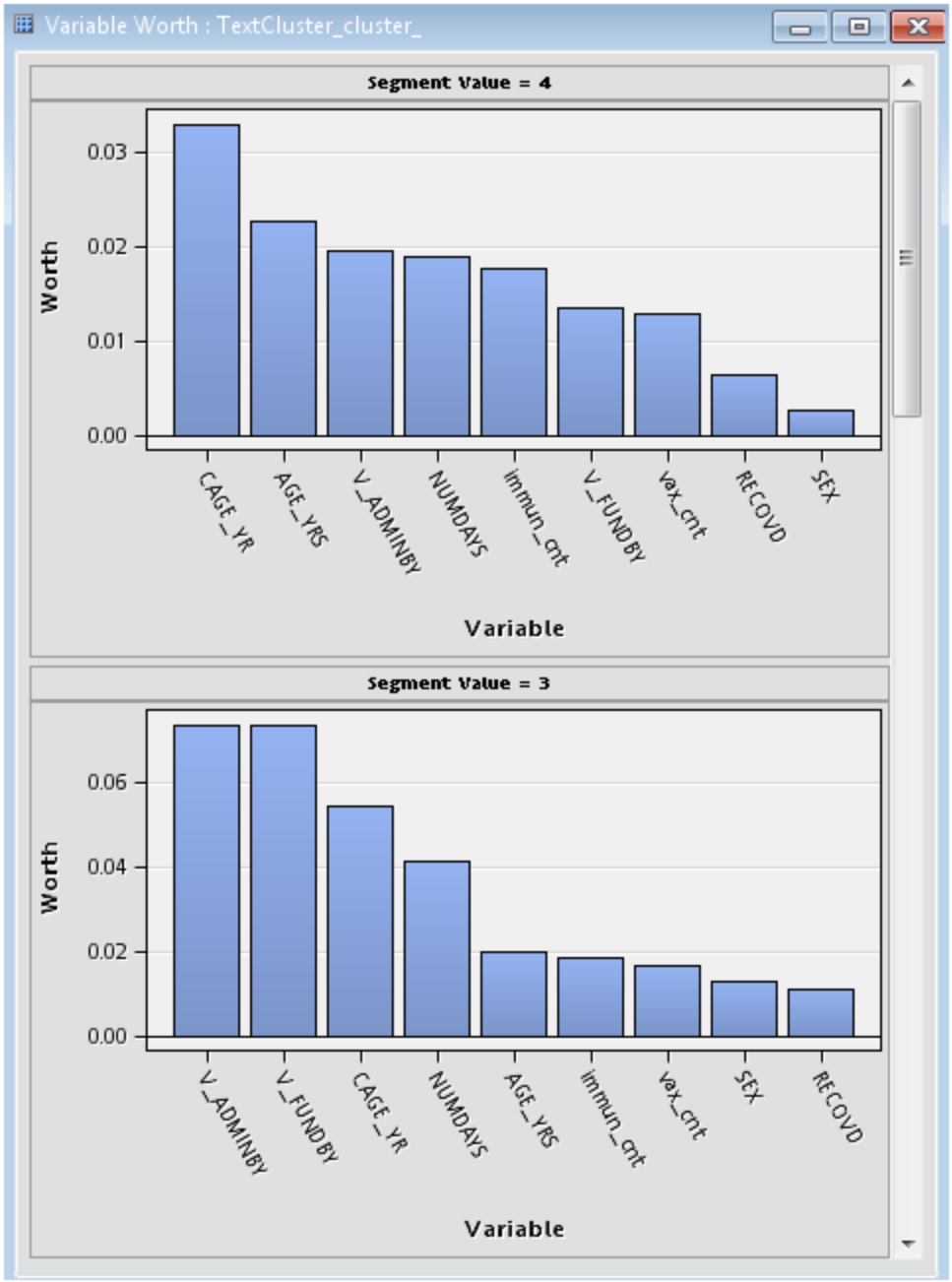
- Interval variable — is displayed as a histogram. The blue shaded region represents the within-segment distribution. The red outline represents the population distribution. The height of the histogram bars can be scaled by count or by percentage of the segment population. When you are using the percentage, the view shows the relative difference between the segment and the population. When you are using the count, the view shows the absolute difference between the segment and the population.

14. Maximize the Segment Size chart.



15. Maximize the Variable Worth window.

    The following shows a portion of this window.



16. Close the **Results** window.

Chapter 5

Cleaning Up Text

About the Tasks That You Will Perform

As demonstrated in the previous chapter, SAS Text Miner does a good job of finding themes that are clear in the data. But, when the data needs cleaning, SAS Text Miner can be less effective at uncovering useful themes. In this chapter, you will encounter manually edited data that contains many misspellings and abbreviations, and you will work on cleaning the data to get better results.

The README.TXT file provided in the Getting Started with SAS Text Miner 13.2 zip file contains a list of abbreviations that are commonly used in the adverse event reports. SAS Text Miner enables you to specify a synonym list. A VAER_ABBREV synonym list is provided for you in the Getting Started with SAS Text Miner 13.2 zip file. So that you can create such a synonym list, the abbreviations list from README.TXT was copied into a Microsoft Excel file. The list was manually edited in the Microsoft Excel file and then imported into a SAS data set. For example, CT was marked as equivalent to computerized axial tomography.

For more information about importing data into a SAS data set, see the following documentation resource:

`http://support.sas.com/documentation/`

You will perform the following tasks to clean the text and examine the results:

1. Use a synonym data set from the Getting Started with SAS Text Miner 13.2 zip file.
2. Create a new synonym data set by using the **SAS Code** node and the %TEXTSYN macro. The %TEXTSYN macro will run through all the terms, automatically identify which ones are misspellings, and create synonyms that map correctly spelled terms to the misspelled terms.
3. Examine results using merged synonym data sets.

Use a Synonym Data Set

To use a synonym data set:

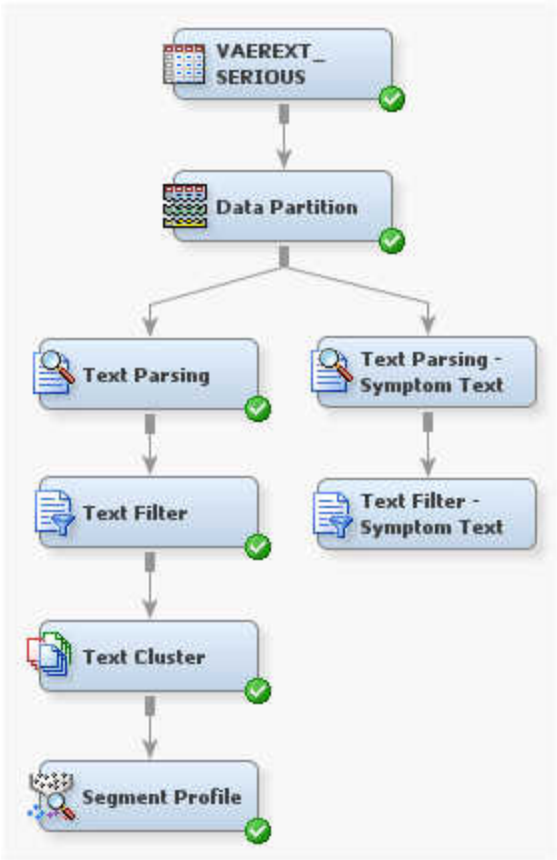
1. Right-click the **Text Parsing** node, and select **Copy**.

   For this example, it is important to copy the node instead of creating a new **Text Parsing** node because the settings you previously specified in the **Text Parsing** node properties panel will be used.

2. Right-click in the empty diagram workspace, and select **Paste**.
3. To distinguish this newly pasted **Text Parsing** node from the first node, right-click it, and select **Rename**.
4. Enter *Text Parsing — Symptom Text* in the **Node Name** field, and then click **OK**.
5. Right-click the **Text Filter** node, and select **Copy**.

   For this example, it is important to copy the node instead of creating a new **Text Filter** node because the settings you previously specified in the **Text Filter** node properties panel will be used.

6. Right-click in the empty diagram workspace, and select **Paste**.
7. To distinguish this newly pasted **Text Filter** node from the first node, right-click it, and select **Rename**.
8. Enter *Text Filter — Symptom Text* in the **Node Name** field, and then click **OK**.
9. Connect the **Data Partition** node to the **Text Parsing — Symptom Text** node.
10. Connect the **Text Parsing — Symptom Text** node to the **Text Filter — Symptom Text** node.



11. Select the **Text Parsing — Symptom Text** node.
12. Select the ... for the **Synonyms** property.

    A dialog box appears.
13. Click **Replace Table**.

    The Select a SAS Table dialog box appears.
14. Select the **Mylib** library in the folder tree.

    The contents of the **Mylib** library appear.
15. Select **Vaer_abbrev**, and then click **OK**.

    The contents of the **Vaer_abbrev** data source appear in a dialog box.
16. Click **Yes** in the Confirmation window.
17. Click **OK**.

    Leave all other settings the same as in the original **Text Parsing** node.
18. Right-click the **Text Filter — Symptom Text** node in the diagram workspace, and select **Run**. Click **Yes** in the Confirmation dialog box.
19. Click **OK** in the Run Status dialog box when the node has finished running.
20. Click the ... for the **Filter Viewer** property of the **Text Filter — Symptom Text** node.

    The Interactive Filter Viewer window appears.
21. Click the TERM column heading to sort the Terms table.

22. Select **abdomen** under the TERM column in the Terms window.

    You might need to scroll down to see the term. In the Terms window, there should be a plus (+) sign next to **abdomen**. Click on the plus sign to expand the term. This shows all synonyms and stems that are mapped to that term. A stem is the root form of a term. The child term **abd** is included. Both **abdomen** and **abd** will be treated the same.

	TERM ▲	FREQ	# DOCS	KEEP	WEIGHT	ROLE
	abdome	1	1	☐	0.0	
⊟	abdomen	661	600	☑	0.041	
	abd	100	93			
	abdomen	561	513			
	abdomen area	1	1	☐	0.0	Noun Group

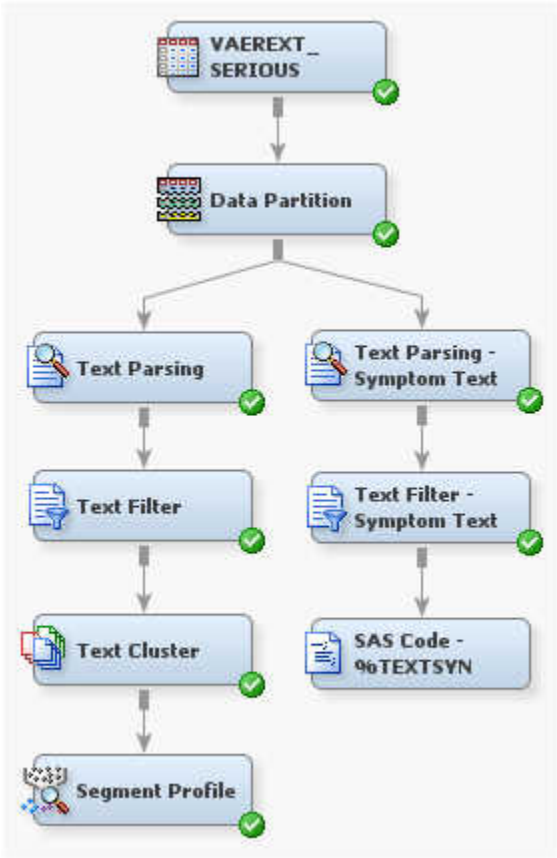
23. Close the Interactive Filter Viewer window.

Create a New Synonym Data Set

You can use the SAS Text Miner %TEXTSYN macro to create a new synonym data set. The %TEXTSYN macro evaluates all the terms, automatically identifies which terms are misspellings, and creates synonyms that map correctly spelled terms to misspelled terms.

To create a new synonym data set:

1. Select the **Utility** tab on the node toolbar and drag a **SAS Code** node into the diagram workspace.
2. Right-click the **SAS Code** node, and select **Rename**.
3. Enter *SAS Code — %TEXTSYN* in the **Node Name** field, and then click **OK**.
4. Connect the **Text Filter — Symptom Text** node to the **SAS Code — %TEXTSYN** node.



5. Select the **SAS Code — %TEXTSYN** node, and then click the ... for the **Code Editor** property in the Properties Panel.

   The Code Editor window appears.

6. Enter the following code in the Code Editor:

```
%textsyn( termds=<libref>.<nodeID>_terms
        , docds=&em_import_data
        , outds=&em_import_transaction
        , textvar=symptom_text
        , mnpardoc=8
        , mxchddoc=10
        , synds=mylib.vaerextsyns
        , dict=mylib.engdict
        , maxsped=15
        ) ;
```

   Note: You will need to replace **<libref>** and **<nodeID>** in the first line in the above code with the correct library name and node ID. To determine what these values are, close the Code Editor window, and then select the arrow that connects the **Text Filter — Symptom Text** node to the **SAS Code — %TEXTSYN** node. The value for <libref> will be the first part of the table name that appears in the Properties panel, such as **emws**, **emws2**, and so on. The node ID will appear after the value for <libref>, and will be **TextFilter**, **TextFilter2**, and so on. After you determine the value for <libref> and <nodeID>, a possible first line might be **termds=emws2.textfilter2_terms**. Your libref and node ID values could differ depending on how many **Text Filter** nodes and diagrams have been created in your workspace.

   For details about the %TEXTSYN macro, see SAS Text Miner Help documentation.

7. After you have added the %TEXTSYN macro code to the Code Editor window, and modified it to add values for **<libref>** and **<nodeID>**, click the  to save the changes.
8. Click the  to run the **SAS Code — %TEXTSYN** node.
9. Click **Yes** in the Confirmation dialog box.
10. Click **OK** in the dialog box that indicates that the node has finished running.
11. Close the **Code Editor** window.
12. Select **View** ⇨ **Explorer** from the main menu.

    The Explorer window appears.
13. Click **Mylib** in the SAS Libraries tree, and then select **Vaerextsyns**.

    Note: If the **Mylib** library is already selected and you do not see the Vaerextsyns data set, you might need to click **Show Project Data** or refresh the Explorer window to see the **Vaerextsyns** data set.
14. Double-click **Vaerextsyns** to see its contents.

	example1	example2	Term	parent
116	... 4 days following. Heat, !!anti-inflammatories!!. 11/18/2002, Muscle relaxant...	... over the counter non steroidal !!anti-inflam...	anti-inflammatories	anti-inflammatory
117	... TO TOUCH. WAS GIVEN !!ANTIBOTICS!! IF SWELLING INCREASES.	... where they gave her !!antibotics!!/steroids.	antibotics	antibiotics
118	... erythematous papules over B/L !!anticubital!! spaces, elbows RLQABD, Axilla,...	... hives on face, neck, ! !anticubital!! and right foot and...	anticubital	antecubital

    Here is a list of what the Vaerextsyns columns provide:

    - **Term** is the misspelled word.
    - **parent** is a guess at the word that was meant.
    - **example1** and **example2** are two examples of the term in a document.
    - **childndocs** is the number of documents that contained that term.
    - **numdocs** is the number of documents that contained the parent.
    - **minsped** is an indication of how close the terms are.
    - **dict** indicates whether the term is a legitimate English word. Legitimate words can still be deemed misspellings, but only if they occur rarely and are very close in spelling to a frequent target term.

    For example, Observation 117 shows **antibotics** to be a misspelling of **antibiotics**. Four documents contain **antibotics**, and 745 documents contain the parent. Note that double exclamation marks (!!) both precede and follow the child term in the example text so that you can see the term in context.
15. Examine the Vaerextsyns table to see whether you disagree with some of the choices made. For this example, however, assume that the %TEXTSYN macro has done a good enough job of detecting misspellings.

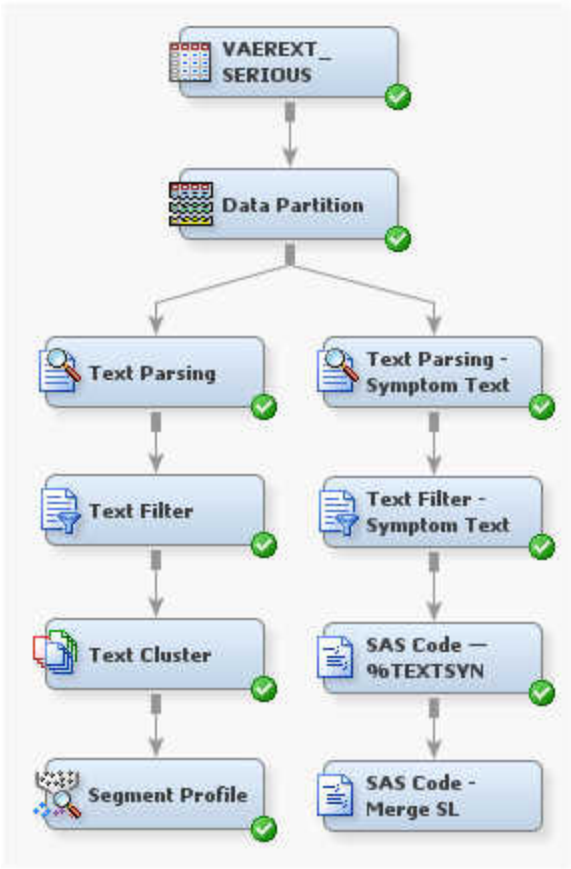
    Note: The Vaerextsyns table can be edited using any SAS table editor. You cannot edit this table in the SAS Enterprise Miner GUI. You can change a parent for any misspellings that appear incorrect or delete a row if the Term column contains a valid term.
16. Close the Mylib.Vaerextsyns table and the Explorer window.

Use Merged Synonym Data Sets

In this set of tasks, you will create a new data set that contains all the observations from both the Mylib.Vaerextsyns and Mylib.Vaer_abbrev data sets. You will examine the results by using the merged synonym data set. Complete the following steps:

1. Select the **Utility** tab on the node toolbar and drag a **SAS Code** node into the diagram workspace.
2. Right-click the **SAS Code** node, and select **Rename**.
3. Enter *SAS Code — Merge SL* in the **Node Name** field, and then click **OK**.

   SL stands for *Synonym Lists*.
4. Connect the **SAS Code — %TEXTSYN** node to the **SAS Code — Merge SL** node.



5. Select the **SAS Code — Merge SL** node.
6. Click the ... for the **Code Editor** property.

   The Code Editor appears.
7. Enter the following code in the Code Editor:

```
data mylib.vaerextsyns_new;
     set mylib.vaerextsyns mylib.vaer_abbrev;
run;
```

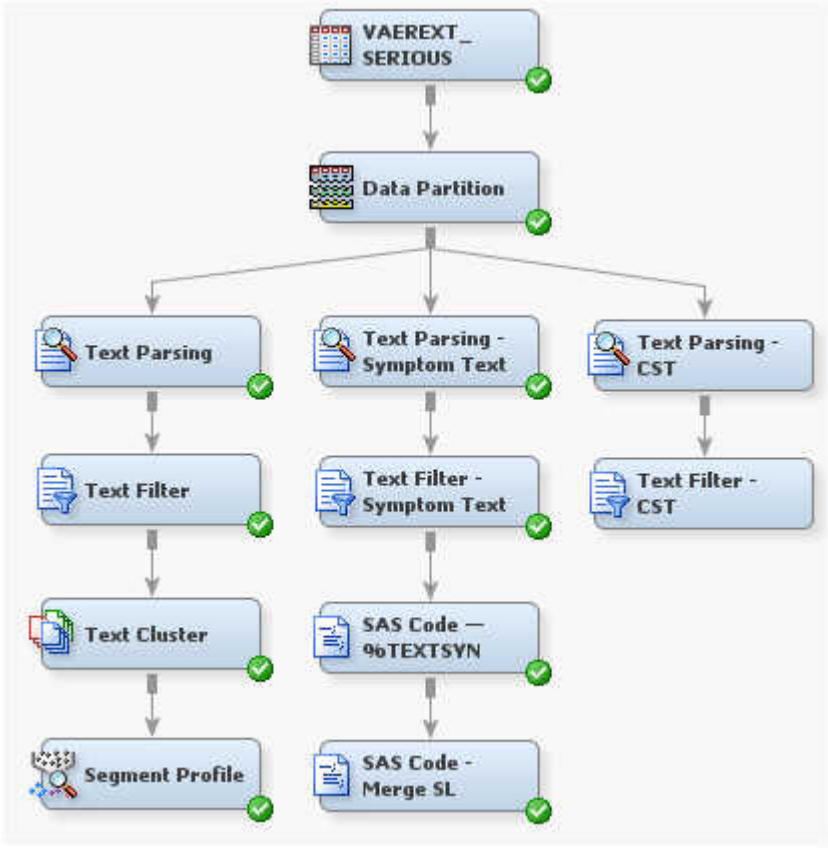
This code merges the resulting synonyms data set from the first **SAS Code — %TEXTSYN** node with the abbreviations data set.

8. Click  .
9. Close the Code Editor window.
10. Right-click the **SAS Code — Merge SL** node, and select **Run**. Click **Yes** in the Confirmation dialog box.
11. Click **Results** in the Run Status dialog box when the node has finished running.
12. From the Results window, select **View** ⇨ **SAS Results** ⇨ **Log** to see the SAS code where the new data set is created.

    Close the Results window.
13. Right-click the **Text Parsing — Symptom Text** node, and then select **Copy**.

    Note: It is important to copy the **Text Parsing — Symptom Text** node instead of creating a new Text Miner node. You do this in order to keep the same property settings you previously configured for the **Text Parsing — Symptom Text** node.
14. Right-click an empty space in the diagram workspace and select **Paste**.
15. Right-click the new **Text Parsing** node, and select **Rename**.
16. Enter *Text Parsing — CST* in the **Node Name** field.

    CST stands for *Cleaned Symptom Text*.
17. Click **OK**.
18. Right-click the **Text Filter — Symptom Text** node, and then select **Copy**.
19. Right-click an empty space in the diagram workspace and select **Paste**.
20. Right-click the new **Text Filter** node, and select **Rename**.
21. Enter *Text Filter — CST* in the **Node Name** field.
22. Click **OK**.
23. Connect the **Data Partition** node to the **Text Parsing — CST** node.
24. Connect the **Text Parsing — CST** node to the **Text Filter — CST** node.



25. Select the **Text Parsing — CST** node.
26. Click the ... for the **Synonyms** property.
27. Click **Replace Table**.
28. Click **Mylib** in the SAS Libraries tree.

    The contents of the Mylib library appear.
29. Select **Mylib.Vaerextsyns_new**.
30. Click **OK**.
31. Click **Yes** in the Confirmation dialog box.

    The contents of the data set appear in the dialog box.
32. Click **OK**.
33. Right-click the **Text Filter — CST** node, and select **Run**. Click **Yes** in the Confirmation dialog box.
34. Click **OK** in the Run Status dialog box when the node has finished running.
35. Select the **Text Filter — CST** node.
36. Click the ... for the **Filter Viewer** property.

    The Interactive Filter Viewer window appears.
37. Scroll down, and select the plus sign (+) next to **patient** in the Terms table. Note that the misspellings **patien**, **patietn**, and **paitent** are included as child terms.

	TERM ▲	FREQ	# DOCS
⊟	patient	22707	9000
	ptsd	4	4
	patietn	5	5
	patien	7	7
	patients	828	598
	patient	21580	8640
	paitent	5	5
	pts	278	215

38. Close the Interactive Filter Viewer.

Chapter 6

Create Topics and Rules

About the Tasks That You Will Perform

This chapter shows how you can create topics and rules from filtered terms using the **Text Topic** node, and the **Text Rule Builder** node.

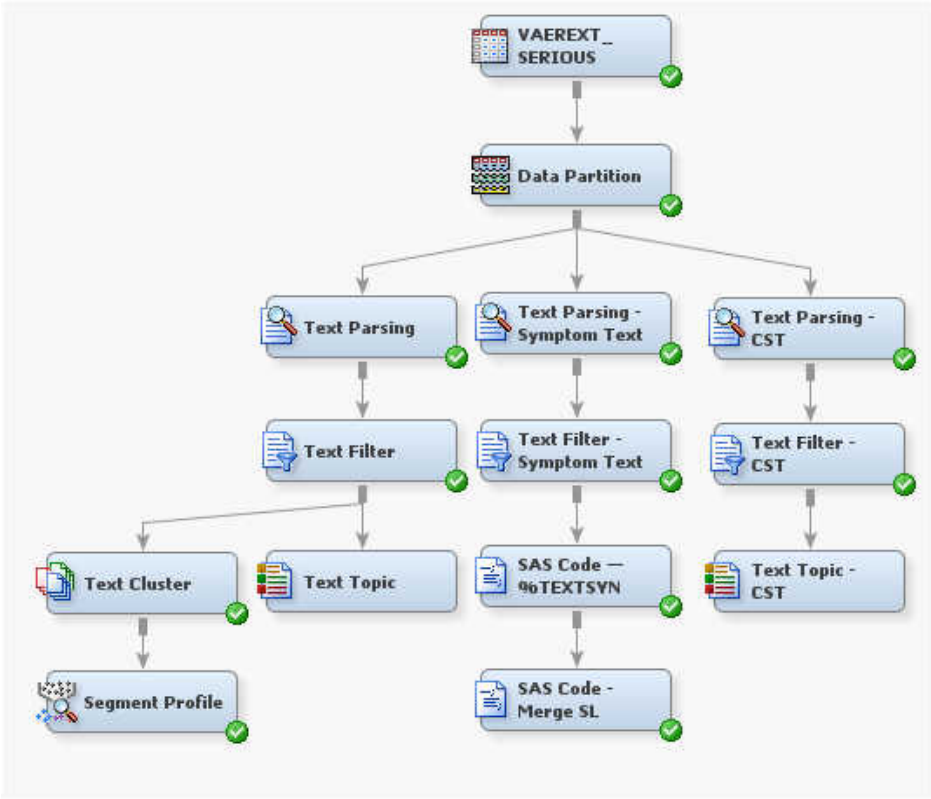
The **Text Topic** node enables you to explore the document collection by automatically associating terms and documents according to both discovered and user-defined topics. Topics are collections of terms that describe and characterize a main theme or idea. The goal in creating a list of topics is to establish combinations of words that you are interested in analyzing. The ability to combine individual terms into topics can improve your text mining analysis. Through combining, you can narrow the amount of text that is subject to analysis to specific groupings of words that you are interested in. For more information about the **Text Topic** node, see the SAS Text Miner Help.

The **Text Rule Builder** node generates an ordered set of rules from small subsets of terms that together are useful in describing and predicting a target variable. Each rule in the set is associated with a specific target category that consists of a conjunction that indicates the presence or absence of one or a small subset of terms (for example, “term1” AND “term2” AND (NOT “term3”)). A particular document matches this rule if and only if it contains at least one occurrence of term1 and of term2 but no occurrences of term3. This set of derived rules creates a model that is both descriptive and predictive. When categorizing a new document, it proceeds through the ordered set and chooses the target that is associated with the first rule that matches that document. The rules are provided in the syntax that can be used within SAS Content Categorization Studio, and can be deployed there. For more information about the **Text Rule Builder** node, see the SAS Text Miner help.

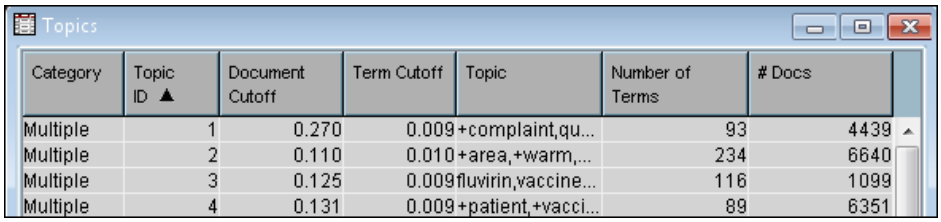
Create Topics

After filtering text, you can use the **Text Topic** node to create topics. Perform the following steps to use **Text Topic** nodes in the analysis:

1. Select the **Text Mining** tab on the node toolbar and drag a **Text Topic** node into the diagram workspace.
2. Connect the **Text Filter** node to the **Text Topic** node.
3. Select the **Text Mining** tab on the node toolbar and drag a **Text Topic** node into the diagram workspace.
4. Right-click the **Text Topic** node, and select **Rename**.
5. Enter *Text Topic — CST* in the **Node Name** field, and then click **OK**.
6. Select the **Text Topic — CST** node.
7. Enter `50` as the value for the **Number of Multi-term Topics** property.
8. Connect the **Text Filter — CST** node to the **Text Topic — CST** node.



9. Right-click the **Text Topic** node in the diagram workspace, and select **Run**. Click **Yes** in the Confirmation dialog box.
10. Click **Results** in the Run Status dialog box when the node has finished running.
11. Review the topics in the Topics table to see which terms make up each topic.

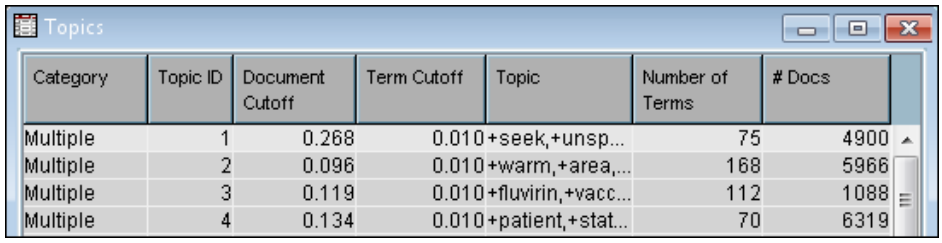

Topics

Category	Topic ID ▲	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Multiple	1	0.270	0.009	+complaint,qu...	93	4439
Multiple	2	0.110	0.010	+area,+warm,...	234	6640
Multiple	3	0.125	0.009	fluvirin,vaccine...	116	1099
Multiple	4	0.131	0.009	+patient,+vacci...	89	6351

12. Close the Results window.
13. Right-click the **Text Topic — CST** node in the diagram workspace, and select **Run**. Click **Yes** in the Confirmation dialog box.

14. Click **Results** in the Run Status dialog box when the node has finished running.
15. Review the topics in the Topics table to see which terms make up each topic.

    Notice that the topics that were generated by running the **Text Topic — CST** are different from those that were generated by running the **Text Topic** node.


Topics

Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Multiple	1	0.268	0.010	+seek,+unsp...	75	4900
Multiple	2	0.096	0.010	+warm,+area,...	168	5966
Multiple	3	0.119	0.010	+fluvirin,+vacc...	112	1088
Multiple	4	0.134	0.010	+patient,+stat...	70	6319

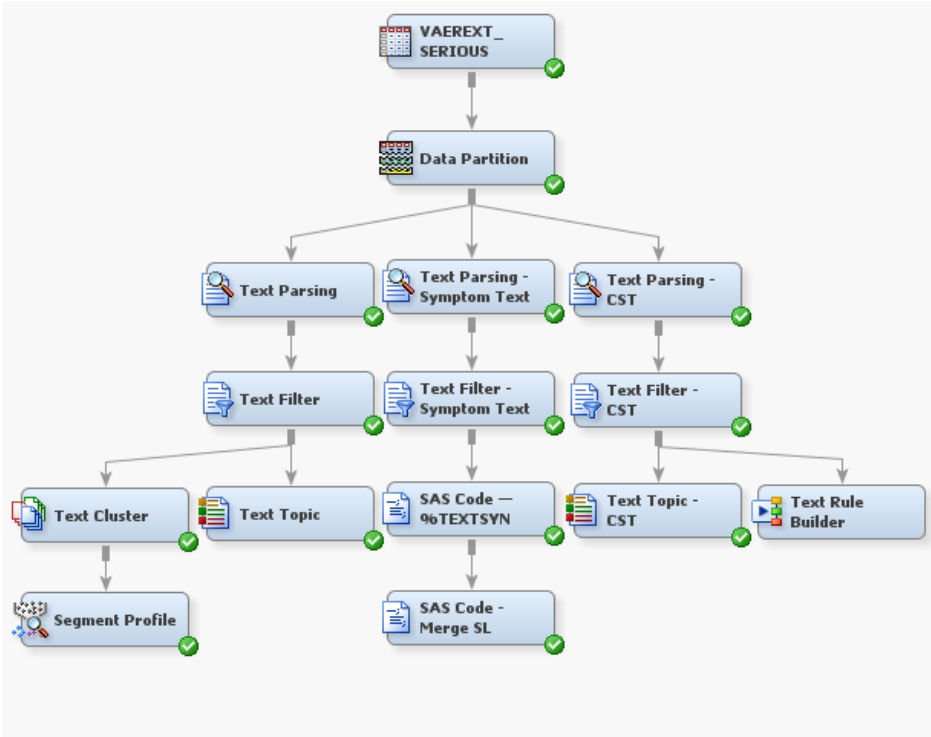
    The differences follow from previously performed text cleaning activities, and the number of multi-term topics.

16. Close the Results window.

Create Rules

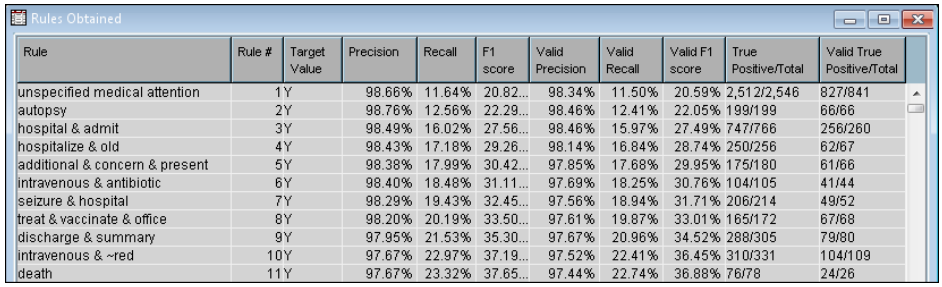
After filtering text, you can use the **Text Rule Builder** node to create rules. Perform the following steps to use a **Text Rule Builder** node in the analysis:

1. Select the **Text Mining** tab on the node toolbar and drag a **Text Rule Builder** node into the diagram workspace.
2. Connect the **Text Filter — CST** node to the **Text Rule Builder** node.



3. Right-click the **Text Rule Builder** node in the diagram workspace, and select **Run**. Click **Yes** in the Confirmation dialog box.

4. Click **Results** in the Run Status dialog box when the node has finished running.
5. Select the Rules Obtained window.


Rules Obtained

Rule	Rule #	Target Value	Precision	Recall	F1 score	Valid Precision	Valid Recall	Valid F1 score	True Positive/Total	Valid True Positive/Total
unspecified medical attention	1	Y	98.66%	11.64%	20.82...	98.34%	11.50%	20.59%	2,512/2,546	827/841
autopsy	2	Y	98.76%	12.56%	22.29...	98.46%	12.41%	22.05%	199/199	66/66
hospital & admit	3	Y	98.49%	16.02%	27.56...	98.46%	15.97%	27.49%	747/766	256/260
hospitalize & old	4	Y	98.43%	17.18%	29.26...	98.14%	16.84%	28.74%	250/256	62/67
additional & concern & present	5	Y	98.38%	17.99%	30.42...	97.85%	17.68%	29.95%	175/180	61/66
intravenous & antibiotic	6	Y	98.40%	18.48%	31.11...	97.69%	18.25%	30.76%	104/105	41/44
seizure & hospital	7	Y	98.29%	19.43%	32.45...	97.56%	18.94%	31.71%	206/214	49/52
treat & vaccinate & office	8	Y	98.20%	20.19%	33.50...	97.61%	19.87%	33.01%	165/172	67/68
discharge & summary	9	Y	97.95%	21.53%	35.30...	97.67%	20.96%	34.52%	288/305	79/80
intravenous & ~red	10	Y	97.67%	22.97%	37.19...	97.52%	22.41%	36.45%	310/331	104/109
death	11	Y	97.67%	23.32%	37.65...	97.44%	22.74%	36.88%	76/78	24/26

In the 10th column above, the True Positive (the first number) is the number of documents that were correctly assigned to the rule. The Total (the second number) is the total positive.

In the 11th column above, the Valid True Positive (the first number) is the total number of remaining documents in the category. The Total (the second number) is the total number of documents remaining.

For more information about the Rules Obtained window or the **Text Rule Builder** node, see "The Text Rule Builder Node" on page 79, and the SAS Text Miner Help.

6. Close the Results window.

Chapter 7

Create Models and Compare Them

About the Tasks That You Will Perform

This section shows how to use **Decision Tree** nodes to create models, and compare them with a **Model Comparison** node.

A **Decision Tree** node can be used to classify observations based on the values of nominal, binary, or ordinal targets. It can also predict outcomes for interval targets or the appropriate decision when you specify decision alternatives. An empirical tree represents a segmentation of the data that is created by applying a series of simple rules. Each rule assigns an observation to a segment based on the value of one input.

One rule is applied after another, resulting in a hierarchy of segments within segments. The hierarchy is called a tree, and each segment is called a node. The original segment contains the entire data set and is called the root node of the tree. A node with all its successors forms a branch of the node that created it.

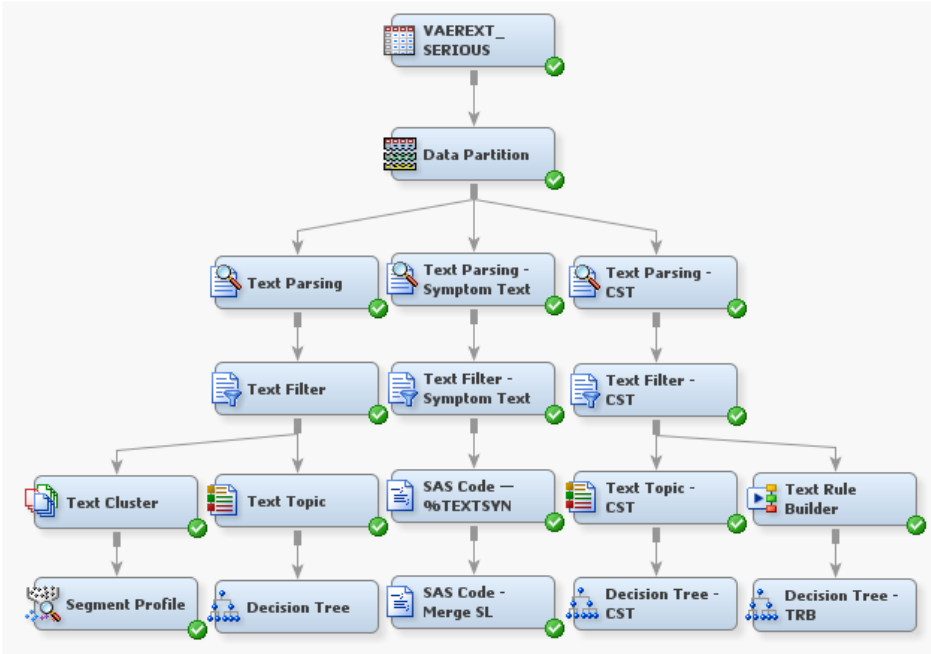
The final nodes are called leaves. For each leaf, a decision is made and applied to all observations in the leaf. The type of decision depends on the context. In predictive modeling, the decision is the predicted value. For more information about **Decision Tree** nodes, see the SAS Enterprise Miner help.

Create Models

To create models by using **Decision Tree** nodes:

1. Select the **Model** tab on the node toolbar and drag a **Decision Tree** node into the diagram workspace.
2. Connect the **Text Topic** node to the **Decision Tree** node.
3. Select the **Model** tab on the node toolbar and drag a **Decision Tree** node into the diagram workspace.
4. Right-click the **Decision Tree** node, and select **Rename**.

5. Enter *Decision Tree — CST* in the **Node Name** field, and then click **OK**.
6. Connect the **Text Topic — CST** node to the **Decision Tree — CST** node.
7. Select the **Model** tab on the node toolbar and drag a **Decision Tree** node into the diagram workspace.
8. Right-click the **Decision Tree** node, and select **Rename**.
9. Enter *Decision Tree — TRB* in the **Node Name** field, and then click **OK**.
10. Connect the **Text Rule Builder** node to the **Decision Tree — TRB** node.

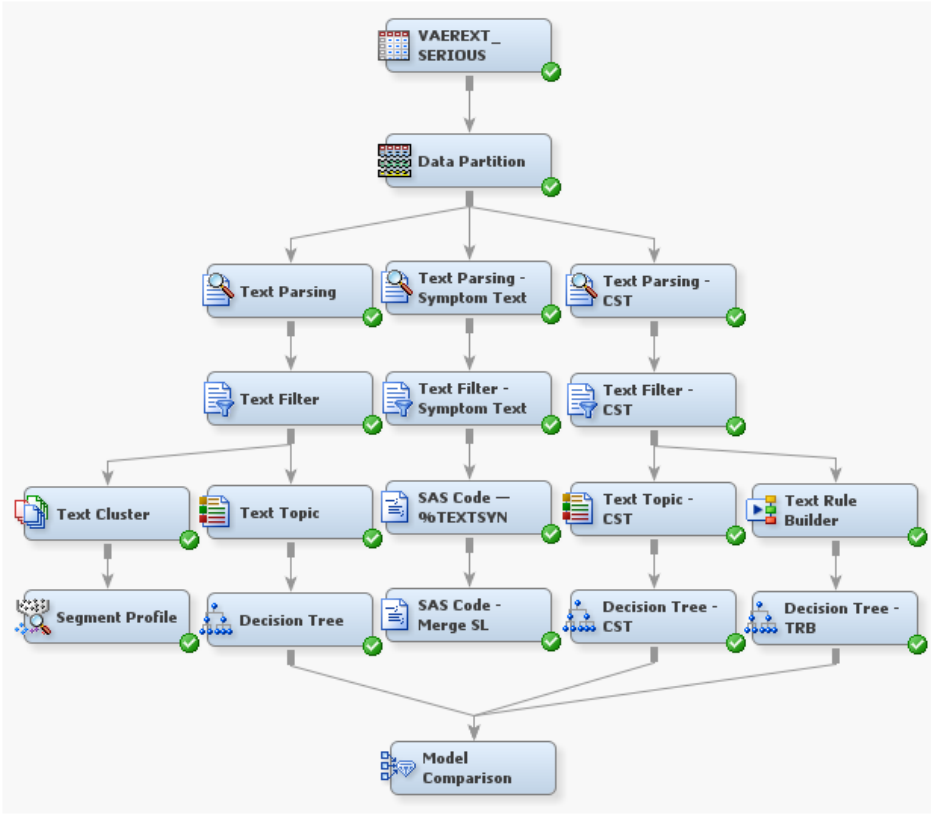


11. Right-click the **Decision Tree** node in the diagram workspace, and select **Run**. Click **Yes** in the Confirmation dialog box.
12. Click **Results** in the Run Status dialog box when the node has finished running.
13. Select the Tree window, and explore the tree that was obtained.
14. Close the Results window.
15. Right-click the **Decision Tree — CST** node in the diagram workspace, and select **Run**. Click **Yes** in the Confirmation dialog box.
16. Click **Results** in the Run Status dialog box when the node has finished running.
17. Select the Tree window, and explore the tree that was obtained. How does this tree differ from the previous tree? The primary difference is that different topics were used for each decision point.
18. Close the Results window.
19. Right-click the **Decision Tree — TRB** node in the diagram workspace, and select **Run**. Click **Yes** in the Confirmation dialog box.
20. Click **Results** in the Run Status dialog box when the node has finished running.
21. Select the Tree window, and explore the tree that was obtained. How does this tree differ from the previous two trees? The primary difference is that instead of topics, single or multi-term rules were used for each decision point.
22. Close the Results window.

Compare the Models

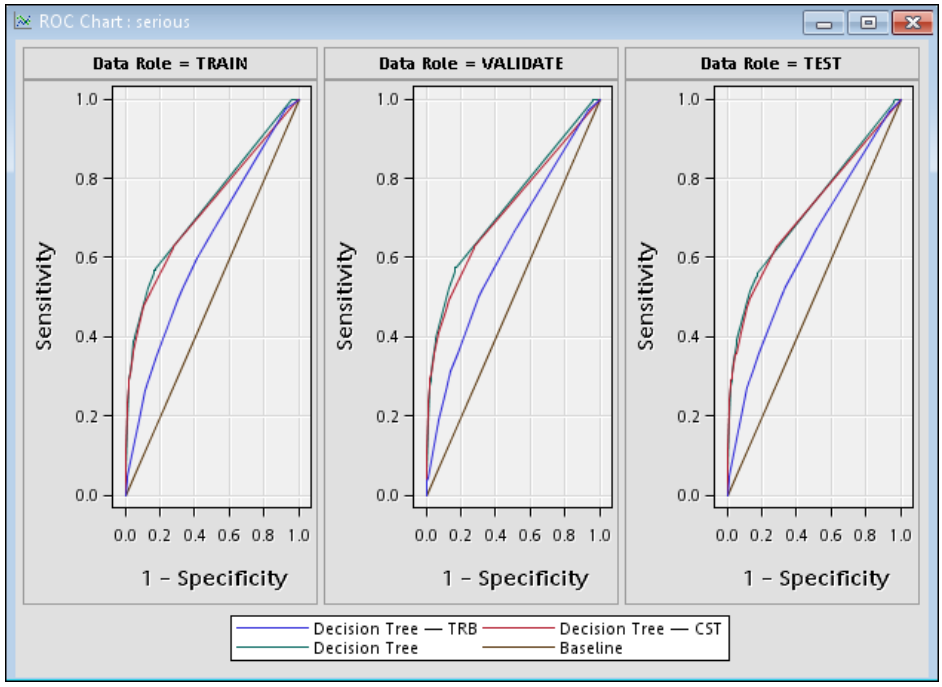
To compare models using a **Model Comparison** node:

1. Select the **Assess** tab on the node toolbar and drag a **Model Comparison** node into the diagram workspace.
2. Connect the **Decision Tree** node, the **Decision Tree — CST** node, and the **Decision Tree — TRB** node to the **Model Comparison** node.



3. Right-click the **Model Comparison** node in the diagram workspace, and select **Run**. Click **Yes** in the Confirmation dialog box.
4. Click **Results** in the Run Status dialog box when the node has finished running.

   The Results window appears.
5. Select the ROC Chart.



The greater the area under the curve, the better the model. The brown line represents the baseline to compare the models by. The blue line represents how well the **Decision Tree — TRB** model did at predicting the target SERIOUS. This model used input from the **Text Rule Builder** node. The green line represents how well the **Decision Tree** model did at predicting the target SERIOUS. The red line represents how well the **Decision Tree — CST** model did at predicting the target SERIOUS.

Both the **Decision Tree** and the **Decision Tree — CST** models performed better than the **Decision Tree — TRB** model. The **Decision Tree** and the **Decision Tree — CST** models performed similarly with respect to each other.

As an additional exercise, you could try modifying the number of multiple or single term topics in the **Text Topic** or **Text Topic — CST** nodes. Then rerun the **Decision Tree** and the **Decision Tree — CST** nodes to see whether the models have been improved.

Chapter 8

The Text Import Node

About the Text Import Node

The **Text Import** node serves as a replacement for an Input Data node by enabling you to create data sets dynamically from files contained in a directory or from the web. The **Text Import** node takes an import directory containing text files in potentially proprietary formats such as MS Word and PDF files as input. The tool traverses this directory and filters or extracts the text from the files, places a copy of the text in a plain text file, and a snippet (or possibly even all) of the text in a SAS data set. If a URL is specified, the node will crawl websites and retrieve files from the web and move them to the import directory before doing this filtering process. The output of a **Text Import** node is a data set that can be imported into the **Text Parsing** node.

In addition to filtering the text, the **Text Import** node can also identify the language that the document is in and take care of transcoding documents to the session encoding. For more on encoding and transcoding, see the SAS Text Miner and SAS Session Encoding topic in the SAS Text Miner Help.

The **Text Import** node relies on the SAS Document Conversion Server installed and running on a Windows machine. The machine must be accessible from the SAS Enterprise Miner server via the host name and port number that were specified at install time.

Note:

- If you run the **Text Import** node in a UTF-8 SAS session, then the node attempts to transcode all filtered text to UTF-8 encoding so that the result data set can be used in a UTF-8 SAS session. In all other SAS session encodings, the **Text Import** node does not transcode the data but instead assumes that the input data is in the same encoding as the SAS session. For more information, see the SAS Text Miner and SAS Session Encoding topic in the SAS Text Miner Help.
- The **Text Import** node is not supported for use in group processing (Start Groups and End Groups nodes).

For more information about the **Text Import** node, see the SAS Text Miner Help.

The rest of this chapter presents examples of how you can use the **Text Import** node.

Using the Text Import Node

Contents

You can use the **Text Import** node to import documents from a directory or the Web. See the following for examples of how to use the **Text Import** node. For information about creating a project and a diagram, see "Setting Up Your Project" on page 9.

Import Documents from a Directory

This example assumes that SAS Enterprise Miner is running, the SAS Document Conversion server is running, and a diagram workspace has been opened in a project. For information about creating a project and a diagram, see "Setting Up Your Project" on page 9.

Perform the following steps to import documents from a directory:

1. Select the **Text Mining** tab, and drag a **Text Import** node into the diagram workspace.

2. Click the ellipsis button next to the **Import File Directory** property of the **Text Import** node.

   A Select Server Directory dialog box opens.

3. Navigate to a folder that contains documents you want to create a data set from, select it, and then click **OK**.

   Note: To see the file types you want to select, you might need to select `All Files` in the type drop-down menu.

4. Click the ellipsis button next to the **Language** property.

   The Language dialog box opens.

5. Select one or more licensed languages in which to require the language identifier to assign each document's language, and then click **OK**.

6. (Optional) Specify the file types to process for the **Extensions** property.

   For example, if you want to look at only documents with a .txt and a .pdf extension, specify *.txt .pdf* for the **Extensions** property, and click **Enter** on your keyboard.

   Note: If you do not specify file types to process, the **Text Import** node will process all file types in the specified import file directory.

7. Right-click the **Text Import** node, and select **Run**.

8. Click **Yes** in the Confirmation dialog box.

9. Click **Results** in the Run Status dialog box when the node has finished running.

10. Examine results from the documents you imported.

    You can now use the **Text Import** node as an input data source for your text mining analysis.

11. Select the **Text Mining** tab, and drag a **Text Parsing** node into the diagram workspace.

12. Connect the **Text Import** node to the **Text Parsing** node.

13. Right-click the **Text Parsing** node, and select **Run**.

14. Click **Yes** in the Confirmation dialog box.

15. Click **Results** in the Run Status dialog box when the node has finished running.

Import Documents from the Web

This example assumes that SAS Enterprise Miner is running, the SAS Document Conversion server is running, and a diagram workspace has been opened in a project. For information about creating a project and a diagram, see "Setting Up Your Project" on page 9.

Note: Web crawling is supported only on Windows operating systems.

Perform the following steps to import documents from the Web:

1. Select the **Text Mining** tab, and drag a **Text Import** node into the diagram workspace.

2. Click the ellipses button for the **Import File Directory** property of the **Text Import** node.

   A Select Server Directory dialog box appears.

3. Navigate to a folder, select it, and then click **OK**.

   The documents are first written to the **Import File Directory** location. The files are processed from the **Import File Directory** location, and then are written to the **Destination Directory** location.

4. Type the uniform resource locator (URL) of a Web page you want to crawl in the **URL** property of the **Text Import** node. For example, try *www.sas.com*.

5. Type *1* as the number of levels to crawl in the **Depth** property.

6. Set the **Domain** property to **`Unrestricted`**.

Note: If you want to crawl a password-protected Web site, set the **Domain** property to **`Restricted`**, and provide a user name for the **User Name** property, and a password for the **Password** property.

7. Right-click the **Text Import** node and select **Run**.

8. Click **Yes** in the Confirmation dialog box.

9. Click **Results** in the Run Status dialog box when the node has finished running.

Chapter 9

The Text Parsing Node

About the Text Parsing Node

The **Text Parsing** node enables you to parse a document collection in order to quantify information about the terms that are contained therein. You can use the **Text Parsing** node with volumes of textual data such as e-mail messages, news articles, web pages, research papers, and surveys. For more information about the **Text Parsing** node, see the SAS Text Miner Help.

The rest of this chapter presents an example of how you can use the **Text Parsing** node.

Using the Text Parsing Node

This example shows you how to identify terms and their instances in a data set containing text using the **Text Parsing** node. This example assumes that SAS Enterprise Miner is running, and a diagram workspace has been opened in a project. For information about creating a project and a diagram, see "Setting Up Your Project" on page 9.

1. The SAS data set SAMPSIO.ABSTRACT contains the titles and text of abstracts from conferences. Create the ABSTRACT data source and add it to your diagram workspace. Set the Role value of the TEXT and TITLE variables to `Text`.

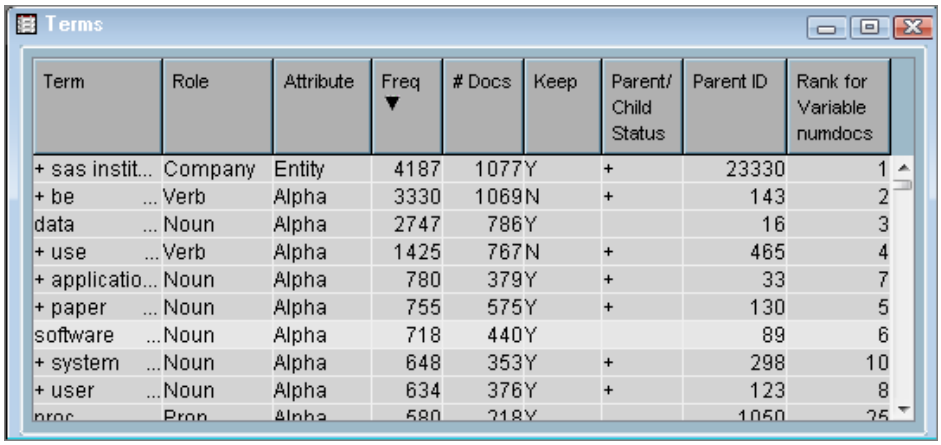
2. Select the **Text Mining** tab on the toolbar, and drag a **Text Parsing** node into the diagram workspace.

3. Connect the ABSTRACT data source to the **Text Parsing** node.

4. In the diagram workspace, right-click the **Text Parsing** node and select **Run**. Click **Yes** in the Confirmation dialog box that appears.

5. Click **Results** in the Run Status dialog box when the node finishes running. The Results window displays a variety of tabular and graphical output to help you analyze the terms and their instances in the ABSTRACT data source.

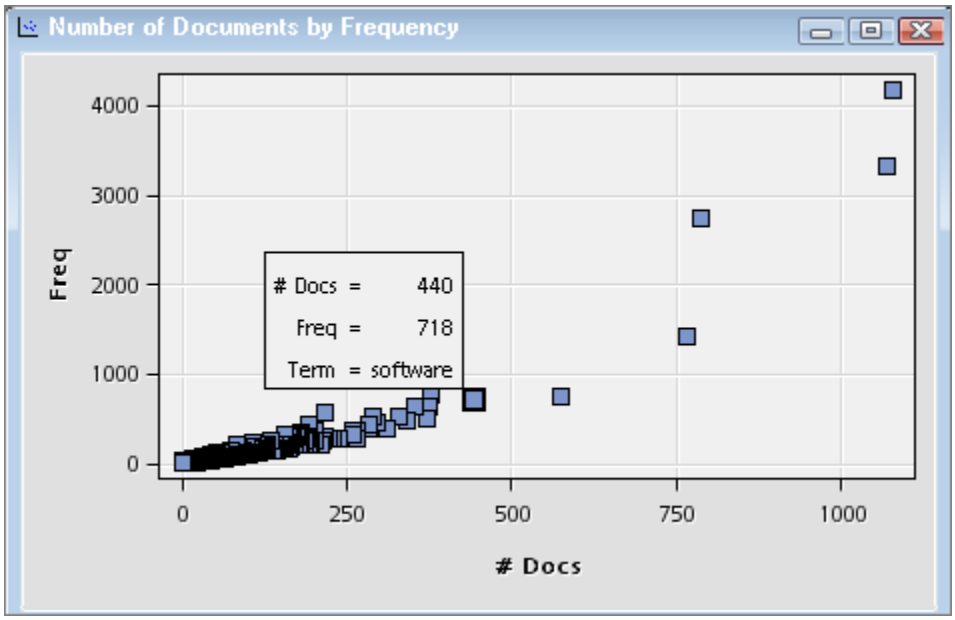
6. Sort the terms in the Terms table by frequency, and then select the term "software." As the Terms table illustrates, the term "software" is a noun that occurs in 440 documents in the ABSTRACT data source, and appears a total number of 718 times.



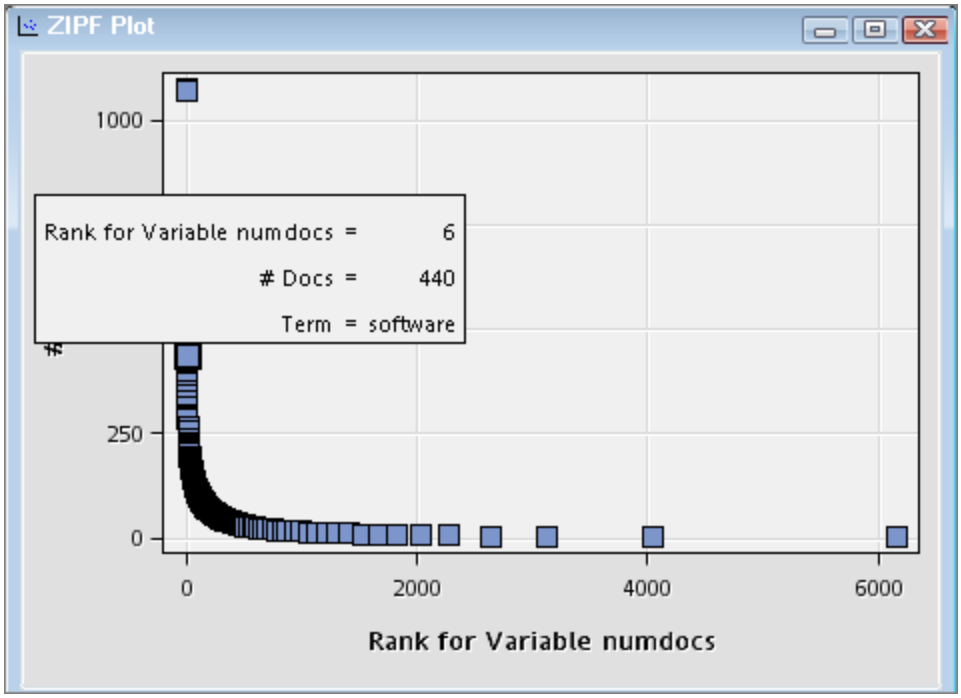
Term	Role	Attribute	Freq ▼	# Docs	Keep	Parent/ Child Status	Parent ID	Rank for Variable numdocs
+ sas instit...	Company	Entity	4187	1077	Y	+	23330	1
+ be ...	Verb	Alpha	3330	1069	N	+	143	2
data ...	Noun	Alpha	2747	786	Y		16	3
+ use ...	Verb	Alpha	1425	767	N	+	465	4
+ applicatio...	Noun	Alpha	780	379	Y	+	33	7
+ paper ...	Noun	Alpha	755	575	Y	+	130	5
software ...	Noun	Alpha	718	440	Y		89	6
+ system ...	Noun	Alpha	648	353	Y	+	298	10
+ user ...	Noun	Alpha	634	376	Y	+	123	8
proc	Prop	Alpha	680	218	Y		1060	25

When you select a term in the Terms table, the point corresponding to that term in the Text Parsing Results plots is highlighted.

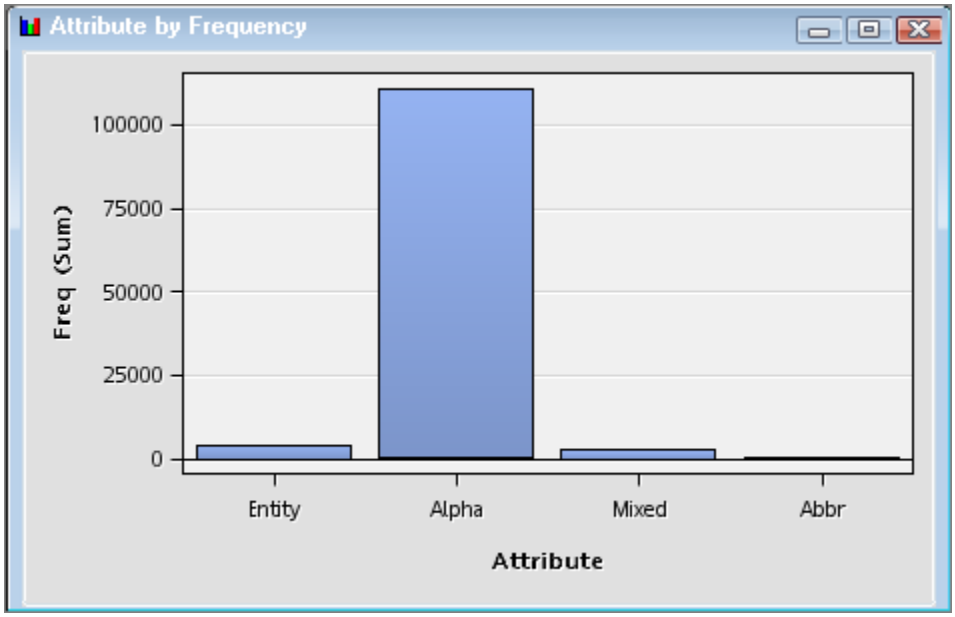
7. Select the Number of Documents by Frequency plot, and position the cursor over the highlighted point for information about the term "software."



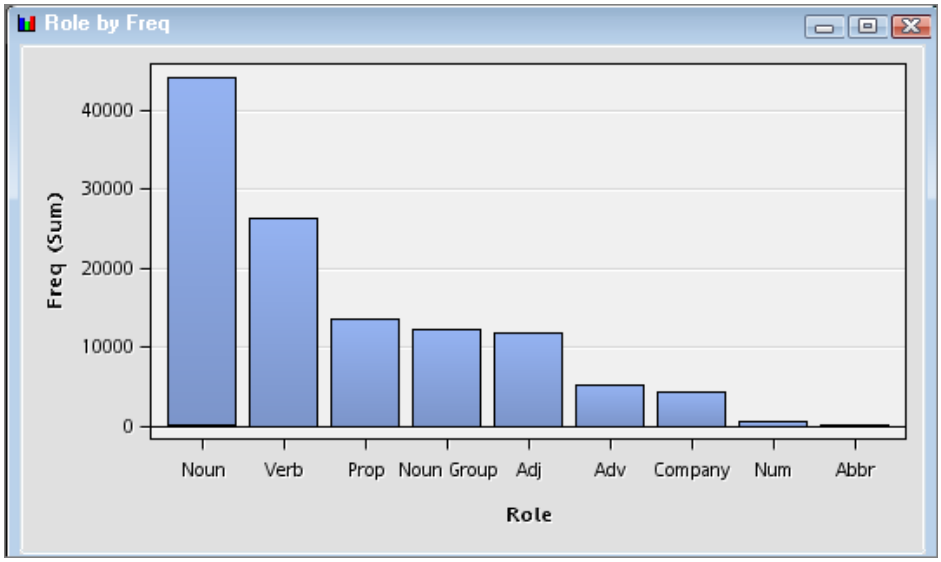
Similar information is also presented in a ZIPF plot.



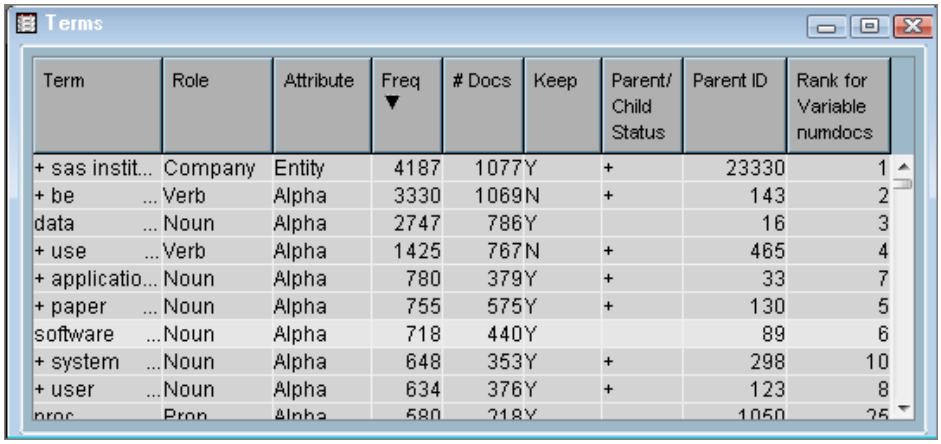
The Attribute by Frequency chart shows that **`Alpha`** has the highest frequency among attributes in the document collection.



The Role by Freq chart illustrates that **`Noun`** has the highest frequency among roles in the document collection.



8. Return to the Terms table, and notice that the term "software" is kept in the text parsing analysis. This is illustrated by the value of **`Y`** in the Keep column. Notice that not all terms are kept when you run the **Text Parsing** node with default settings.


Terms

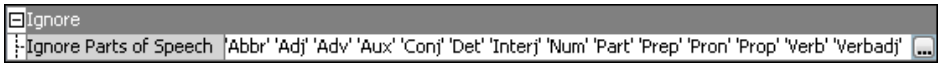
Term	Role	Attribute	Freq ▼	# Docs	Keep	Parent/ Child Status	Parent ID	Rank for Variable numdocs
+ sas instit...	Company	Entity	4187	1077	Y	+	23330	1
+ be ...	Verb	Alpha	3330	1069	N	+	143	2
data ...	Noun	Alpha	2747	786	Y		16	3
+ use ...	Verb	Alpha	1425	767	N	+	465	4
+ applicatio...	Noun	Alpha	780	379	Y	+	33	7
+ paper ...	Noun	Alpha	755	575	Y	+	130	5
software ...	Noun	Alpha	718	440	Y		89	6
+ system ...	Noun	Alpha	648	353	Y	+	298	10
+ user ...	Noun	Alpha	634	376	Y	+	123	8
proc ...	Prop	Alpha	580	218	Y		1060	25

The **Text Parsing** node not only enables you to gather statistical data about the terms in a document collection, but it also enables you to modify your output set of parsed terms by dropping terms that are a certain part of speech, type of entity, or attribute. Scroll down the list of terms in the Terms table and notice that many of the terms with a role other than **`Noun`** are kept. Let us assume that we want to limit our text parsing results to terms with a role of **`Noun`**.

9. Close the Results window.

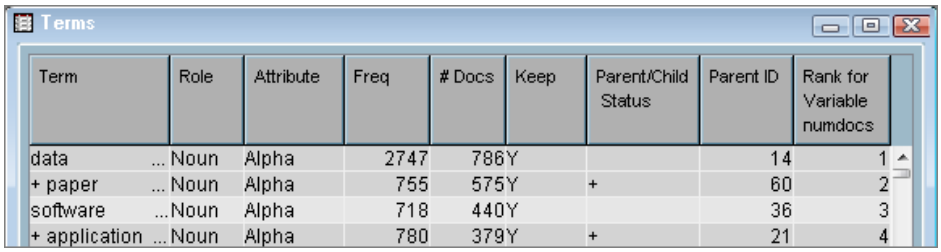
10. Select the **Text Parsing** node, and then select the ellipsis for the **Ignore Parts of Speech** property.

11. In the Ignore Parts of Speech dialog box, select all parts of speech except for **`Noun`** by holding Ctrl down on your keyboard and clicking on each option. Click **OK**. Notice that the value for the **Ignore Parts of Speech** property is updated with your selection.



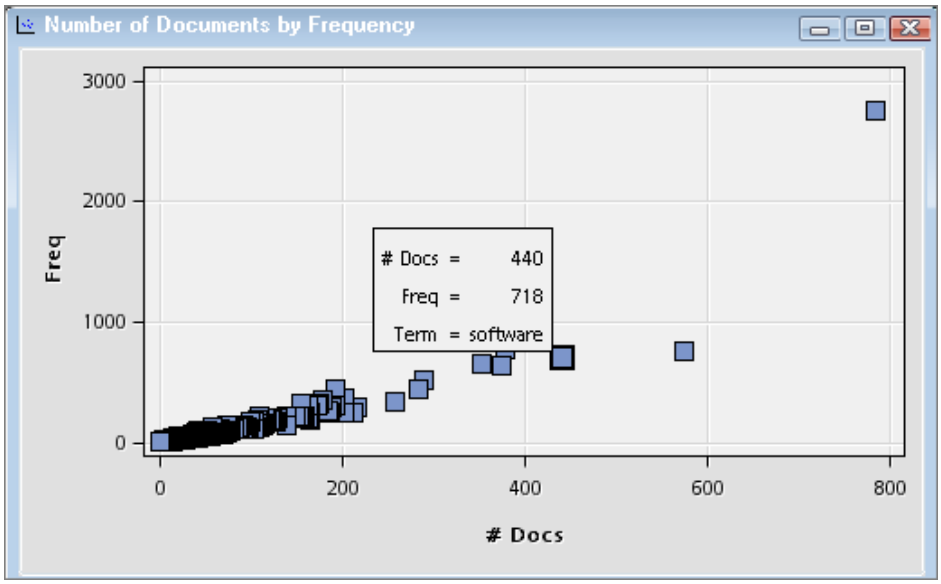
12. In addition to nouns, let us also keep noun groups. Set the **Noun Groups** property to **`Yes`**.

13. Right-click the **Text Parsing** node and select **Run**. Click **Yes** in the Confirmation dialog box that appears. Select **Results** in the Run Status dialog box when the node has finished running. Notice that the term "software" has a higher rank among terms with a role of just "noun" or "noun group" than it did when other roles were included. If you scroll down in the Terms table, you can see that just terms with a **`Noun`** or **`Noun Group`** role are included.

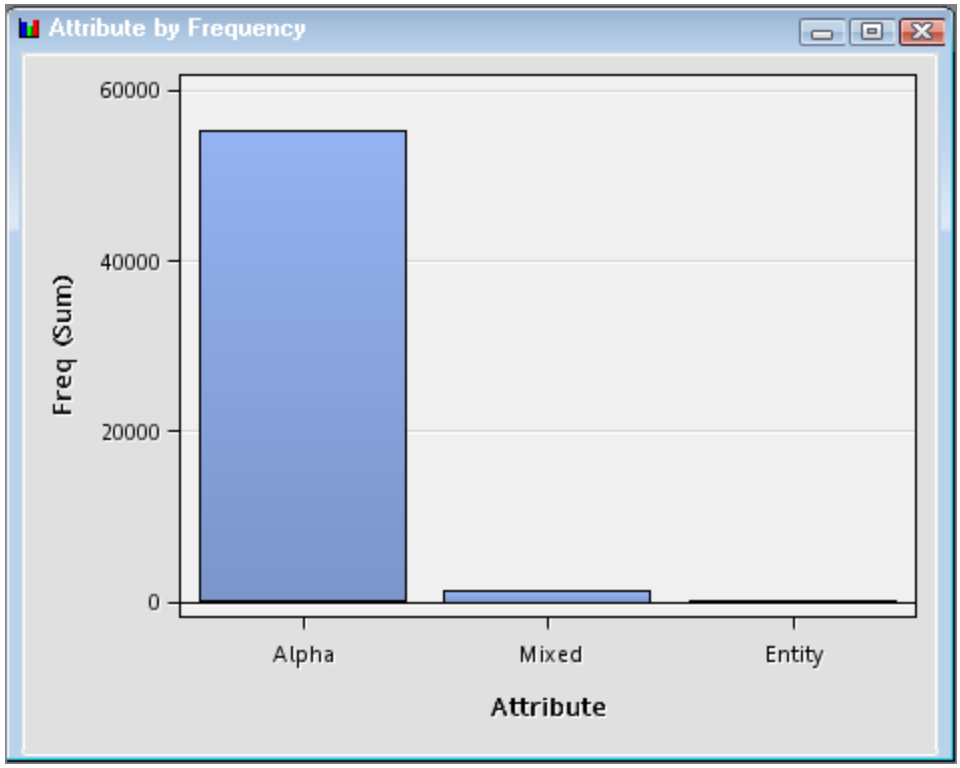

Terms

Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
data	Noun	Alpha	2747	786	Y		14	1
+ paper	Noun	Alpha	755	575	Y	+	60	2
software	Noun	Alpha	718	440	Y		36	3
+ application	Noun	Alpha	780	379	Y	+	21	4

As we would expect, there are fewer terms plotted in the Number of Documents by Frequency plot:



Similarly, the total number of terms in the output results with an attribute of **`Alpha`** has decreased, as can be seen in the Attribute by Frequency chart:


Attribute by Frequency
60000
40000
20000
0
Freq (Sum)
Alpha
Mixed
Entity
Attribute

Chapter 10

The Text Filter Node

About the Text Filter Node

You can use the **Text Filter** node to reduce the total number of parsed terms or documents that are analyzed. Therefore, you can eliminate extraneous information so that only the most valuable and relevant information is considered. For example, the **Text Filter** node can be used to remove unwanted terms and to keep only documents that discuss a particular issue. This reduced data set can be orders of magnitude smaller than the one representing the original collection that might contain hundreds of thousands of documents and hundreds of thousands of distinct terms.

For more information about the **Text Filter** node, see the SAS Text Miner Help.

The rest of this chapter presents an example of how you can use the **Text Filter** node.

Using the Text Filter Node

This example assumes that SAS Enterprise Miner is running, and a diagram workspace has been opened in a project. For information about creating a project and a diagram, see *Getting Started with SAS Enterprise Miner*.

The **Text Filter** node enables you to reduce the total number of terms in your text mining analysis. For example, common or infrequent words might not be useful to analyze, and can be filtered out. This example shows you how to filter out terms using the **Text Filter** node. This example assumes that you have performed "Using the Text Parsing Node" on page 51, and builds off the process flow diagram created there.

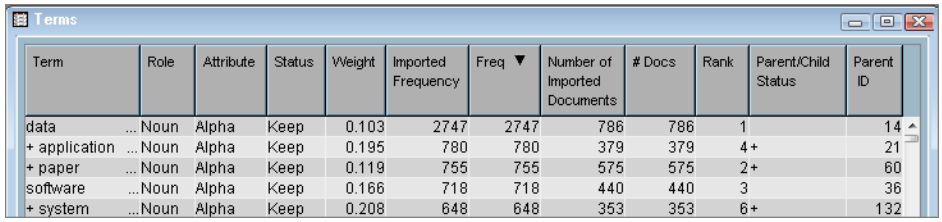
1. Select the **Text Mining** tab on the toolbar, and drag a **Text Filter** node into the diagram workspace.

2. Connect the **Text Parsing** node to the **Text Filter** node.

3. In the diagram workspace, right-click the **Text Filter** node and select **Run**. Click **Yes** in the Confirmation dialog box.

4. Click **Results** in the Run Status dialog box when the node finishes running.

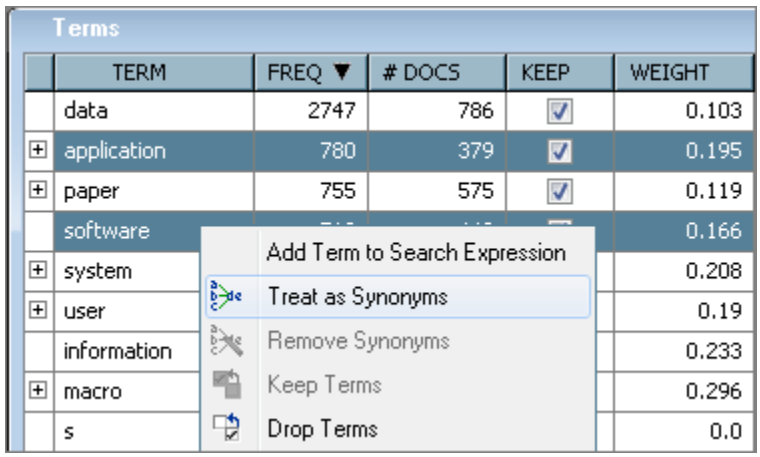
5. Select the Terms table. Sort the terms by frequency by clicking the Freq column heading.



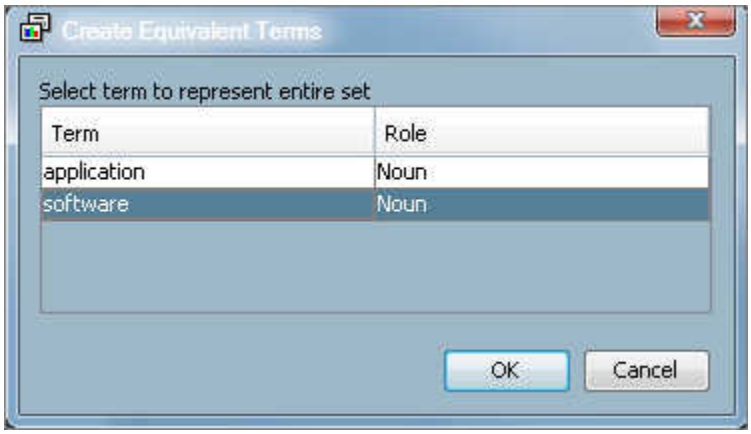
Assume that for purposes of our text mining analysis, we know that "software" and "application" are really used as synonyms in the documents that we want to analyze, and we want to treat them as the same term.

6. Close the Results window. Select the **Text Filter** node, and then click the ellipsis button for the **Filter Viewer** property.

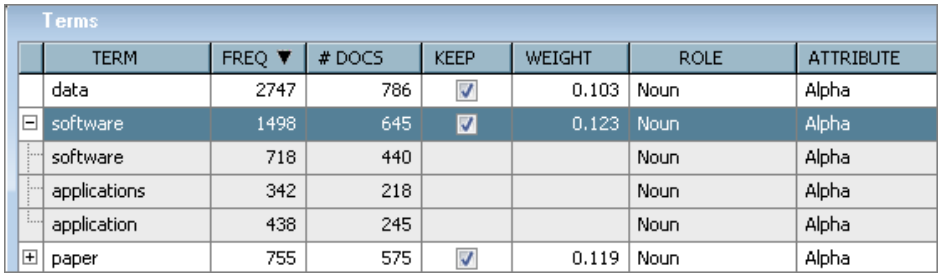
7. In the Interactive Filter Viewer sort the terms in the Terms table by frequency. Hold Ctrl down on your keyboard, select "software" and "application", and then right-click "software" and select **Treat as Synonyms** from the drop-down menu.



8. In the Create Equivalent Terms dialog box, select **software** as the term to represent both terms in the Terms table.



9. Click **OK** in the Create Equivalent Terms dialog box. Notice that the term "software" now represents both terms in the Terms table. Expand the term "software".

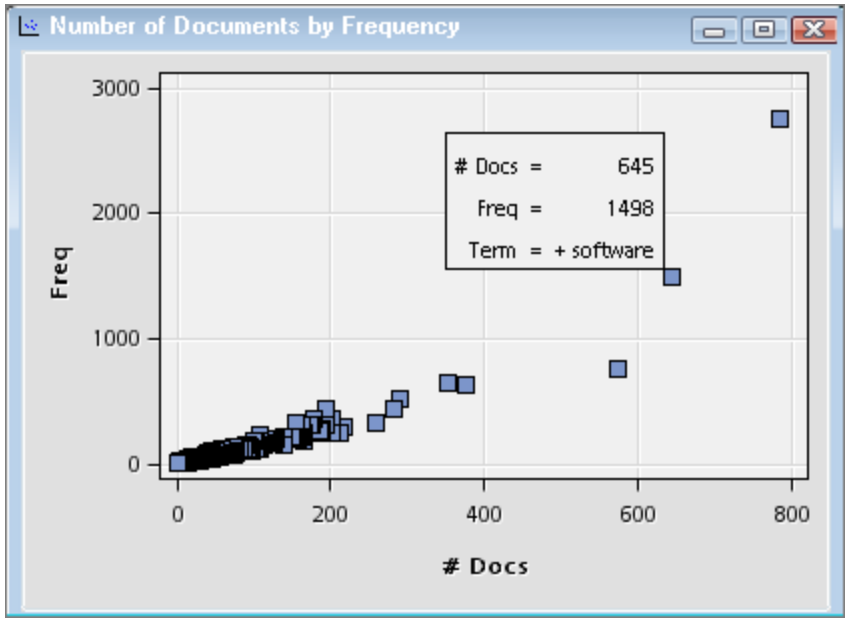


	TERM	FREQ ▼	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
	data	2747	786	☑	0.103	Noun	Alpha
⊟	software	1498	645	☑	0.123	Noun	Alpha
	software	718	440			Noun	Alpha
	applications	342	218			Noun	Alpha
	application	438	245			Noun	Alpha
⊞	paper	755	575	☑	0.119	Noun	Alpha

10. Close the Interactive Filter Viewer. When prompted whether you would like to save your changes, select **Yes**.

11. Right-click the **Text Filter** node, and select **Run**. Select **Yes** in the Confirmation dialog box. Select **Results** in the Run Status dialog box when the node has finished running.

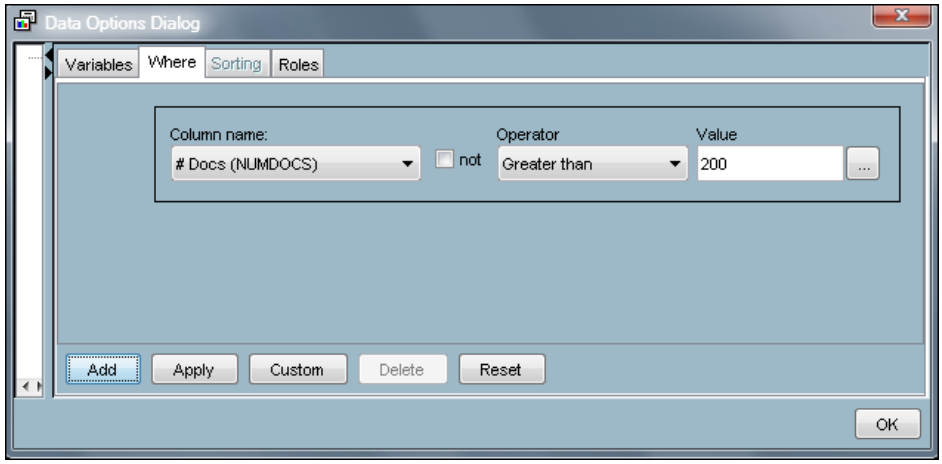
12. Select the Number of Documents by Frequency plot to see how both terms are now treated as the same.



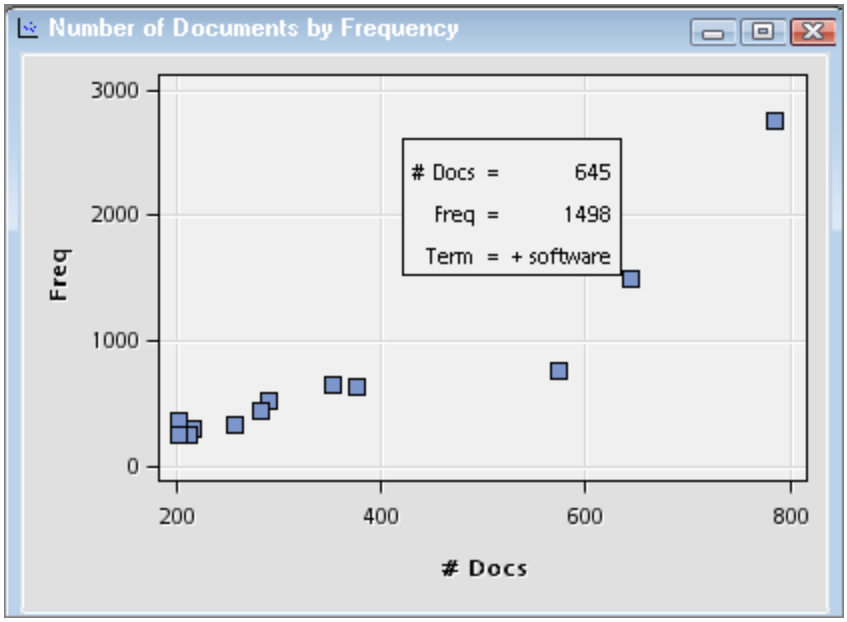
You can also use options to change your view or specify a subset of results to appear in a plot. For example, consider that you want to refine this plot to only show terms that appear in more than 200 documents.

13. Right-click the Number of Documents by Frequency plot, and select **Data Options**.

14. Select the **Where** tab in the Data Options Dialog box. Select `# Docs` from the **Column name** drop-down menu. Select `Greater than` from the **Operator** drop-down menu. Type *200* in the **Value** text box.

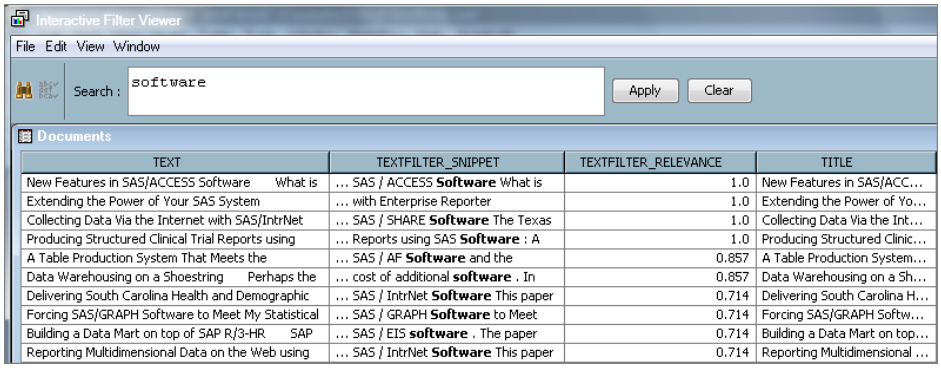


15. Click **Apply**, and then click **OK**. The Number of Documents by Frequency plot resizes and includes only terms that occur in more than 200 documents.



16. Close the Results window. In addition to resizing or subsetting a plot to help focus your analysis, you can also directly search for terms using the Interactive Filter Viewer.

17. Select the **Text Filter** node, and then click the ellipsis button for the **Filter Viewer** property. In the Interactive Filter Viewer, type *software* in the **Search** text box, and click **Apply**.



TEXT	TEXTFILTER_SNIPPET	TEXTFILTER_RELEVANCE	TITLE
New Features in SAS/ACCESS Software What is	... SAS / ACCESS **Software** What is	1.0	New Features in SAS/ACC...
Extending the Power of Your SAS System	... with Enterprise Reporter	1.0	Extending the Power of Yo...
Collecting Data Via the Internet with SAS/IntrNet	... SAS / SHARE **Software** The Texas	1.0	Collecting Data Via the Int...
Producing Structured Clinical Trial Reports using	... Reports using SAS **Software** : A	1.0	Producing Structured Clinic...
A Table Production System That Meets the	... SAS / AF **Software** and the	0.857	A Table Production System...
Data Warehousing on a Shoestring Perhaps the	... cost of additional **software** . In	0.857	Data Warehousing on a Sh...
Delivering South Carolina Health and Demographic	... SAS / IntrNet **Software** This paper	0.714	Delivering South Carolina H...
Forcing SAS/GRAPH Software to Meet My Statistical	... SAS / GRAPH **Software** to Meet	0.714	Forcing SAS/GRAPH Softw...
Building a Data Mart on top of SAP R/3-HR SAP	... SAS / EIS **software** . The paper	0.714	Building a Data Mart on top...
Reporting Multidimensional Data on the Web using	... SAS / IntrNet **Software** This paper	0.714	Reporting Multidimensional ...

The Documents table provides a snippet of text that includes the term that you are searching for. You can use information in the Documents table to help you understand the context in which a term is being used by examining the snippet result in addition to the full text and title of the document. For more information about the Interactive Filter Viewer, see the Interactive Filter Viewer topic in the SAS Text Miner Help.

Searching for a term in the Interactive Filter Viewer raises an interesting problem. As shown above, a search for “software” is case insensitive. However, what if there are instances of a term that we want to find that are misspelled in the document collection? You can also check for spelling when filtering terms using a dictionary data set.

18. Close the Interactive Filter Viewer, and select **No** when prompted for whether you want to save changes.

19. (Optional) Select the **Text Filter** node, and set the **Check Spelling** property to **`Yes`**. When you rerun the **Text Filter** node, terms will be checked for misspellings. You can also specify a data set to use in spell-checking by clicking the ellipsis button for the **Dictionary** property and selecting a data set. For information about creating a dictionary data set, see the How to Create a Dictionary Data Set topic in the SAS Text Miner Help.

    Right-click the **Text Filter** node, and select **Run**. Select **Yes** in the Confirmation dialog box. When the node finishes running, select **OK** in the Run Status dialog box. Click the ellipsis button for the **Spell-Checking Results** property to access a window in which you can view the data set that contains spelling corrections generated during spell-checking. For example, the term "softwae" is identified as a misspelling of the term "software."

	Parent # Docs	Term	# Docs	Parent	Role	Parent Role	Min Distance	Dictionary	Key	Parent ID
1	245.0	applicaion	1.0	application	Noun	Noun	5.0		7823.0	21.0
2	119.0	procedural	2.0	procedure	Noun	Noun	14.0		7850.0	33.0
3	440.0	solftware	1.0	software	Noun	Noun	6.0		879.0	36.0
4	440.0	softwae	2.0	software	Noun	Noun	7.0		5694.0	36.0
5	33.0	entr	1.0	entry	Noun	Noun	8.0		6071.0	58.0
6	5.0	dependence	1.0	dependent	Noun	Noun	14.0		11860.0	80.0
7	83.0	suport	1.0	support	Noun	Noun	4.0		5358.0	82.0
8	22.0	succe	1.0	success	Noun	Noun	14.0		4490.0	87.0
9	76.0	agility	1.0	ability	Noun	Noun	14.0		4974.0	88.0
10	13.0	enduser	1.0	end-user	Noun	Noun	7.0		10790.0	89.0

You can see this relationship in the Terms table in the Interactive Filter Viewer. Click the ellipsis button for the **Filter Viewer** property. Expand the term "software" in the Terms table to view its synonyms. The synonyms include "softwae," which was identified as a misspelled term during spell-checking.

Terms

	TERM	FREQ ▼	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
	data	2747	786	☑	0.103	Noun	Alpha
⊟	software	1502	646	☑	0.123	Noun	Alpha
	applicaion	1	1			Noun	Alpha
	softwae	2	2			Noun	Alpha
	solftware	1	1			Noun	Alpha
	application	438	245			Noun	Alpha
	software	718	440			Noun	Alpha
	applications	342	218			Noun	Alpha
⊞	paper	755	575	☑	0.119	Noun	Alpha
⊞	system	648	353	☑	0.208	Noun	Alpha

Notice that the synonyms also include "application," which was created in steps 7-10 of this example, and "applicaion," which was identified during spell-checking as a misspelling of "application."

Chapter 11

The Text Topic Node

About the Text Topic Node

The **Text Topic** node enables you to explore the document collection by automatically associating terms and documents according to both discovered and user-defined topics. Topics are collections of terms that describe and characterize a main theme or idea. The approach is different from clustering because clustering assigns each document to a unique group while the **Text Topic** node assigns a score for each document and term to each topic. Then thresholds are used to determine if the association is strong enough to consider that the document or term belongs to the topic. As a result, documents and terms may belong to more than one topic or to none at all. The number of topics that you request should be directly related to the size of the document collection (for example, a large number for a large collection).

The most memory-intensive task is computing the singular value decomposition (SVD) of the term-by-document frequency matrix. For more information, see the Singular Value Decomposition topic in the SAS Text Miner Help. When in-memory resources are limited, the **Text Topic** node might use, instead of the full collection, a simple random sample of the documents in an attempt to run the node successfully. Sampling occurs only if the node encounters a memory failure during an attempt to compute the SVD without sampling. Furthermore, because sampling generally occurs when the document collection is extremely large, there is typically not an adverse effect on modeling results. Exactly when sampling occurs depends on a number of parameters including the size of your collection, the platform on which your system is running, and the available RAM.

For more information about the **Text Topic** node, see the SAS Text Miner Help.

Note: The **Text Topic** node is not supported for use in group processing (Start Groups and End Groups nodes).

Using the Text Topic Node

This example assumes that SAS Enterprise Miner is running, and a diagram workspace has been opened in a project. For information about creating a project and a diagram, see *Getting Started with SAS Enterprise Miner*.

The **Text Topic** node enables you to create topics of interest from a list of terms. The goal in creating a list of topics is to establish combinations of words that you are interested in analyzing. For example, you might be interested in mining articles that discuss the activities of a "company president." One way to approach this task is to look at all articles that have the term "company," and all articles that have the term "president." The **Text Topic** node enables you to combine the terms "company" and "president" into the topic "company president."

The ability to combine individual terms into topics can improve your text mining analysis. Through combining, you can narrow the amount of text that is subject to analysis to specific groupings of words that you are interested in. This example shows you how to create topics using the **Text Topic** node.

1. The SAS data set SAMPSIO.ABSTRACT contains the titles and text of abstracts from conferences. Create the ABSTRACT data source and add it to your diagram workspace. Set the Role value of the TEXT and TITLE variables to **`Text`**.

2. Select the **Text Mining** tab on the toolbar, and drag a **Text Parsing** node into the diagram workspace.

3. Connect the ABSTRACT data source to the **Text Parsing** node.

4. Select the **Text Parsing** node, and then select the ellipsis for the **Ignore Parts of Speech** property.

5. In the Ignore Parts of Speech dialog box, select all parts of speech except for **`Noun`** by holding down Ctrl on your keyboard and clicking each option. Click **OK**.

6. Set the **Noun Groups** property to **`Yes`**.

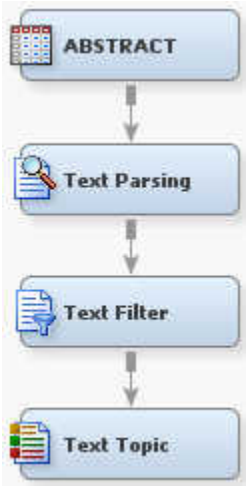
7. Select the **Text Mining** tab on the toolbar, and drag a **Text Filter** node into the diagram workspace.

8. Connect the **Text Parsing** node to the **Text Filter** node.

9. Select the **Text Mining** tab on the toolbar, and drag a **Text Topic** node into the diagram workspace.

10. Connect the **Text Filter** node to the **Text Topic** node.

    Your process flow diagram should resemble the following:



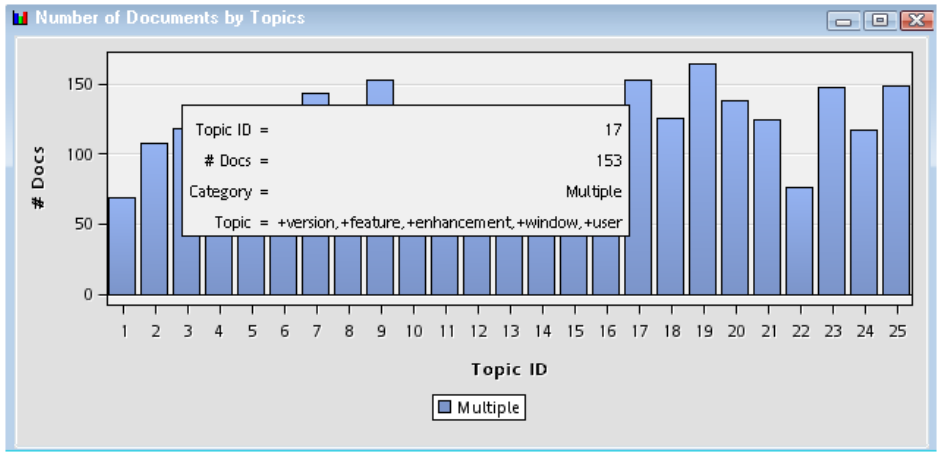
11. In the diagram workspace, right-click the **Text Topic** node and select **Run**. Click **Yes** in the Confirmation dialog box that appears. Click **Results** in the Run Status dialog box when the node finishes running.

12. Select the Topics table to view the topics that have been created with a default run of the **Text Topic** node.

Topics

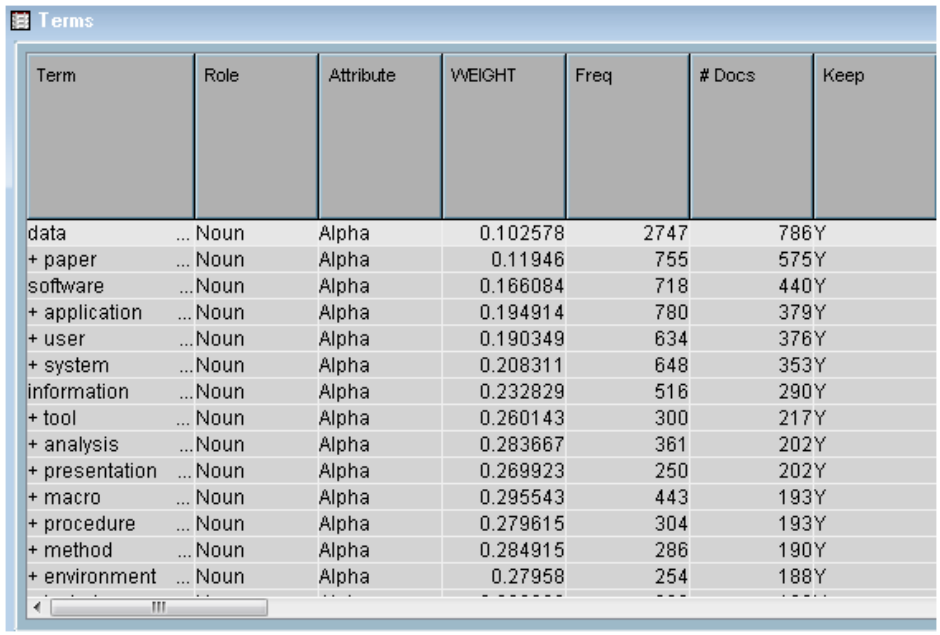
Category	Topic ID ▲	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Multiple	1	0.112	0.035	+select statement,+...	56	69
Multiple	2	0.100	0.037	+customer,+busine...	86	108
Multiple	3	0.103	0.037	+test,+sample,+met...	119	118
Multiple	4	0.104	0.036	+graph,graph softw...	71	100
Multiple	5	0.124	0.035	+macro,+macro vari...	46	110
Multiple	6	0.095	0.037	+application,+devel...	105	134
Multiple	7	0.104	0.037	web,+page,+brows...	74	143
Multiple	8	0.105	0.037	+output,+system,+te...	84	117
Multiple	9	0.121	0.037	+data set,+value,+v...	82	153
Multiple	10	0.102	0.037	+server,+performan...	81	125
Multiple	11	0.076	0.038	+performance,+pro...	114	107
Multiple	12	0.113	0.036	+data warehouse,+...	60	112
Multiple	13	0.106	0.036	+report,print,+tool,+...	85	126
Multiple	14	0.108	0.037	java,+technology,+a...	88	111
Multiple	15	0.089	0.038	+analysis,+develop...	120	126
Multiple	16	0.093	0.037	+group,+treatment,+...	94	104
Multiple	17	0.102	0.037	+version,+feature,+...	102	153
Multiple	18	0.106	0.037	+decision,support,i...	90	126
Multiple	19	0.103	0.038	+programmer,+tech...	98	165
Multiple	20	0.096	0.037	+performance,+data...	97	138
Multiple	21	0.086	0.037	+table,power,+proc...	106	125
Multiple	22	0.096	0.037	+entry,+catalog,+cat...	85	76
Multiple	23	0.085	0.038	health,care,+system...	127	148
Multiple	24	0.092	0.037	+model,+variable,re...	109	117
Multiple	25	0.102	0.037	+mainframe,+progr...	101	149

13. Select the Number of Documents by Topics chart to see a topic by the number of documents that it contains.



Note: You might need to resize the default graph to see the topic ID values.

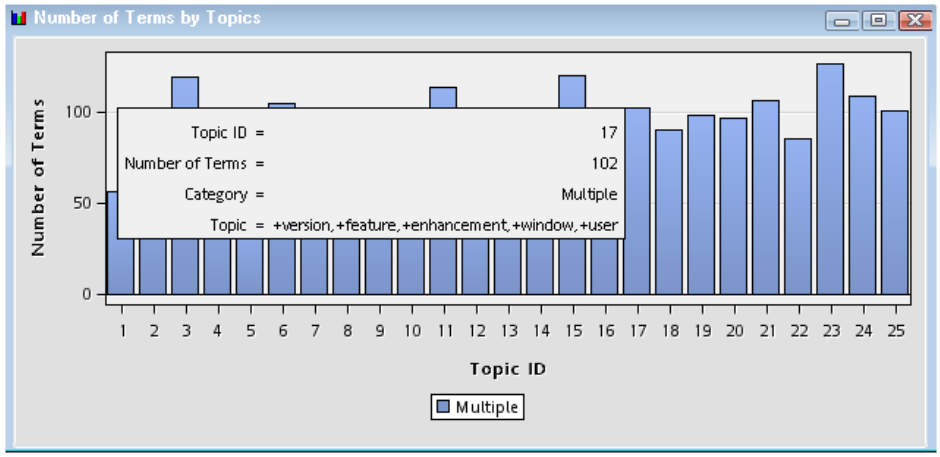
14. Select the **Terms** table. Select the first entry in the table.



Term	Role	Attribute	WEIGHT	Freq	# Docs	Keep
data	Noun	Alpha	0.102578	2747	786	Y
+ paper	Noun	Alpha	0.11946	755	575	Y
software	Noun	Alpha	0.166084	718	440	Y
+ application	Noun	Alpha	0.194914	780	379	Y
+ user	Noun	Alpha	0.190349	634	376	Y
+ system	Noun	Alpha	0.208311	648	353	Y
information	Noun	Alpha	0.232829	516	290	Y
+ tool	Noun	Alpha	0.260143	300	217	Y
+ analysis	Noun	Alpha	0.283667	361	202	Y
+ presentation	Noun	Alpha	0.269923	250	202	Y
+ macro	Noun	Alpha	0.295543	443	193	Y
+ procedure	Noun	Alpha	0.279615	304	193	Y
+ method	Noun	Alpha	0.284915	286	190	Y
+ environment	Noun	Alpha	0.27958	254	188	Y

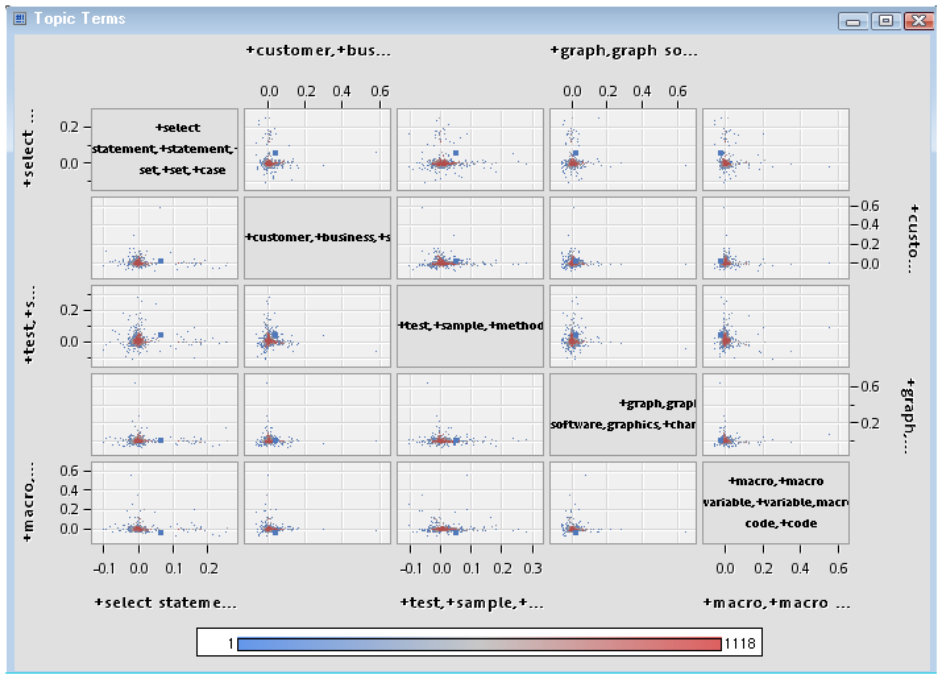
The **Terms** table shows terms and their weights for each topic. Notice that all kept terms have a role of **`Noun`** or **`Noun Group`**.

15. Select the **Number of Terms by Topics** bar chart.



If you position the mouse pointer over a bar, then a tooltip indicates the topic ID, the number of terms included in that topic, the category, and the topic.

16. Select the **Topic Terms** matrix graph.



The **Topic Terms** matrix graph shows the topic values across terms.

Note: You might need to expand this matrix to see the points more clearly.

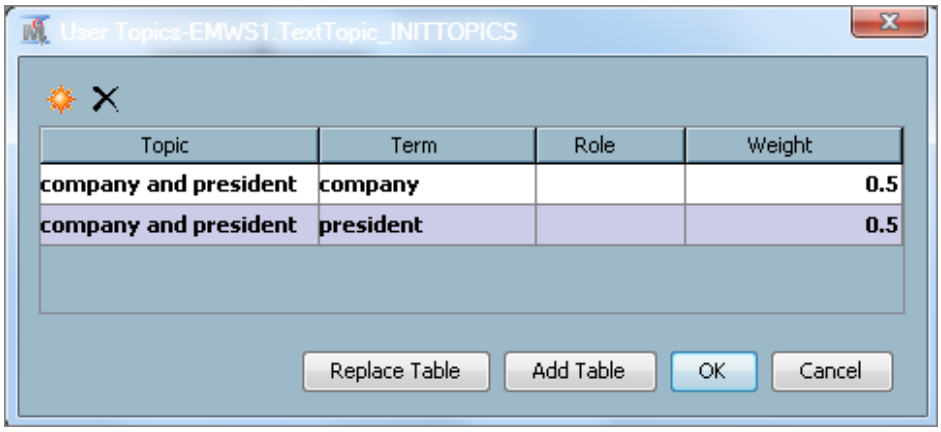
In addition to multi-term topics, you can use the **Text Topic** node to create single-term topics or to create your own topics.

17. Close the Results window, and select the **Text Topic** node.

18. Select the **Number of Single-term Topics** property, type *10*, and press **Enter** on your keyboard.

19. Click the ellipsis button for the **User Topics** property.

20. In the User Topics dialog box, click  to add a row. Enter the term *company*, give it a weight of *0.5*, and specify the topic *company and president*. Click  again to add a second row. Enter the term *president*, give it a weight of *0.5*, and specify the topic *company and president*.



21. Click **OK**.

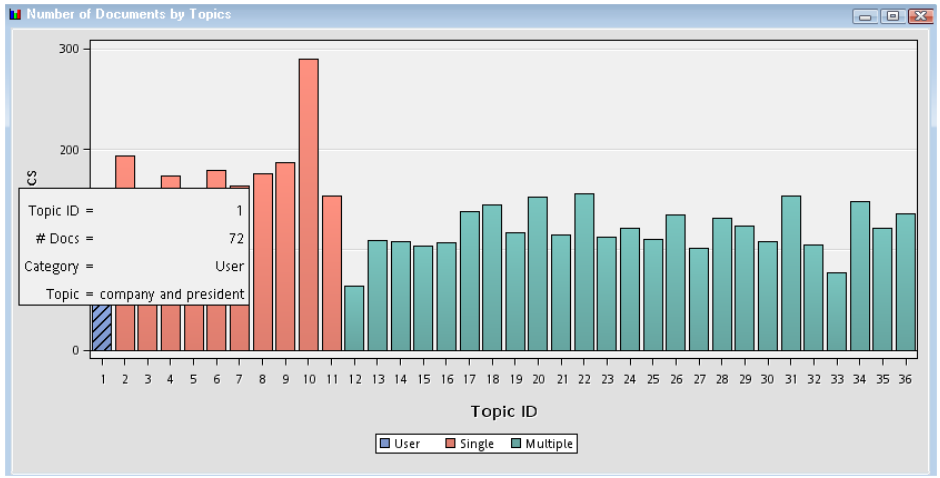
22. Right-click the **Text Topic** node and select **Run**. Select **Yes** in the Confirmation dialog box, and then **Results** in the Run Status dialog box when the node finishes running.

23. Select the Topics table. Notice that 10 new single-term topics have been created along with the topic that you specified in the User Topics dialog box.

Topics

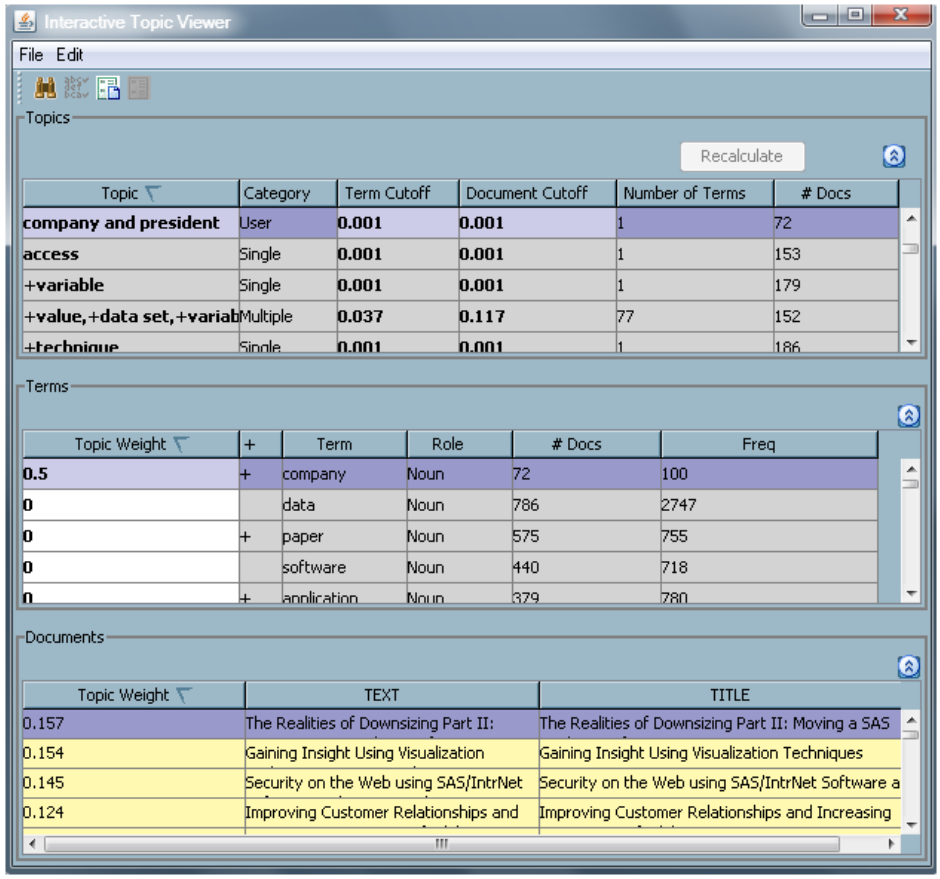
Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
User	1	0.001	0.001	company and president	1	72
Single	2	0.001	0.001	+macro	1	193
Single	3	0.001	0.001	+report	1	156
Single	4	0.001	0.001	+data set	1	173
Single	5	0.001	0.001	+server	1	130
Single	6	0.001	0.001	+variable	1	179
Single	7	0.001	0.001	web	1	163
Single	8	0.001	0.001	+program	1	176
Single	9	0.001	0.001	+technique	1	186
Single	10	0.001	0.001	information	1	290
Single	11	0.001	0.001	access	1	153
Multiple	12	0.111	0.035	+statement,+select stat...	54	64
Multiple	13	0.079	0.038	+program,+time,+date,+...	104	109
Multiple	14	0.101	0.037	+method,+sample,+test...	116	108
Multiple	15	0.102	0.036	+graph,graph software,...	68	103
Multiple	16	0.119	0.035	+macro,+macro variable...	48	107
Multiple	17	0.099	0.037	+application,+developm...	109	138
Multiple	18	0.104	0.037	web,+page,+browser,+s...	83	144
Multiple	19	0.105	0.037	+output,+system,+templ...	86	117
Multiple	20	0.117	0.037	+value,+data set,+variab...	77	152
Multiple	21	0.101	0.037	+performance,+server,+...	103	115
Multiple	22	0.104	0.037	+programmer,+techniqu...	95	155
Multiple	23	0.116	0.036	+data warehouse,+ware...	66	112
Multiple	24	0.106	0.036	+report,print,+tool,+solut...	78	121
Multiple	25	0.108	0.037	java,+technology,+apple...	80	110
Multiple	26	0.096	0.038	+analysis,+developmen...	107	134
Multiple	27	0.090	0.037	+group,+treatment,+stu...	87	101
Multiple	28	0.100	0.037	+feature,+version,+enha...	113	131
Multiple	29	0.105	0.037	+decision,support,infor...	88	123
Multiple	30	0.091	0.037	+model,regression,+var...	95	108
Multiple	31	0.093	0.038	+database,data,+data s...	119	153
Multiple	32	0.081	0.037	+table,power,+worksho...	99	105
Multiple	33	0.096	0.037	+entry,+catalog,+catalog...	81	77
Multiple	34	0.086	0.038	health,care,+system,inf...	134	148
Multiple	35	0.088	0.037	+client,+server,+window...	94	121
Multiple	36	0.102	0.037	+mainframe,+pc,+file,+u...	97	136

24. Select the Number of Documents by Topics window to see the multi-term, single-term, and user-created topics by the number of documents that they contain.



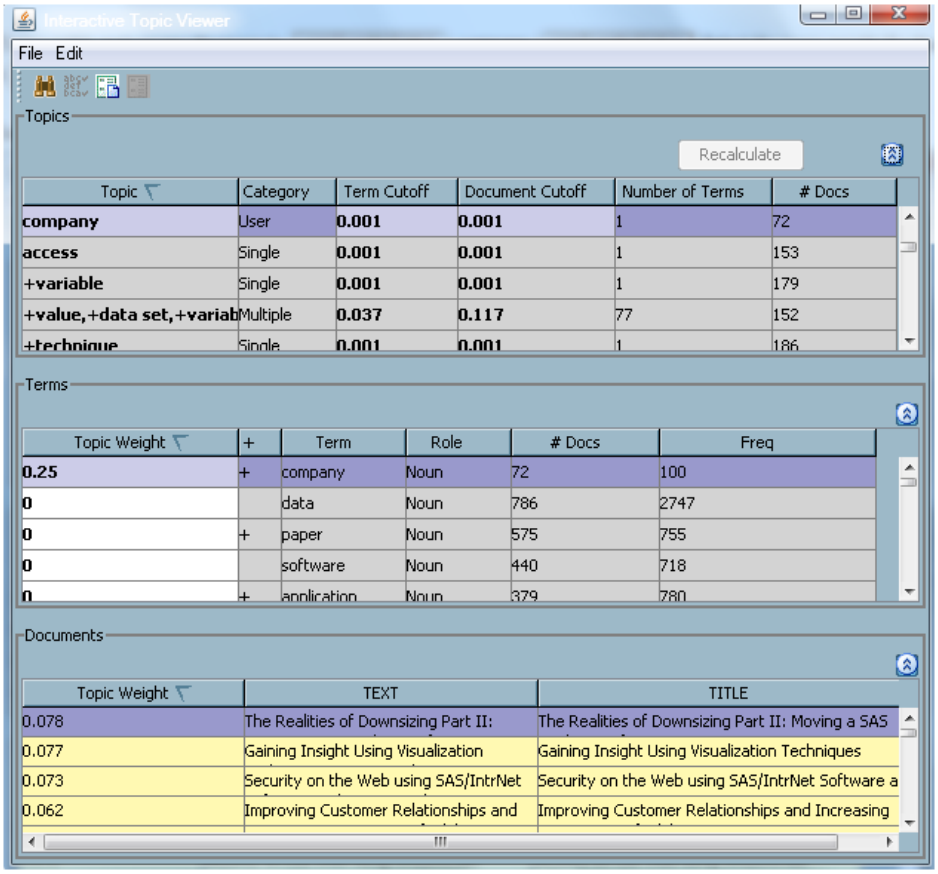
You can use the Interactive Topic Viewer to view and modify topic properties.

25. Close the Results window, and select the **Text Topic** node. Click the ellipsis button for the **Topic Viewer** property. When the Interactive Topic Viewer window appears, sort by the **Topic** column in the **Topics** pane.



In the Interactive Topic Viewer, you can change the topic name, term and document cutoff values, and the topic weight.

26. Select the topic value “company and president” in the Topics table and rename the topic to *company*. Select the topic weight for the term “company” in the Terms table, and change it to *0.25*. Click **Recalculate**.



27. Close the Interactive Topic Viewer, and select **No** when prompted for whether you want to save your changes. For more information about the Interactive Topic Viewer, see the Interactive Topic Viewer topic in the SAS Text Miner Help.

Chapter 12

The Text Cluster Node

About the Text Cluster Node

The **Text Cluster** node clusters documents into disjoint sets of documents and reports on the descriptive terms for those clusters. Two algorithms are available. The Expectation Maximization algorithm clusters documents with a flat representation, and the Hierarchical clustering algorithm groups clusters into a tree hierarchy. Both approaches rely on the singular value decomposition (SVD) to transform the original weighted, term-document frequency matrix into a dense but low dimensional representation.

The most memory-intensive task of the text clustering process is computing the SVD of the weighted term-by-document frequency matrix. When in-memory resources are limited, the node might use, instead of the full collection, a simple random sample of the documents in an attempt to run the node successfully. Sampling occurs only if the node encounters a memory failure during an attempt to compute the SVD without sampling. Furthermore, because sampling generally occurs when the document collection is extremely large, there is typically not an adverse effect on modeling results. Exactly when sampling occurs depends on a number of parameters including the size of your collection, the platform on which your system is running, and the available RAM.

For more information about the **Text Cluster** node, see the SAS Text Miner Help.

The rest of this chapter presents an example of how you can use the **Text Cluster** node.

Using the Text Cluster Node

This example uses the **Text Cluster** node to cluster SAS Users Group International (SUGI) abstracts. This example assumes that SAS Enterprise Miner is running, and that a diagram workspace has been opened in a project. For information about creating a project and a diagram, see "Setting Up Your Project" on page 9.

Note: SAS Users Group International is now SAS Global Forum.

Perform the following steps:

1. Create a data source for SAMPSIO.ABSTRACT. Change the Role of the variable TITLE to `ID`.

   Note: The SAMPSIO.ABSTRACT data set contains information about 1,238 papers prepared for meetings of SUGI from 1998 through 2001 (SUGI 23 through 26). The variable TITLE is the title of the SUGI paper. The variable TEXT contains the abstract of the SUGI paper.

2. Add the SAMPSIO.ABSTRACT data source to the diagram workspace.

3. Select the **Text Mining** tab on the Toolbar, and drag a **Text Parsing** node into the diagram workspace.

4. Connect the Input Data node to the **Text Parsing** node.

5. Select the **Text Parsing** node, and then click the ellipsis for the **Stop List** property.

6. Click the **Replace Table** button, browse to select SAMPSIO.SUGISTOP as the stop list, and then click **OK**. Click **Yes** in the confirmation dialog box. Click **OK** to exit the dialog box for the **Stop List** property.

7. Set the **Find Entities** property to **Standard**.

8. Click the ellipsis button for the **Ignore Types of Entities** property to open the Ignore Types of Entities dialog box.

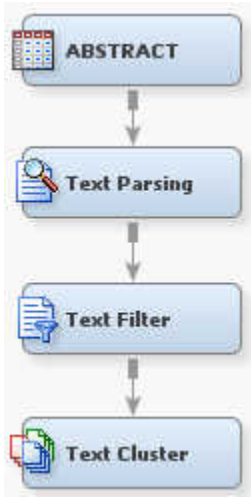
9. Select all entity types except for: `Location`, `Organization`, `Person`, and `Product`. Click **OK**.

10. Select the **Text Mining** tab, and drag a **Text Filter** node into the diagram workspace.

11. Connect the **Text Parsing** node to the **Text Filter** node.

12. Select the **Text Mining** tab, and drag a **Text Cluster** node into the diagram workspace.

13. Connect the **Text Filter** node to the **Text Cluster** node. Your process flow diagram should resemble the following:

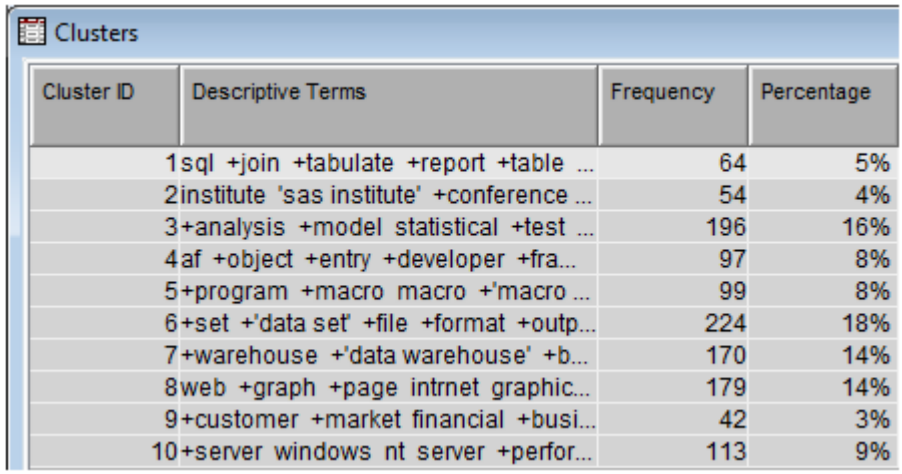


14. Right-click the **Text Cluster** node and select **Run**. Click **Yes** in the Confirmation dialog box.

15. Click **Results** in the Run Status dialog box when the node has finished running.

16. Select the Clusters table.

    The Clusters table contains an ID for each cluster, the descriptive terms that make up that cluster, and statistics for each cluster.

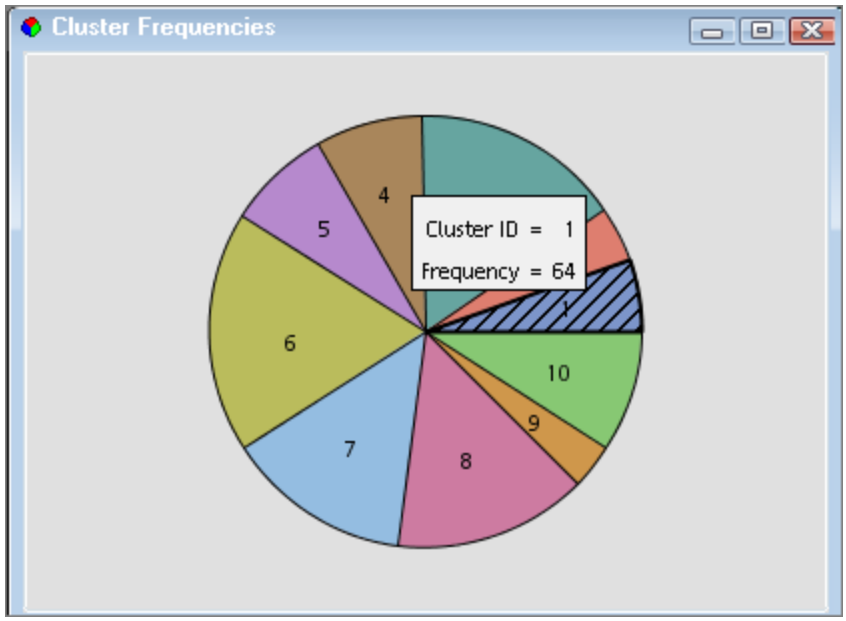


Clusters

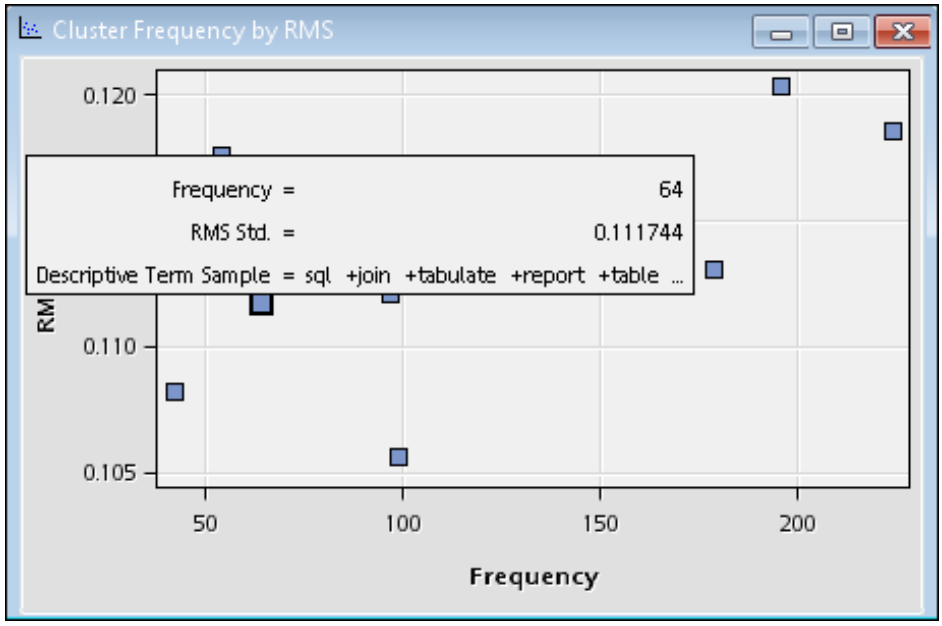
Cluster ID	Descriptive Terms	Frequency	Percentage
1	sql +join +tabulate +report +table ...	64	5%
2	institute 'sas institute' +conference ...	54	4%
3	+analysis +model statistical +test ...	196	16%
4	af +object +entry +developer +fra...	97	8%
5	+program +macro macro +'macro ...	99	8%
6	+set +'data set' +file +format +outp...	224	18%
7	+warehouse +'data warehouse' +b...	170	14%
8	web +graph +page intrnet graphic...	179	14%
9	+customer +market financial +busi...	42	3%
10	+server windows nt server +perfor...	113	9%

17. Select the first cluster in the Clusters table.

18. Select the Cluster Frequencies window to see a pie chart of the clusters by frequency. Position the mouse pointer over a section to see the frequency for that cluster in a tooltip.

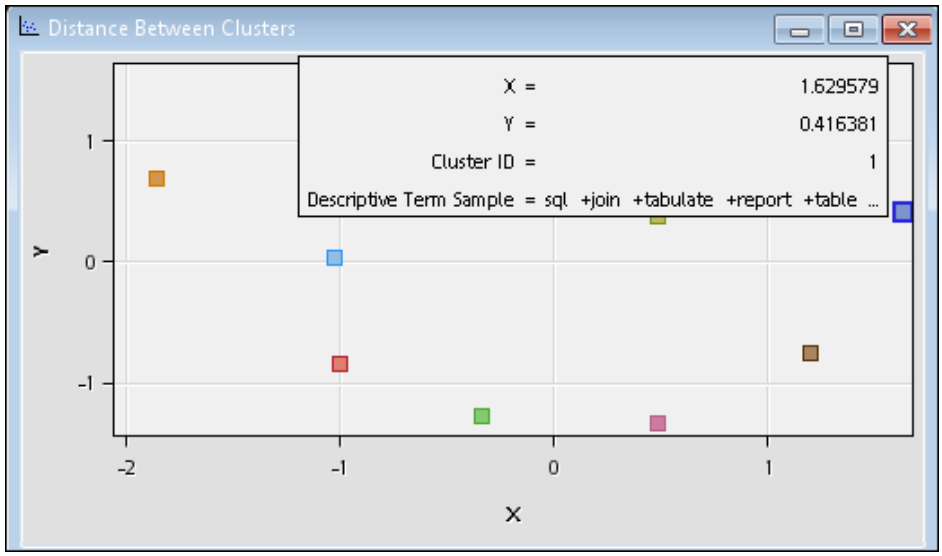


19. Select the Cluster Frequency by RMS window, and then position the mouse pointer over the highlighted cluster.



How does the first cluster compare to the other clusters in terms of distance?

20. Select the Distance Between Clusters window, and then position the mouse pointer over the highlighted cluster to see the position of the first cluster in an X and Y coordinate grid.



Position the mouse pointer over other clusters to compare distances.

21. Close the Results window.

    Now compare the clustering results obtained with the Expectation-Maximization clustering algorithm with using a Hierarchical clustering algorithm.

22. Select the **Text Cluster** node.

23. Select `Exact` for the **Exact or Maximum Number** property.

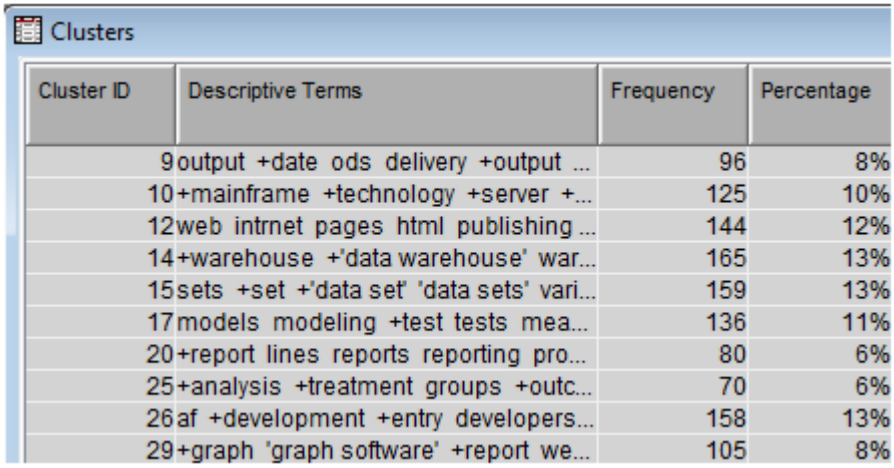
24. Specify *10* for the **Number of Clusters** property.

25. Select **`Hierarchical`** for the **Cluster Algorithm** property.

26. Right-click the **Text Cluster** node and select **Run**. Click **Yes** in the Confirmation dialog box.

27. Click **Results** in the Run Status dialog box when the node has finished running.

28. Select the Clusters table.

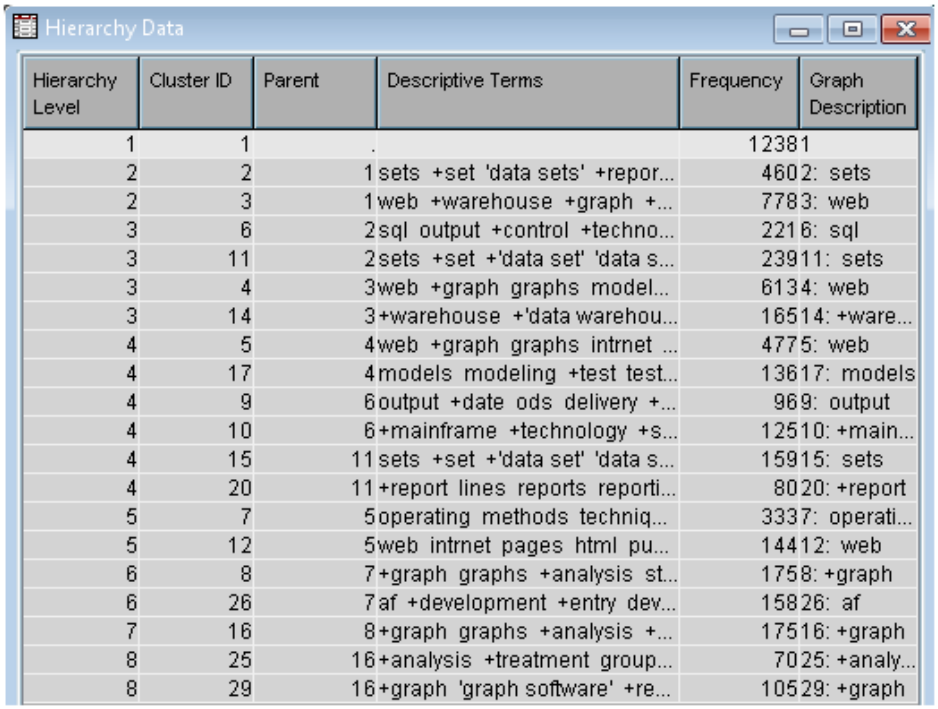


Clusters

Cluster ID	Descriptive Terms	Frequency	Percentage
9	output +date ods delivery +output ...	96	8%
10	+mainframe +technology +server +...	125	10%
12	web intrnet pages html publishing ...	144	12%
14	+warehouse +'data warehouse' war...	165	13%
15	sets +set +'data set' 'data sets' vari...	159	13%
17	models modeling +test tests mea...	136	11%
20	+report lines reports reporting pro...	80	6%
25	+analysis +treatment groups +outc...	70	6%
26	af +development +entry developers...	158	13%
29	+graph 'graph software' +report we...	105	8%

Notice that while there are 10 clusters in the table, the Cluster IDs do not range from 1 to 10.

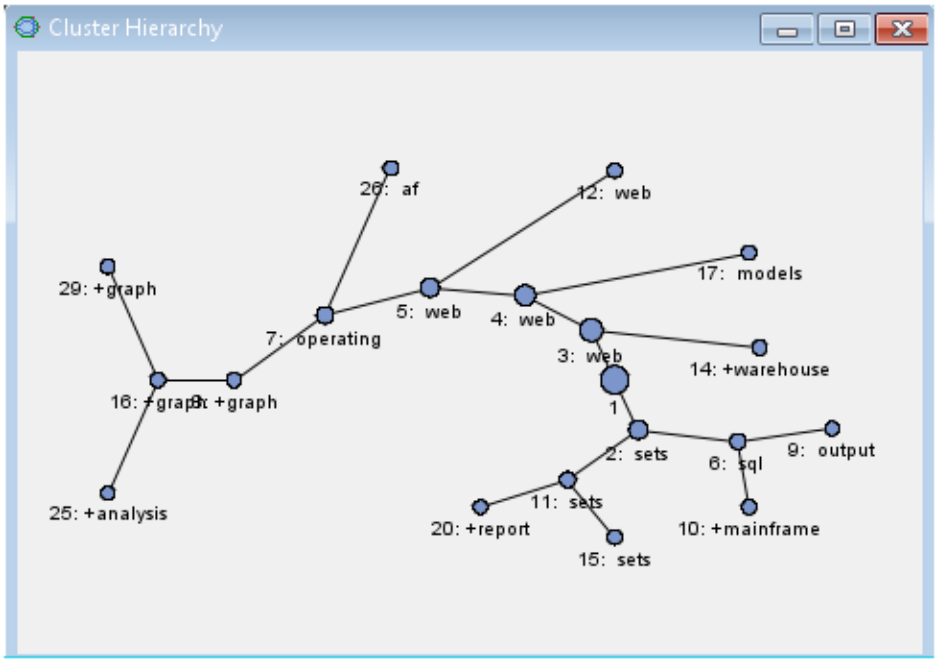
29. Select the Hierarchy Data table for more information about the clusters that appear in the Clusters table.



Hierarchy Data

Hierarchy Level	Cluster ID	Parent	Descriptive Terms	Frequency	Graph Description
1	1	.		12381	
2	2	1	sets +set 'data sets' +repor...	4602	2: sets
2	3	1	web +warehouse +graph +...	7783	3: web
3	6	2	sql output +control +techno...	2216	6: sql
3	11	2	sets +set +'data set' 'data s...	23911	1: sets
3	4	3	web +graph graphs model...	6134	4: web
3	14	3	+warehouse +'data warehou...	16514	4: +ware...
4	5	4	web +graph graphs intrnet ...	4775	5: web
4	17	4	models modeling +test test...	13617	7: models
4	9	6	output +date ods delivery +...	969	9: output
4	10	6	+mainframe +technology +s...	12510	0: +main...
4	15	11	sets +set +'data set' 'data s...	15915	5: sets
4	20	11	+report lines reports reporti...	8020	0: +report
5	7	5	operating methods techniq...	3337	7: operati...
5	12	5	web intrnet pages html pu...	14412	2: web
6	8	7	+graph graphs +analysis st...	1758	8: +graph
6	26	7	af +development +entry dev...	15826	6: af
7	16	8	+graph graphs +analysis +...	17516	6: +graph
8	25	16	+analysis +treatment group...	7025	5: +analy...
8	29	16	+graph 'graph software' +re...	10529	9: +graph

30. Select the Cluster Hierarchy graph for a hierarchical graphical representation of the clusters.



31. Close the Results window.

Chapter 13

The Text Rule Builder Node

About the Text Rule Builder Node

The **Text Rule Builder** node generates an ordered set of rules that together are useful in describing and predicting a target variable. Each rule in the set is associated with a specific target category, consisting of a conjunction that indicates the presence or absence of one or a small subset of terms (for example, “term1” AND “term2” AND (NOT “term3”)). A particular document matches this rule if and only if it contains at least one occurrence of term1 and of term2 but no occurrences of term3.

This set of derived rules creates a model that is both descriptive and predictive. When categorizing a new document, it will proceed through the ordered set and choose the target that is associated with the first rule that matches that document. The rules are provided in the syntax that can be used within SAS Content Categorization Studio, and can be deployed there.

The **Text Rule Builder** node is a standard SAS Enterprise Miner modeling tool, complete with the standard reporting features. You can view which predicted target values are most likely to be wrong based on the generated model. Optionally, you can change the target that is assigned to some of these observations and rerun the results. Thus it facilitates “active learning” in which a user can dynamically interact with an algorithm to iteratively build a predictive model.

For more information about the **Text Rule Builder** node, see the SAS Text Miner Help.

The rest of this chapter presents an example of how you can use the **Text Rule Builder** node.

Using the Text Rule Builder Node

This example assumes that SAS Enterprise Miner is running, and a diagram workspace has been opened in a project. For information about creating a project and a diagram, see "Setting Up Your Project" on page 9.

The **Text Rule Builder** node creates Boolean rules from small subsets of terms to predict a categorical target variable. The node must be preceded by **Text Parsing** and **Text Filter** nodes.

This example uses the SAMPSIO.NEWS data set to show you how to predict a categorical target variable with the **Text Rule Builder** node. The results will also show that the model is highly interpretable and useful for explanatory and summary purposes as well.

The SAMPSIO.NEWS data set consists of 600 brief news articles. Most of the news articles fall into one of these categories: computer graphics, hockey, and medical issues.

The SAMPSIO.NEWS data set contains 600 observations and the following variables:

- **TEXT** is a nominal variable that contains the text of the news article.
- **graphics** is a binary variable that indicates whether the document belongs to the computer graphics category (1-yes, 0-no).
- **hockey** is a binary variable that indicates whether the document belongs to the hockey category (1-yes, 0-no).
- **medical** is a binary variable that indicates whether the document is related to medical issues (1-yes, 0-no).
- **newsgroup** is a nominal variable that contains the group that a news article fits into.

To use the **Text Rule Builder** node to predict the categorical target variable, **newsgroup**, in the SAMPSIO.NEWS data set:

1. Use the Data Source Wizard to define a data source for the data set SAMPSIO.NEWS.

   a. Set the measurement levels of the variables **graphics**, **hockey**, and **medical** to **`Binary`**.

   b. Set the model role of the variable **newsgroup** to **`Target`** and leave the roles of **graphics**, **hockey**, and **medical** as **`Input`**.

   c. Set the variable **TEXT** to have a role of **`Text`**.

   d. Select **`No`** in the Data Source Wizard — Decision Configuration dialog box.

   e. Use the default target profile for the target **newsgroup**.

2. After you create the **NEWS** data source, drag it to the diagram workspace.

3. Select the **Text Mining** tab on the toolbar, and drag a **Text Parsing** node into the diagram workspace.

4. Connect the **NEWS** data source to the **Text Parsing** node.

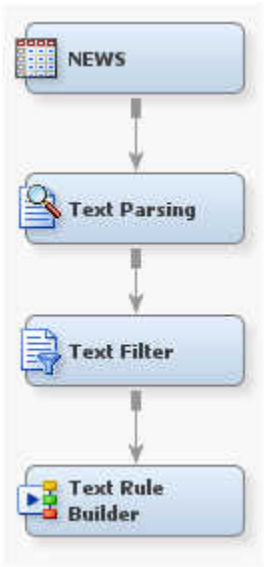
5. Select the **Text Mining** tab on the toolbar, and drag a **Text Filter** node into the diagram workspace.

6. Connect the **Text Parsing** node to the **Text Filter** node.

7. Select the **Text Mining** tab on the toolbar, and drag a **Text Rule Builder** node into the diagram workspace.

8. Connect the **Text Filter** node to the **Text Rule Builder** node.

   Your process flow diagram should resemble the following:



9. Select the **Text Rule Builder** node in the process flow diagram.

10. Click the value for the **Generalization Error** property, and select `Very Low`.

11. Click the value for the **Purity of Rules** property, and select `Very Low`.

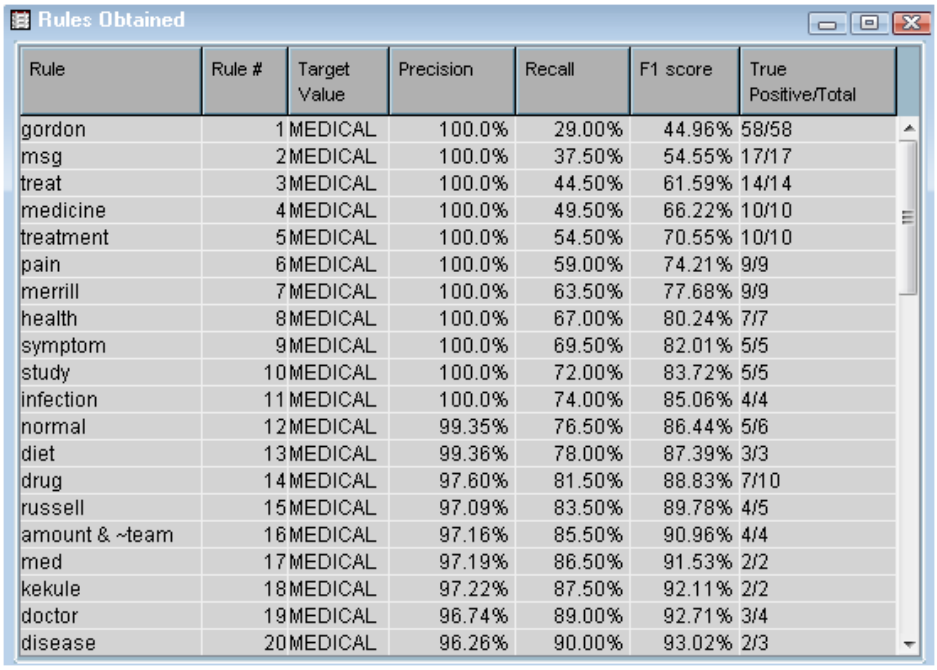
12. Click the value for the **Exhaustiveness** property, and select `Very Low`.

13. In the diagram workspace, right-click the **Text Rule Builder** node and select **Run**. Click **Yes** in the Confirmation dialog box that appears.

14. Click **Results** in the Run Status dialog box when the node finishes running.

15. Select the Rules Obtained table to see information about the rules that were obtained.

    The words in the Rule column have the corresponding precision at implying the target, **newsgroup**.



Rules Obtained

Rule	Rule #	Target Value	Precision	Recall	F1 score	True Positive/Total
gordon	1	MEDICAL	100.0%	29.00%	44.96%	58/58
msg	2	MEDICAL	100.0%	37.50%	54.55%	17/17
treat	3	MEDICAL	100.0%	44.50%	61.59%	14/14
medicine	4	MEDICAL	100.0%	49.50%	66.22%	10/10
treatment	5	MEDICAL	100.0%	54.50%	70.55%	10/10
pain	6	MEDICAL	100.0%	59.00%	74.21%	9/9
merrill	7	MEDICAL	100.0%	63.50%	77.68%	9/9
health	8	MEDICAL	100.0%	67.00%	80.24%	7/7
symptom	9	MEDICAL	100.0%	69.50%	82.01%	5/5
study	10	MEDICAL	100.0%	72.00%	83.72%	5/5
infection	11	MEDICAL	100.0%	74.00%	85.06%	4/4
normal	12	MEDICAL	99.35%	76.50%	86.44%	5/6
diet	13	MEDICAL	99.36%	78.00%	87.39%	3/3
drug	14	MEDICAL	97.60%	81.50%	88.83%	7/10
russell	15	MEDICAL	97.09%	83.50%	89.78%	4/5
amount & ~team	16	MEDICAL	97.16%	85.50%	90.96%	4/4
med	17	MEDICAL	97.19%	86.50%	91.53%	2/2
kekule	18	MEDICAL	97.22%	87.50%	92.11%	2/2
doctor	19	MEDICAL	96.74%	89.00%	92.71%	3/4
disease	20	MEDICAL	96.26%	90.00%	93.02%	2/3

In the seventh column above, the True Positive (the first number) is the number of documents that were correctly assigned to the rule. The Total (the second number) is the total positive.

In the above example, in the first row, 58 of the documents were assigned to the rule "gordon" (58 were correctly assigned). This means that if a document contains the word "gordon," and you assign all those documents to the MEDICAL newsgroup, 58 out of 58 will be assigned correctly. In the next row, 17 documents are correctly assigned to the rule "msg." This means that if a document contains the term "msg," and you assign all those documents to the MEDICAL newsgroup, 17 out of 17 will be assigned correctly.

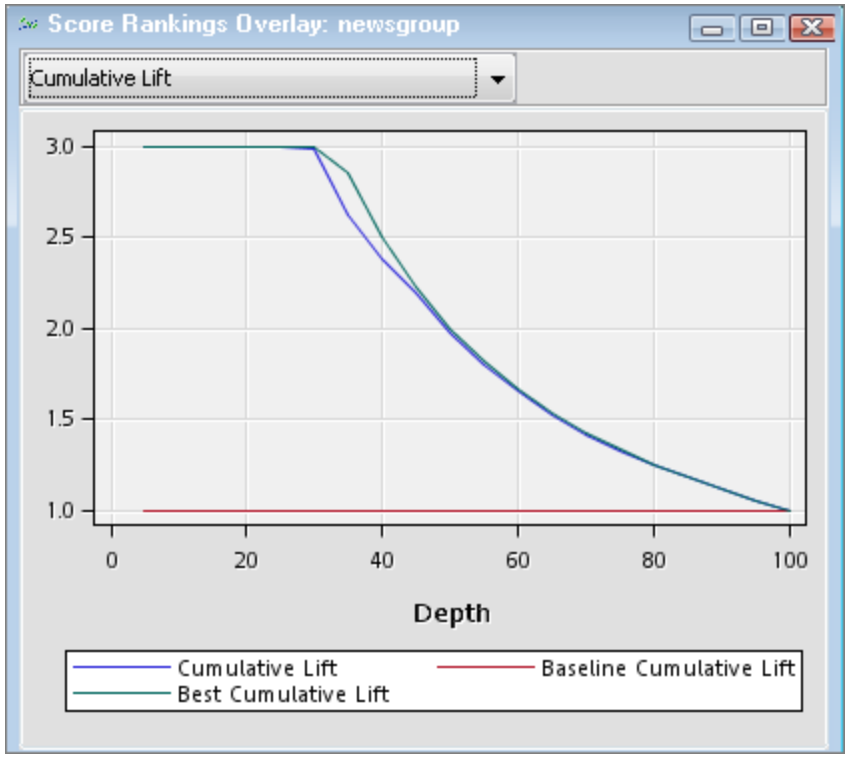
Most of the rules are single term rules because the NEWS data set is limited in size. However, there is one multiple term rule above. In the 16th row, the rule "amount & ~team" means that if a document contains the word "amount" and does not contain the word "team", then 4 of the remaining documents will be correctly assigned to the MEDICAL newsgroup.

Note: ~ means logical not.

16. Select the Score Rankings Overlay graph to view the following types of information about the target variable:

    - Cumulative Lift
    - Lift
    - Gain
    - % Response
    - Cumulative % Response
    - % Captured Response
    - Cumulative % Captured Response

    Note: To change the statistic, select one of the above options from the drop-down menu.

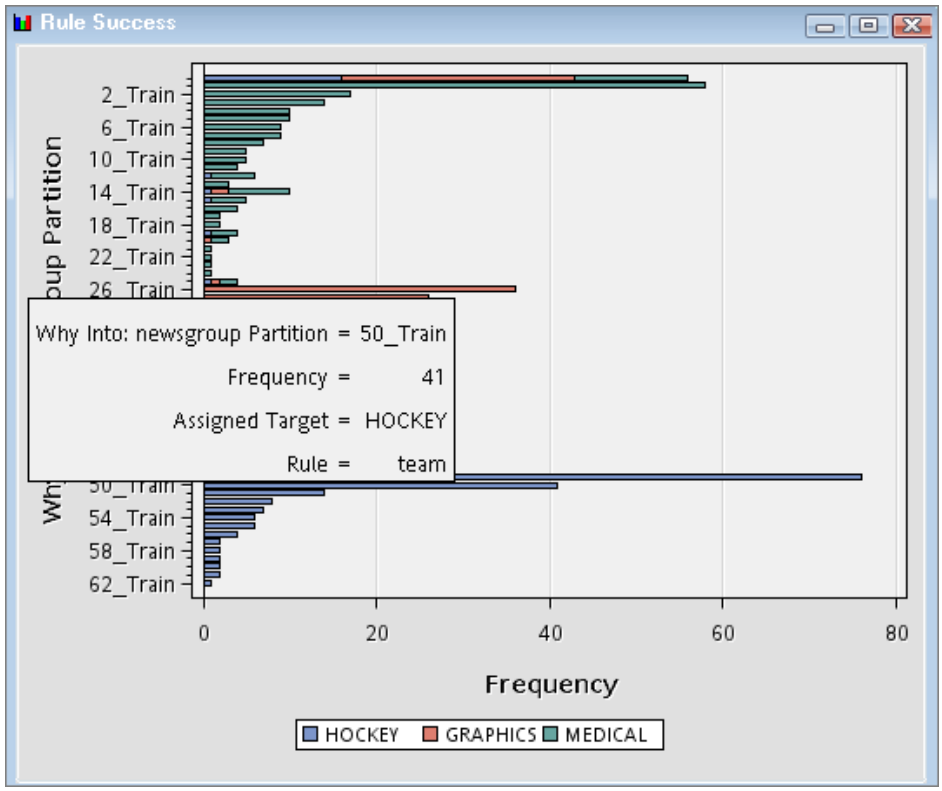


17. Select the Fit Statistics window for statistical information about the target variable, **newsgroup**.

Fit Statistics

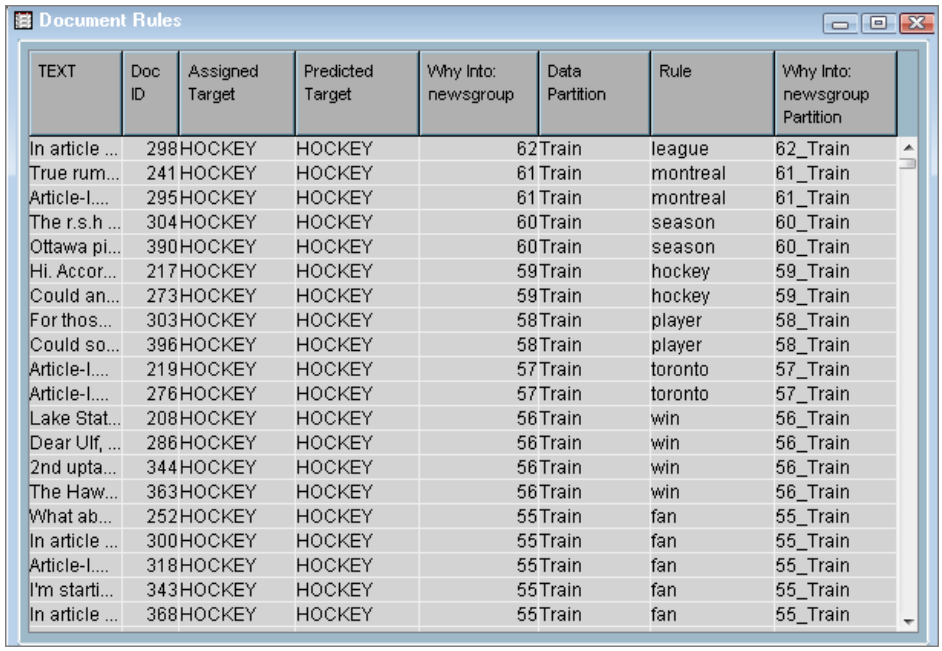
Target	Target Label	Fit Statistics	Statistics Label	Train
newsgroup		_ASE_	Average Squared Error	0.037681
newsgroup		_DIV_	Divisor for ASE	1800
newsgroup		_MAX_	Maximum Absolute Error	1
newsgroup		_NOBS_	Sum of Frequencies	600
newsgroup		_RASE_	Root Average Squared Error	0.194116
newsgroup		_SSE_	Sum of Squared Errors	67.82585
newsgroup		_DISF_	Frequency of Classified Cases	600
newsgroup		_MISC_	Misclassification Rate	0.075
newsgroup		_WRONG_	Number of Wrong Classifications	45

18. Select the Rule Success graph, and position the cursor over a bar for more information.



19. Select **View** ⇨ **Rules** ⇨ **Document Rules** from the menu.

    The Document Rules table appears with more information about the rules in the Rule Success graph.



TEXT	Doc ID	Assigned Target	Predicted Target	Why Into: newsgroup	Data Partition	Rule	Why Into: newsgroup Partition
In article ...	298	HOCKEY	HOCKEY	62	Train	league	62_Train
True rum...	241	HOCKEY	HOCKEY	61	Train	montreal	61_Train
Article-I....	295	HOCKEY	HOCKEY	61	Train	montreal	61_Train
The r.s.h ...	304	HOCKEY	HOCKEY	60	Train	season	60_Train
Ottawa pi...	390	HOCKEY	HOCKEY	60	Train	season	60_Train
Hi. Accor...	217	HOCKEY	HOCKEY	59	Train	hockey	59_Train
Could an...	273	HOCKEY	HOCKEY	59	Train	hockey	59_Train
For thos...	303	HOCKEY	HOCKEY	58	Train	player	58_Train
Could so...	396	HOCKEY	HOCKEY	58	Train	player	58_Train
Article-I....	219	HOCKEY	HOCKEY	57	Train	toronto	57_Train
Article-I....	276	HOCKEY	HOCKEY	57	Train	toronto	57_Train
Lake Stat...	208	HOCKEY	HOCKEY	56	Train	win	56_Train
Dear Ulf, ...	286	HOCKEY	HOCKEY	56	Train	win	56_Train
2nd upta...	344	HOCKEY	HOCKEY	56	Train	win	56_Train
The Haw...	363	HOCKEY	HOCKEY	56	Train	win	56_Train
What ab...	252	HOCKEY	HOCKEY	55	Train	fan	55_Train
In article ...	300	HOCKEY	HOCKEY	55	Train	fan	55_Train
Article-I....	318	HOCKEY	HOCKEY	55	Train	fan	55_Train
I'm starti...	343	HOCKEY	HOCKEY	55	Train	fan	55_Train
In article ...	368	HOCKEY	HOCKEY	55	Train	fan	55_Train

20. Close the Results window.

21. Click the value for the **Generalization Error** property, and select **`Medium`**.

22. Click the value for the **Purity of Rules** property, and select **`Medium`**.

23. Click the value for the **Exhaustiveness** property, and select **`Medium`**.

24. Select the **NEWS** data source.

25. Click the ellipsis button for the **Variables** property.

26. Change the role of the HOCKEY variable to **`Target`**, and change the role of the NEWSGROUP variable to **`Input`**.

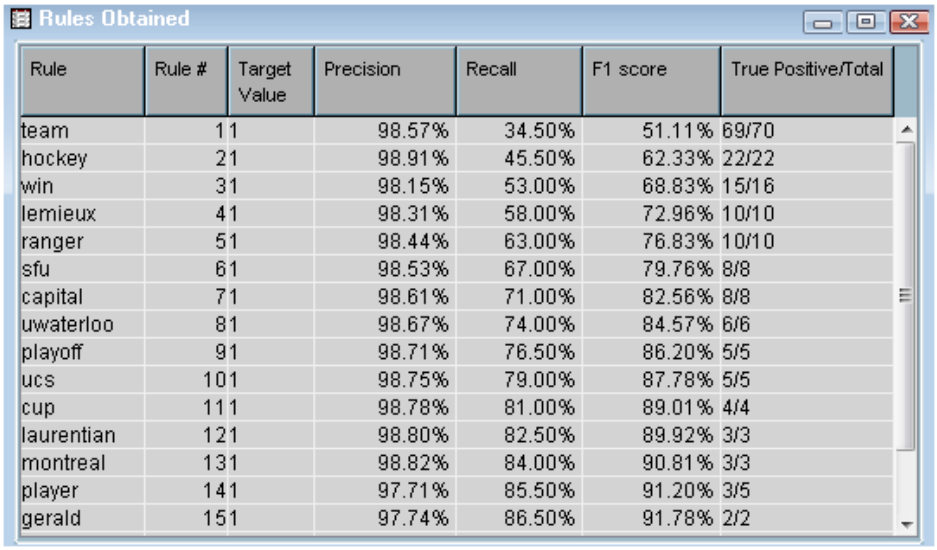
27. Click **OK**.

28. In the diagram workspace, right-click the **Text Rule Builder** node and select **Run**. Click **Yes** in the Confirmation dialog box that appears.

29. Click **Results** in the Run Status dialog box when the node finishes running.

30. Select the Rules Obtained table to see information about the rules that predicted the target — the HOCKEY newsgroup.

    The words in the Rule column have the corresponding precision at implying the hockey target.



Rules Obtained

Rule	Rule #	Target Value	Precision	Recall	F1 score	True Positive/Total
team	11		98.57%	34.50%	51.11%	69/70
hockey	21		98.91%	45.50%	62.33%	22/22
win	31		98.15%	53.00%	68.83%	15/16
lemieux	41		98.31%	58.00%	72.96%	10/10
ranger	51		98.44%	63.00%	76.83%	10/10
sfu	61		98.53%	67.00%	79.76%	8/8
capital	71		98.61%	71.00%	82.56%	8/8
uwaterloo	81		98.67%	74.00%	84.57%	6/6
playoff	91		98.71%	76.50%	86.20%	5/5
ucs	101		98.75%	79.00%	87.78%	5/5
cup	111		98.78%	81.00%	89.01%	4/4
laurentian	121		98.80%	82.50%	89.92%	3/3
montreal	131		98.82%	84.00%	90.81%	3/3
player	141		97.71%	85.50%	91.20%	3/5
gerald	151		97.74%	86.50%	91.78%	2/2

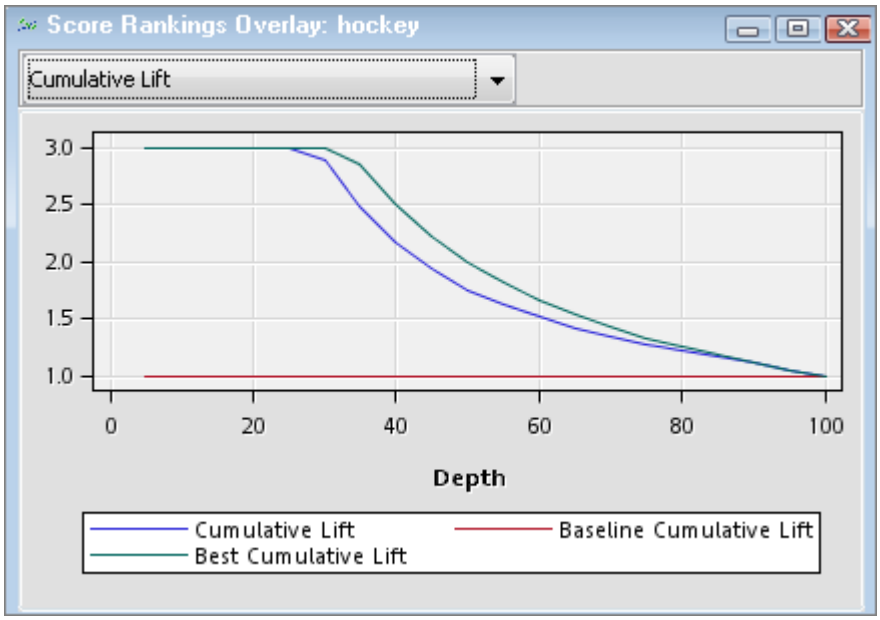
    In the above example, the target value is **`1`**, instead of "HOCKEY," because we set the **hockey** variable to be the target instead of the **newsgroup** variable. 70 of the documents were assigned to the rule "team" (69 were correctly assigned). This means that if a document contains the word "team," and you assign all those documents to the HOCKEY newsgroup, 69 out of 70 will be assigned correctly. In the next row, 22 documents are correctly assigned to the rule "hockey." This means that if a document contains the word "hockey," and you assign all those documents to the HOCKEY newsgroup, 22 out of 22 will be assigned correctly.

31. Select the Score Rankings Overlay graph to view the following types of information about the target variable:

    - Cumulative Lift
    - Lift
    - Gain
    - % Response

- Cumulative % Response
- % Captured Response
- Cumulative % Captured Response

Note: To change the statistic, select one of the above options from the drop-down menu.

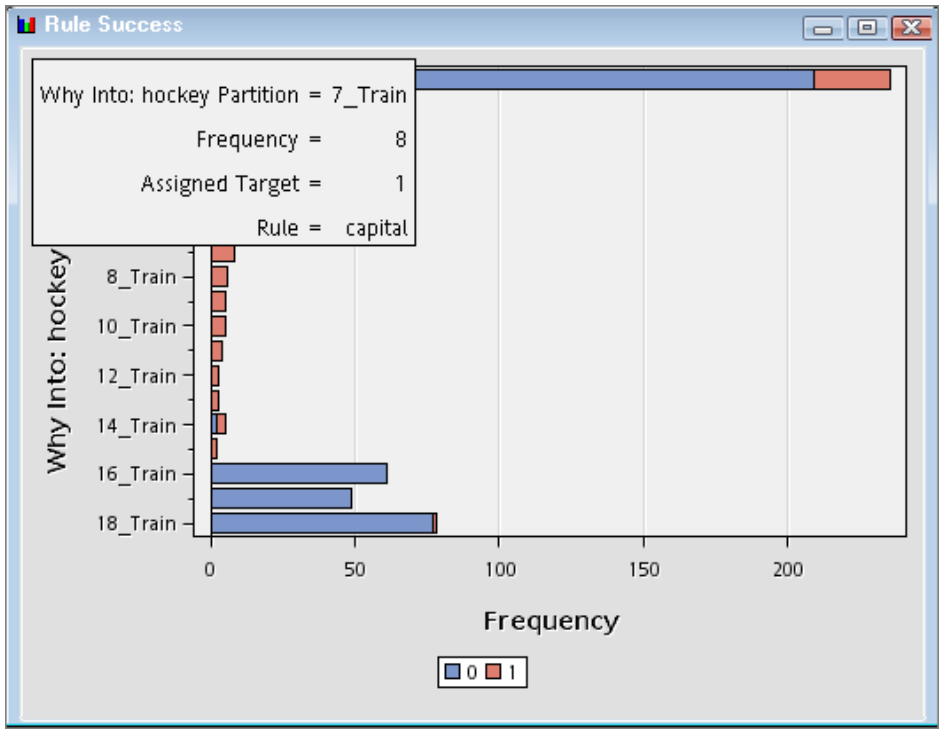


32. Select the Fit Statistics table for statistical information about the hockey target variable.

Fit Statistics

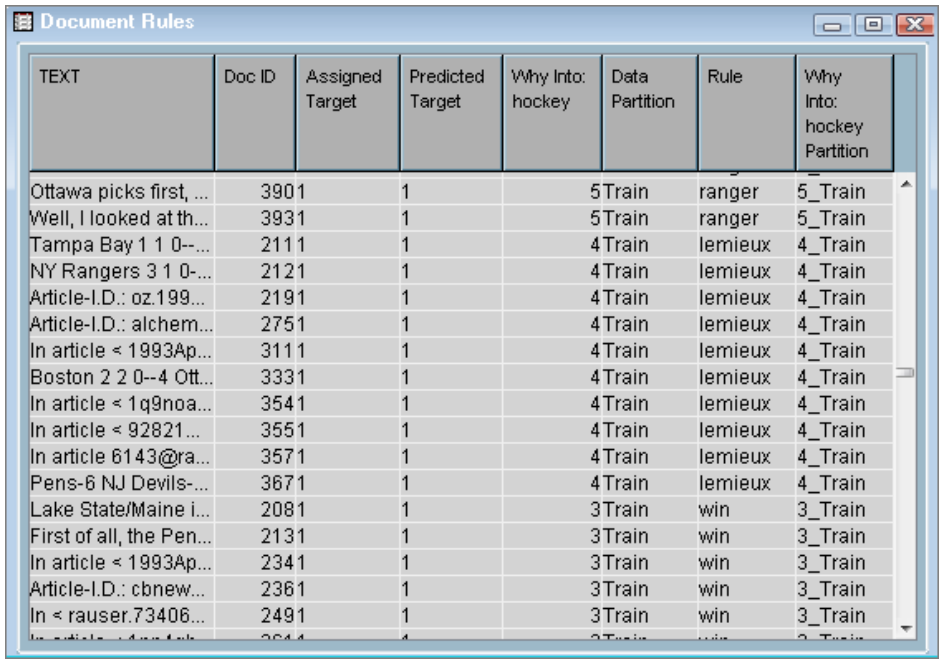
Target	Target Label	Fit Statistics	Statistics Label	Train
hockey		_ASE_	Average Squared Error	0.009274
hockey		_DIV_	Divisor for ASE	1200
hockey		_MAX_	Maximum Absolute Error	0.957888
hockey		_NOBS_	Sum of Frequencies	600
hockey		_RASE_	Root Average Squared Error	0.096302
hockey		_SSE_	Sum of Squared Errors	11.12879
hockey		_DISF_	Frequency of Classified Cases	600
hockey		_MISC_	Misclassification Rate	0.051667
hockey		_WRONG_	Number of Wrong Classifications	31

33. Select the Rule Success graph, and position the cursor over a bar for more information.



34. Select **View** ⇨ **Rules** ⇨ **Document Rules** from the menu.

    The Document Rules table appears with more information about the rules in the Rule Success graph.



35. Close the Results window.

36. Click the ellipsis button for the **Content Categorization Code** property.

    The Content Categorization Code window appears. The code provided in this window is the code that is output for SAS Content Categorization and is ready for compilation.

37. Click **Cancel**.

38. Click the ellipsis button for the **Change Target Values** property.

    The Change Target Values window appears.

    You can use the Change Target Values window to improve the model.

39. Select one or more cells in the **Assigned Target** column, and select a new target value.

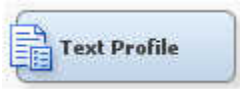
40. Click **OK**.

41. Rerun the **Text Rule Builder** node, and then check whether your model has been improved.

Chapter 14

The Text Profile Node

About the Text Profile Node



The **Text Profile** node enables you to profile a target variable using terms found in the documents. For each level of a target variable, the node outputs a list of terms from the collection that characterize or describe that level.

The approach uses a hierarchical Bayesian model using proc TMBelief to predict which terms are the most likely terms to describe the level. In order to avoid merely selecting the most common terms, prior probabilities are used to down-weight terms that are common in more than one level of the target variable. For binary target variables, a two-way comparison is used to enhance the selection of terms. For nominal variables, an n-way comparison is used. Finally, for ordinal and time variables (which are converted to ordinal internally), a sequential, two-way comparison is done. This means that the reported terms for level n are compared to those at level n-1. The exception for this is the first level, which is compared to level 2 since there is no preceding level to compare it to.

In all cases of variable types, a corpus level profile output is also provided. This can be interpreted as the best descriptive terms for the entire collection itself.

For more information about the **Text Profile** node, see the SAS Text Miner Help.

The rest of this chapter presents an example of how you can use the **Text Profile** node.

Using the Text Profile Node

This example uses the SAMPSIO.NEWS data set to profile terms using the **Text Profile** node. This example assumes that SAS Enterprise Miner is running, and a diagram workspace has been opened in a project. For information about creating a project and a diagram, see “Setting Up Your Project” on page 9.

The **Text Profile** node must be preceded by one **Text Parsing** node and at least one **Text Filter** node in the diagram.

The SAMPSIO.NEWS data set consists of 600 brief news articles. Most of the news articles fall into one of these categories: computer graphics, hockey, and medical issues.

The SAMPSIO.NEWS data set contains 600 observations and the following variables:

- **TEXT** is a nominal variable that contains the text of the news article.
- **graphics** is a binary variable that indicates whether the document belongs to the computer graphics category (1-yes, 0-no).
- **hockey** is a binary variable that indicates whether the document belongs to the hockey category (1-yes, 0-no).
- **medical** is a binary variable that indicates whether the document is related to medical issues (1-yes, 0-no).
- **newsgroup** is a nominal variable that contains the group that a news article fits into.

1. Use the Data Source Wizard to define a data source for the data set SAMPSIO.NEWS.

   a. Set the measurement levels of the variables **graphics**, **hockey**, and **medical** to **`Binary`**.

   b. Set the model role of the variable **hockey** to **`Target`** and leave the roles of **newsgroup**, **graphics**, and **medical** as **`Input`**.

   c. Set the variable **TEXT** to have a role of **`Text`**.

   d. Select **`No`** in the Data Source Wizard — Decision Configuration dialog box.

   e. Use the default target profile for the target **hockey**.

2. After you create the **NEWS** data source, drag it to the diagram workspace.

3. Select the **Text Mining** tab on the toolbar, and drag a **Text Parsing** node to the diagram workspace.

4. Connect the **NEWS** data source to the **Text Parsing** node.

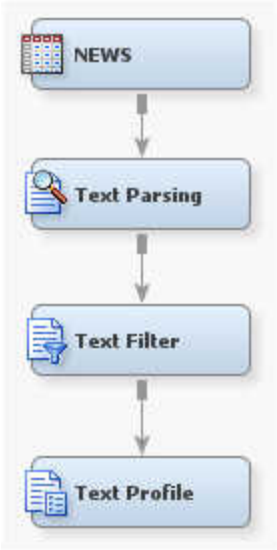
5. Select the **Text Mining** tab on the toolbar, and drag a **Text Filter** node to the diagram workspace.

6. Connect the **Text Parsing** node to the **Text Filter** node.

7. Select the **Text Mining** tab on the toolbar, and drag a **Text Profile** node to the diagram workspace.

8. Connect the **Text Filter** node to the **Text Profile** node.

   Your process flow diagram should resemble the following:

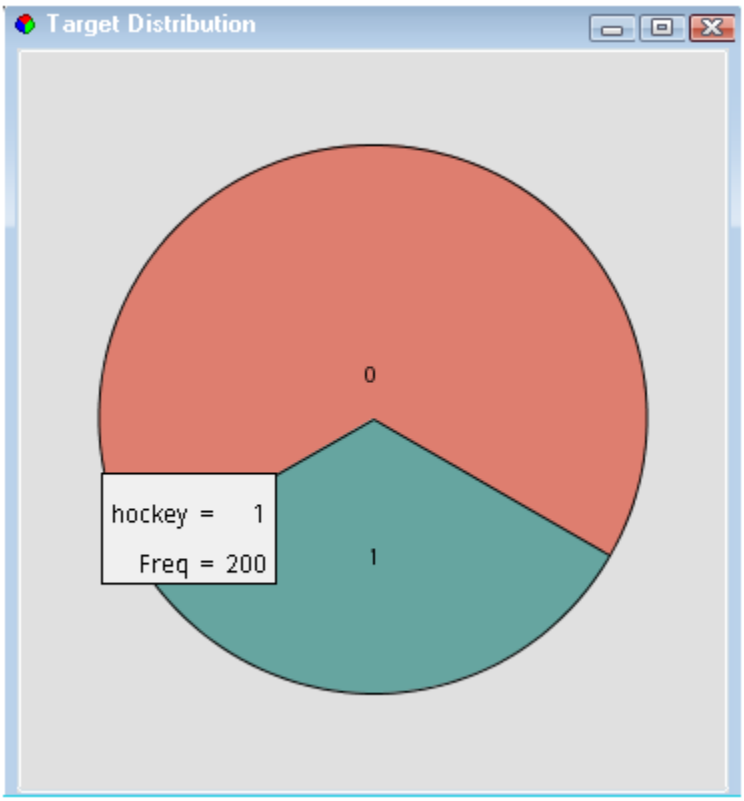


9. Select the **Text Profile** node in the process flow diagram.

10. In the diagram workspace, right-click the **Text Profile** node and select **Run**. Click **Yes** in the Confirmation dialog box that appears.

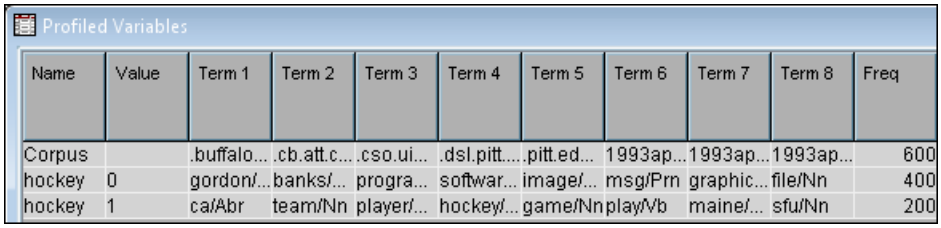
11. Click **Results** in the Run Status dialog box when the node finishes running.

12. Select the **Target Distribution** pie chart.



The **Target Distribution** pie chart shows the frequencies of the target values. It is grouped by hierarchy level and is linked to the Profiled Variables table.

13. Select the **Profiled Variables** table.



Profiled Variables

Name	Value	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Freq
Corpus		.buffalo...	.cb.att.c...	.cso.ui...	.dsl.pitt....	.pitt.ed...	1993ap...	1993ap...	1993ap...	600
hockey	0	gordon/...	banks/...	progra...	softwar...	image/...	msg/Prn	graphic...	file/Nn	400
hockey	1	ca/Abr	team/Nn	player/...	hockey/...	game/Nn	play/Vb	maine/...	sfu/Nn	200

The **Profiled Variables** table displays each combination of values of the target variables and their associated highest-belief terms. Each observation has at most the specified maximum number of terms associated it, but can have less. All graphical results in the **Text Profile** node Results window are linked to this table. Therefore, you can select an observation in the Profiled Variables table, and the associated data points are highlighted in the graphics. Or, you can select data points in the graphics, and the associated observations are highlighted in the Profiled Variables table.

14. Close the Results window.

15. Select the **News** data source.

16. Click the ellipsis button for the **`Variables`** property.

    The Variables dialog box appears.

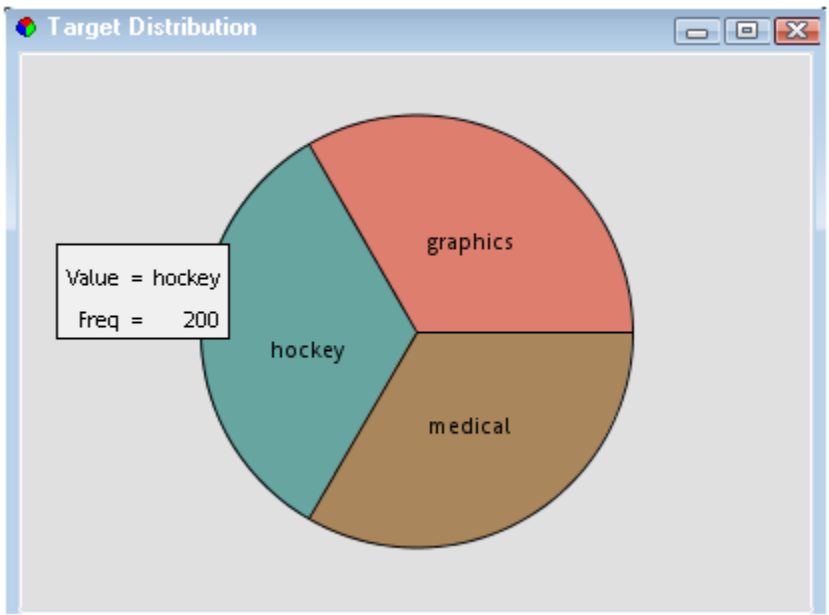
17. Set the role of the **newsgroup** variable to **`Target`**, and set the role of the **hockey** variable to **Input**.

18. Click **OK**.

19. In the diagram workspace, right-click the **Text Profile** node and click **Run**. Click **Yes** in the Confirmation dialog box that appears.

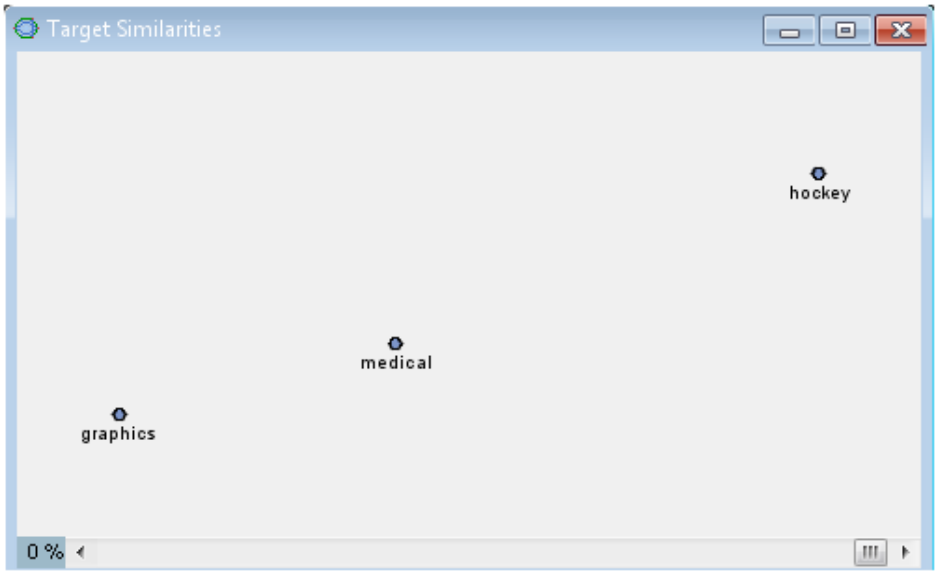
20. Click **Results** in the Run Status dialog box when the node finishes running.

21. Select the **Target Distribution** pie chart.



Since we changed the value of the Target from the binary hockey variable to the nominal newsgroup variable, we can now see the distribution of the three possible newsgroup values (hockey, medical, and graphics).

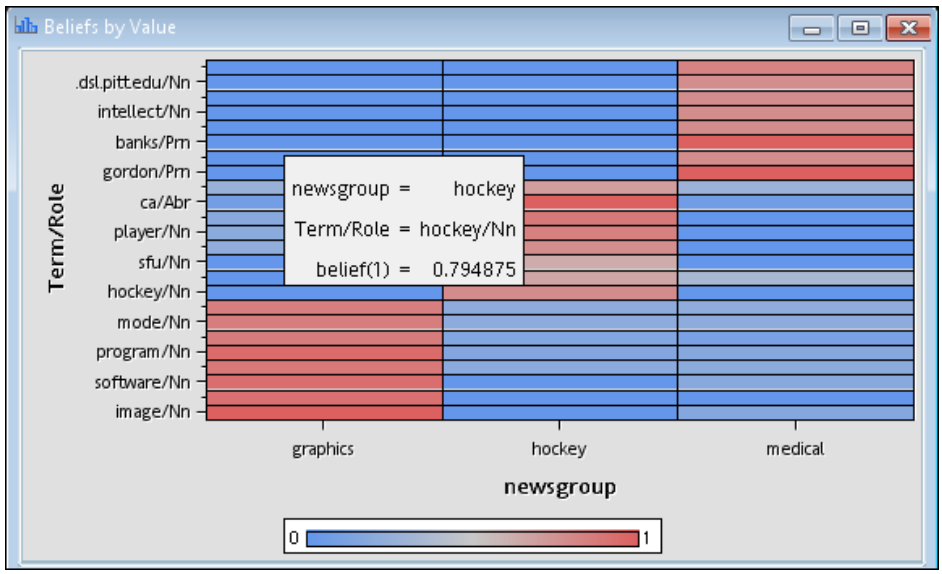
22. Select the **Target Similarities** constellation plot.



The **Target Similarities** constellation plot shows similarities among different target values. The similarities are measured using PROC DISTANCE on the term beliefs. Links are shown only between those target values on the same level of the hierarchy. The constellation plot is linked to the Profiled Variables table.

Note: The **Target Similarities** constellation plot is available for nominal and ordinal targets.

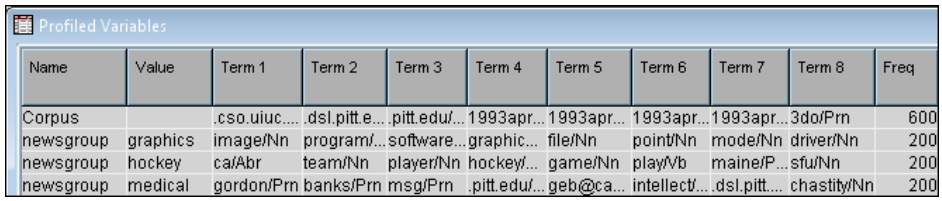
23. Select the **Beliefs by Value** graph.



The **Beliefs by Value** graph shows belief values for term and role pairs for various target values. If you position the mouse pointer over a cell, then a tooltip indicates the target value, term and role pair, and belief value.

Note: The **Beliefs by Value** graph is displayed for nominal and ordinal targets.

24. Select the **Profiled Variables** table.


Profiled Variables

Name	Value	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Freq
Corpus		.cso.uiuc....	.dsl.pitt.e...	pitt.edu/...	1993apr...	1993apr...	1993apr...	1993apr...	3do/Prn	600
newsgroup	graphics	image/Nn	program/...	software...	graphic...	file/Nn	point/Nn	mode/Nn	driver/Nn	200
newsgroup	hockey	ca/Abr	team/Nn	player/Nn	hockey/...	game/Nn	play/Vb	maine/P...	sfu/Nn	200
newsgroup	medical	gordon/Prn	banks/Prn	msg/Prn	.pitt.edu/...	geb@ca...	intellect/...	dsl.pitt....	chastity/Nn	200

25. Close the Results window.

26. Select the **Text Profile** node.

27. Click the value for the **Maximum Number of Terms** property, and enter `16`.

    Click **Enter** on your keyboard.

28. In the diagram workspace, right-click the **Text Profile** node and click **Run**. Click **Yes** in the Confirmation dialog box that appears.

29. Click **Results** in the Run Status dialog box when the node finishes running.

30. Select the **Profiled Variables** table. Notice that there are now 16 term and role pairs for each target value.

31. Close the Results window.

32. (Optional) Run the **Text Profile** node with a target variable that has the format type DATE or DATETIME. This will produce a Term Time Series line plot in the results.

    a. Create the data source **VRTEXT** from the **Sampsio** library. Set the role value of the **VAX_DATE** variable to `Target`.

       This data source contains adverse event reactions to vaccinations. For example, some side effects reported might include soreness, redness, fever, and so on.

    b. Add the **VRTEXT** node to the diagram workspace.

    c. Select the **Text Mining** tab on the toolbar, and drag a **Text Parsing** node to the diagram workspace.

    d. Connect the **VRTEXT** data source to the **Text Parsing** node.

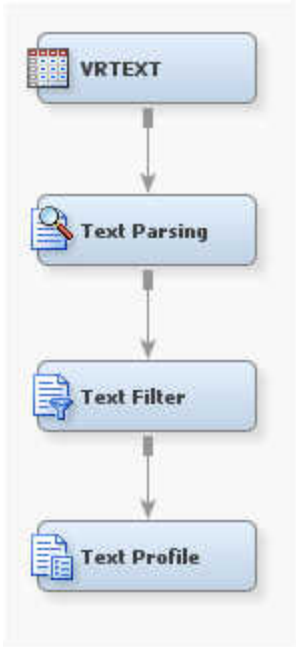
    e. Select the **Text Mining** tab on the toolbar, and drag a **Text Filter** node to the diagram workspace.

    f. Connect the **Text Parsing** node to the **Text Filter** node.

    g. Select the **Text Mining** tab on the toolbar, and drag a **Text Profile** node to the diagram workspace.

    h. Connect the **Text Filter** node to the **Text Profile** node.

       Your process flow diagram should resemble the following:

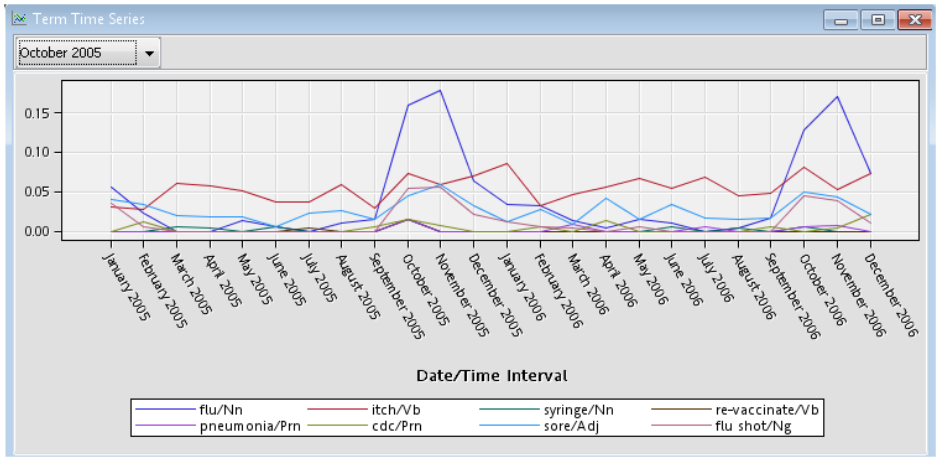


i. Select the **Text Profile** node, and select `Monthly` as the value for the **Date Binning Interval** property.

j. In the diagram workspace, right-click the **Text Profile** node and click **Run**. Click **Yes** in the Confirmation dialog box that appears.

k. Click **Results** in the Run Status dialog box when the node finishes running.

l. Select the Term Time Series line plot, and then select `Oct 2005` from the menu.



In this example, you can see that the flu starts climbing and stays prominent for several months.

33. (Optional) Run the **Text Profile** node with a target variable that is ordinal. This will produce a Term Ordinal Series line plot in the results.

Chapter 15

Tips for Text Mining

Processing a Large Collection of Documents

Using SAS Text Miner nodes to process a 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.
- Set some of the parse properties to **No** or **None**, such as **Noun Groups** or **Find Entities**.
- Reduce the number of SVD dimensions or roll-up terms. If you are running into memory problems with the SVD approach, you can roll up a certain number of terms, and then the remaining terms are automatically dropped.
- Limit parsing to high information words by turning off all parts of speech other than nouns, proper nouns, noun groups, and verbs.
- Structure sentences properly for best results, including correct grammar, punctuation, and capitalization. Entity extraction does not always generate reasonable results.

Dealing with Long Documents

SAS Text Miner uses the "bag-of-words" approach to represent documents. That means that documents are represented with a vector that contains the frequency with which each term occurs in each document. In addition, word order is ignored. This approach is very effective for short, paragraph-sized documents, but it can cause a harmful loss of information with longer documents. You might want to consider preprocessing your long documents in order to isolate the content that is really of use in your model. For example, if you are analyzing journal papers, you might find that analyzing only the abstract gives the best results. Consider using the SAS DATA step or an alternative programming language such as Perl to extract the relevant content from long documents.

Processing Documents from an Unsupported Language or Encoding

If you have a collection of documents from an unsupported language or encoding, you might still be able to successfully process the text and get useful results. Follow these steps:

1. Set the language to **English**.
2. Turn off these parse properties:
   - **Different Parts of Speech**
   - **Noun Groups**
   - **Find Entities**
   - **Stem Terms**
3. Run the **Text Parsing** node.

Chapter 16

Next Steps: A Quick Look at Additional Features

The %TEXTSYN Macro

The %TEXTSYN macro is provided with SAS Text Miner. You can use this macro after a **Text Parsing** node has been successfully run to find and correct misspellings that appear in the input data source. It is not supported for use with the Chinese language.

The macro creates a synonym data set, which you can use in SAS Text Miner, that contains misspelled terms and candidate parents (correctly spelled terms). The data set includes the variables “term,” “parent,” and “category.” Using optional arguments, you can also specify that the synonym data set include example usages (from up to two documents) of the misspelled terms.

See the SAS Text Miner Help for more information about the %TEXTSYN macro.

The %TMFILTER Macro

The %TMFILTER macro is a SAS macro that enables you to convert files into SAS data sets. The %TMFILTER macro is provided with SAS Text Miner. It is supported in all operating systems for filtering and on Windows for crawling. The %TMFILTER macro relies on the SAS Document Conversion Server that is installed and running on a Windows machine. See SAS Document Conversion server for more information. You can use the macro to perform the following tasks:

- filter a collection of documents that is saved in any supported file format and output a SAS data set that can be used to create a SAS Text Miner data source.
- Web crawl and output a SAS data set that can be used to create a SAS Text Miner data source. Web crawling retrieves the text of a starting Web page, extracts the URL links within that page, and then repeats the process within the linked pages recursively. You can restrict a crawl to the domain of the starting URL, or you can let a crawl process any linked pages that are not in the domain of the starting URL. The crawl continues until a specified number of levels of drill-down is reached or until all the Web pages that satisfy the domain constraint are found. Web crawling is supported only on Windows operating systems.

- identify the languages of all documents in a collection.

See the SAS Text Miner Help for more information about the %TMFILTER macro.

Glossary

catalog directory
a part of a SAS catalog that stores and maintains information about the name, type, description, and update status of each member of the catalog.

clustering
the process of dividing a data set into mutually exclusive groups so that the observations for each group are as close as possible to one another and different groups are as far as possible from one another. In SAS Text Miner, clustering involves discovering groups of documents that are more similar to each other than they are to the rest of the documents in the collection. When the clusters are determined, examining the words that occur in the cluster reveals the focus of the cluster. Forming clusters within the document collection can help you understand and summarize the collection without reading every document. The clusters can reveal the central themes and key concepts that are emphasized by the collection.

concept linking
finding and displaying the terms that are highly associated with the selected term in the Terms table.

data source
a data object that represents a SAS data set in the Java-based Enterprise Miner GUI. A data source contains all the metadata for a SAS data set that Enterprise Miner needs in order to use the data set in a data mining process flow diagram. The SAS data set metadata that is required to create an SAS Enterprise data source includes the name and location of the data set; the SAS code that is used to define its library path; and the variable roles, measurement levels, and associated attributes that are used in the data mining process.

diagram
See process flow diagram.

entity
any of several types of information that SAS Text Miner is able to distinguish from general text. For example, SAS Text Miner can identify names (of people, places, companies, or products, for example), addresses (including street addresses, post office addresses, e-mail addresses, and URLs), dates, measurements, currency amounts, and many other types of entities.

libref
a name that is temporarily associated with a SAS library. The complete name of a SAS file consists of two words, separated by a period. The libref, which is the first word, indicates the library. The second word is the name of the specific SAS file. For example, in VLIB.NEWBDAY, the libref VLIB tells SAS which library contains the file NEWBDAY. You assign a libref with a LIBNAME statement or with an operating system command.

model
a formula or algorithm that computes outputs from inputs. A data mining model includes information about the conditional distribution of the target variables, given the input variables.

node
(1) in the SAS Enterprise Miner user interface, a graphical object that represents a data mining task in a process flow diagram. The statistical tools that perform the data mining tasks are called nodes when they are placed on a data mining process flow diagram. Each node performs a mathematical or graphical operation as a component of an analytical and predictive data model. (2) in a neural network, a linear or nonlinear computing element that accepts one or more inputs, computes a function of the inputs, and optionally directs the result to one or more other neurons. Nodes are also known as neurons or units. (3) a leaf in a tree diagram. The terms leaf, node, and segment are closely related and sometimes refer to the same part of a tree.

parsing
to analyze text for the purpose of separating it into its constituent words, phrases, multiword terms, punctuation marks, or other types of information.

partitioning
to divide available data into training, validation, and test data sets.

process flow diagram
a graphical representation of the various data mining tasks that are performed by individual Enterprise Miner nodes during a data mining analysis. A process flow diagram consists of two or more individual nodes that are connected in the order in which the data miner wants the corresponding statistical operations to be performed. Short form: PFD.

roll-up terms
the highest-weighted terms in the document collection.

SAS data set
a file whose contents are in one of the native SAS file formats. There are two types of SAS data sets: SAS data files and SAS data views. SAS data files contain data values in addition to descriptor information that is associated with the data. SAS data views contain only the descriptor information plus other information that is required for retrieving data values from other SAS data sets or from files that are stored in other software vendors' file formats.

scoring
the process of applying a model to new data in order to compute output. Scoring is the last process that is performed in data mining.

segmentation
the process of dividing a population into sub-populations of similar individuals. Segmentation can be done in a supervisory mode (using a target variable and various

techniques, including decision trees) or without supervision (using clustering or a Kohonen network).

singular value decomposition
a technique through which high-dimensional data is transformed into lower-dimensional data.

source-level debugger
an interactive environment in SAS that enables you to detect and resolve logical errors in programs that are being developed. The debugger consists of windows and a group of commands.

stemming
the process of finding and returning the root form of a word. For example, the root form of grind, grinds, grinding, and ground is grind.

stop list
a SAS data set that contains a simple collection of low-information or extraneous words that you want to remove from text mining analysis.

test data
currently available data that contains input values and target values that are not used during training, but which instead are used for generalization and model comparisons.

training data
currently available data that contains input values and target values that are used for model training.

validation data
data that is used to validate the suitability of a data model that was developed using training data. Both training data sets and validation data sets contain target variable values. Target variable values in the training data are used to train the model. Target variable values in the validation data set are used to compare the training model's predictions to the known target values, assessing the model's fit before using the model to score new data.

variable
a column in a SAS data set or in a SAS data view. The data values for each variable describe a single characteristic for all observations. Each SAS variable can have the following attributes: name, data type (character or numeric), length, format, informat, and label.

Index