



THE
POWER
TO KNOW.

Getting Started with SAS[®] Text Miner 4.2



The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2010. *Getting Started with SAS® Text Miner 4.2*. Cary, NC: SAS Institute Inc.

Getting Started with SAS® Text Miner 4.2

Copyright © 2010, SAS Institute Inc., Cary, NC, USA

ISBN (electronic book)

All rights reserved. Produced in the United States of America.

For a hardcopy book: No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

For a Web download or e-book: Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

U.S. Government Restricted Rights Notice: Use, duplication, or disclosure of this software and related documentation by the U.S. government is subject to the Agreement with SAS Institute and the restrictions set forth in FAR 52.227–19 Commercial Computer Software-Restricted Rights (June 1987).

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

1st printing, November 2010

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.

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

Contents

<i>Recommended Reading</i>	<i>v</i>
Chapter 1 • Introduction to Text Mining and SAS Text Miner 4.2	1
What Is Text Mining?	1
What Is SAS Text Miner?	2
The Text Mining Process	3
Accessibility Features of SAS Text Miner 4.2	4
Chapter 2 • Learning by Example: Using SAS Text Miner 4.2	5
About the Scenarios in This Book	5
Prerequisites for This Scenario	8
How to Get Help for SAS Text Miner	9
Chapter 3 • Setting Up Your Project	11
About the Tasks That You Will Perform	11
Create a Project	11
Create a Library	12
Create a Data Source	13
Create a Diagram	15
Chapter 4 • Analyzing the SYMPTOM_TEXT Variable	17
About the Tasks That You Will Perform	17
Identify Input Data	17
Partition Input Data	18
Set Text Miner Node Properties	18
View Interactive Results	21
Examine Data Segments	24
Chapter 5 • Cleaning Up Text	29
About the Tasks That You Will Perform	29
Use a Synonym Data Set	30
Create a New Synonym Data Set	31
Examine Results Using Merged Synonym Data Sets	34
Create a Stop List	36
Explore Results Improvements	39
Chapter 6 • Predictive Modeling with Text Variables	41
About the Tasks That You Will Perform	41
Use the COSTRING Variable to Model	41
Use the SYMPTOM_TEXT Variable to Model	45
Compare the Models	47
Additional Exercises	48
Chapter 7 • Using the Text Parsing Node	51
About the New Nodes	51
About the Text Parsing Node	51
About the Tasks That You Will Perform	52
Creating a Project	52
Creating a Data Source	52

Creating a Diagram	53
Using the Text Parsing Node	53
Chapter 8 • Using the Text Filter Node	55
About the Text Filter Node	55
About the Tasks That You Will Perform	55
Using The Text Filter Node	55
Chapter 9 • Using the Text Topic Node	61
About the Text Topic Node	61
About the Tasks That You Will Perform	61
Using the Text Topic Node	61
Chapter 10 • Tips for Text Mining	65
Processing a Large Collection of Documents	65
Dealing with Long Documents	65
Processing Documents from an Unsupported Language or Encoding	66
Chapter 11 • Next Steps: A Quick Look at Additional Features	67
The %TMFILTER Macro	67
Appendix 1 • Vaccine Adverse Event Reporting System Data Preprocessing	69
Index	75

Recommended Reading

- *Many of the concepts and topics that are discussed in additional product documentation for SAS Text Miner 4.2 (<http://support.sas.com/documentation/onlinedoc/txtminer>) and SAS Enterprise Miner (<http://support.sas.com/documentation/onlinedoc/miner>) might also help you use SAS Text Miner 4.2.*

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

SAS Publishing Sales
SAS Campus Drive
Cary, NC 27513
Telephone: 1-800-727-3228
Fax: 1-919-531-9439
E-mail: sasbook@sas.com
Web address: support.sas.com/bookstore

Chapter 1

Introduction to Text Mining and SAS Text Miner 4.2

What Is Text Mining?	1
What Is SAS Text Miner?	2
The Text Mining Process	3
Accessibility Features of SAS Text Miner 4.2	4

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. Additionally, 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?

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.

The Text Miner node encompasses the parsing and exploration aspects of text mining and prepares data for predictive mining and further exploration using other SAS Enterprise Miner nodes. The Text Miner node enables you to analyze structured text information, and combine the structured output of a Text Miner node with other structured data as desired. The Text Miner node is highly customizable and enables you to choose among a variety of parsing options. It is possible to parse documents for detailed information about the terms, phrases, and other entities in the collection. You can also cluster documents into meaningful groups and report concepts that you discover in the clusters. You can use the Text Miner node in an environment that enables you to interact with the collection. Sorting, searching, filtering (subsetting), and finding similar terms or documents all enhance the exploration process.

Also available are the Text Parsing, Text Filter, and Text Topic nodes. Each of these nodes performs a specific task of the text mining process. The Text Parsing node performs the same parsing operations as the Text Miner node and can be configured in much the same way. The Text Filter node enables you to remove terms that are deemed to have low information value or occur in too few documents to be relevant. The Text Topic node creates a set of topics based on the most highly correlated terms in the document collection. This is similar to the process of clustering the document collection that is done in the Text Miner node.

The Text Miner and Text Parsing nodes' extensive parsing capabilities include the following:

- stemming
- automatic recognition of multi-word terms
- normalization of various entities such as dates, currencies, percentages, and years
- part-of-speech tagging
- extraction of entities such as organizations, products, Social Security numbers, time, titles, and more
- support for synonyms
- language-specific analysis for Arabic, Chinese, Dutch, English, French, German, Italian, Japanese, Korean Polish, Portuguese, Spanish, and Swedish

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 very 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 Miner node or the Text Parsing node, and might contain the actual text or paths to the actual text.	%TMFILTER macro — a SAS macro for extracting text from documents and creating a predefined SAS data set with a text variable
Text parsing	Decomposes textual data and generates a quantitative representation suitable for data mining purposes.	Text Miner node, Text Parsing node
Transformation (dimension reduction)	Transforms the quantitative representation into a compact and informative format.	Text Miner node, Text Filter node
Document analysis	Performs classification, prediction, or concept linking of the document collection. Creates clusters or topics from the data.	Text Miner node, Text Topic node, or SAS Enterprise Miner predictive modeling nodes

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

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

Accessibility Features of SAS Text Miner 4.2

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.	<p>The software supports keyboard equivalents for all user actions with the exceptions noted below:</p> <p>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.</p> <p>The Explore action in the data source pop-up menu cannot be invoked directly from the keyboard, but there is an alternative way to invoke the data source explorer using the View ⇒ Explorer menu.</p>
Color coding shall not be used as the only means of conveying information, indicating an action, prompting a response, or distinguishing a visual element.	Supported with exception.	Node run or failure indication relies on color, but there is 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 4.2

About the Scenarios in This Book	5
Prerequisites for This Scenario	8
How to Get Help for SAS Text Miner	9

About the Scenarios in This Book

This book is divided into two examples, and each describes an extended scenario that is intended to familiarize you with the many features of SAS Text Miner. Within each example, the current topic builds on the previous topics, so you must work through the chapters in sequence. Several key components of the SAS Text Miner process flow diagram are covered. In these step-by-step examples, you learn to do basic tasks in SAS Text Miner, such as how to create a project and build 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 examples in this book are designed to be used in conjunction with SAS Text Miner software.

The first example illustrates the use of the Text Mining node and uses the Vaccine Adverse Event Reporting System (VAERS) data. This 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 anybody, but most reports come from vaccine manufacturers (42%) and health care providers (30%). 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. Please see the “Guide to Interpreting Case Report Information Obtained from the Vaccine Adverse Event Reporting System (VAERS)” available from HHS (<http://vaers.hhs.gov/data/index>).

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 8](#) 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 the table of VAERS data for 2005. 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	231786	01/01/2005	MA	63	63	.	F	01/01/2005
2	231787	01/02/2005	MD	30	30	.	F	01/02/2005
3	231788	01/02/2005	VA	18	18	.	F	01/02/2005
4	231789	01/02/2005	PA	1.3	1	0.3	M	01/02/2005
5	231790	01/02/2005	CA	16	15	.	M	01/02/2005
6	231791	01/02/2005		20	19	.	M	01/02/2005
7	231829	01/03/2005	DC	45	45	.	M	12/28/2004
8	231830	01/03/2005	TN	90	89	.	F	12/22/2004
9	231838	01/03/2005	CA	1.1	1	0.1	F	12/27/2004
10	231839	01/03/2005	LA	59	58	.	F	12/17/2004

In analyzing adverse reactions to medications, both in clinical trials and in post-release monitoring of reactions, keyword or word-spotting techniques combined with a thesaurus are most often used to characterize the symptoms. The Coding Symbols for Thesaurus of Adverse Reaction Terms (COSTART) has traditionally been the categorization technique of choice, but it has been largely replaced by the Medical Dictionary for Regulatory Affairs (MedDRA). COSTART is a term developed by the U.S. Food and Drug Administration (FDA) for the coding, filing, and retrieving of post-marketing adverse reports. It provides a keyword-spotting technique that deals with the variations in terms used by those who submit adverse event reports to the FDA.

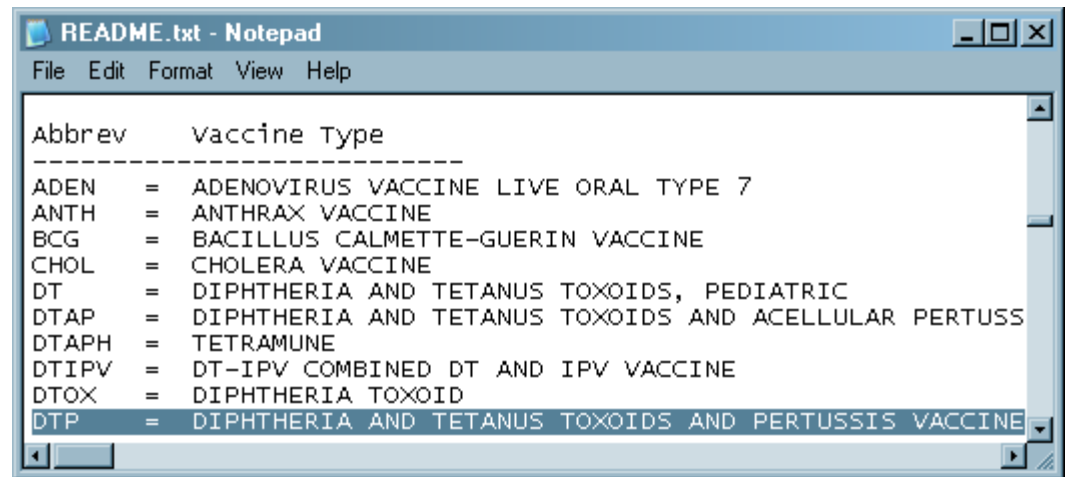
In the case of vaccinations, the COSTART system has been used. The FDA has used a program to extract COSTART categories from the SYMPTOM_TEXT column. Here are some of the variables used by the program:

- SYMPTOM_TEXT — reported symptom text
- SYM01- SYM20 — extracted COSTART categories
- SYM_CNT — number of SYM fields that are populated for a particular vaccination
- VAERS_ID — VAERS identification number

If you open the VAERS data for 2005, you can see that VAERS_ID **231844** has SYMPTOM_TEXT of **101 fever, stiff neck, cold**. The program has automatically extracted the COSTART terms that appear in column SYM01 to column SYM20 in the data file.

The VAERS table contains other columns, including a variety of flags that indicate the seriousness of the event (life-threatening illness, emergency room or doctor visit, hospitalized, disability, and recovered), the number of days after the vaccine that the event occurred, how many different vaccinations were given, and a list of codes (VAX1-VAX8) for each of the shots given. There are also columns that indicate where the shots were given, who funded them, what medications the patient was taking, and so on.

The README file taken from the VAERS Web site decodes the vaccine abbreviations. Note that some vaccinations contain multiple vaccines (for example, DTP contains diphtheria, tetanus, and pertussis). Here is a portion of the README file:

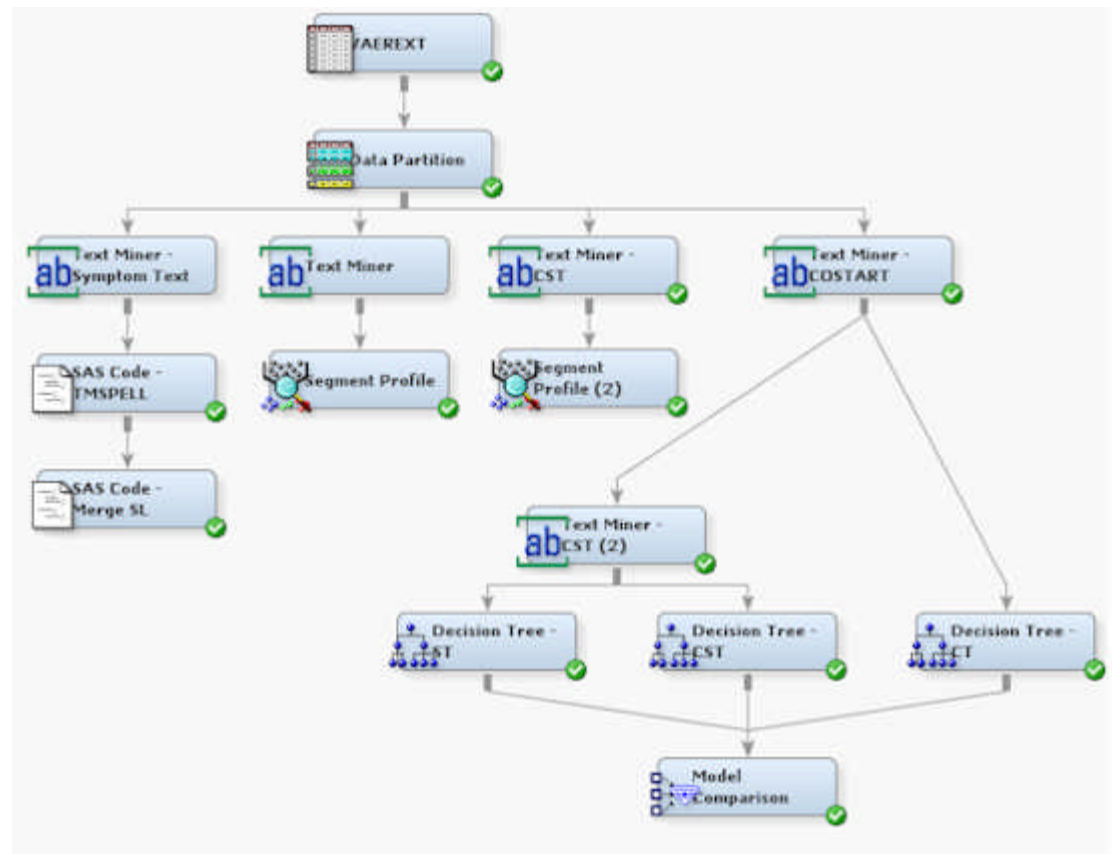


Abbrev	Vaccine Type
ADEN	= ADENOVIRUS VACCINE LIVE ORAL TYPE 7
ANTH	= ANTHRAX VACCINE
BCG	= BACILLUS CALMETTE-GUERIN VACCINE
CHOL	= CHOLERA VACCINE
DT	= DIPHTHERIA AND TETANUS TOXOIDS, PEDIATRIC
DTAP	= DIPHTHERIA AND TETANUS TOXOIDS AND ACELLULAR PERTUSS
DTAPH	= TETRAMUNE
DTIPV	= DT-IPV COMBINED DT AND IPV VACCINE
DTOX	= DIPHTHERIA TOXOID
DTP	= DIPHTHERIA AND TETANUS TOXOIDS AND PERTUSSIS VACCINE

As you go through this example, imagine you are a researcher trying to discover what information is contained within this data set and 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, an unsanitized needle). Some of the adverse reactions 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, and find answers to the following questions:

- What are some categories of reactions that people are experiencing?
- How do these relate to the vaccination that was given, the age of the recipient, the place that they received the vaccine, or other pertinent information?
- What factors influence whether a reaction becomes serious?
- How well are these factors captured by the automatically extracted COSTART terms?
- Is there any important information contained in the adverse reaction text that is not represented by the COSTART terms?

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



After you complete the VAERS data example, you will use the Text Parsing, Text Filter, and Text Topic nodes to analyze the abstracts from a collection of SAS Users Group International (now called the SAS Global Forum) papers. The goal of this example is to determine whether any themes are present in the papers. You will use the Text Topic node to create a set of topics that will describe the document collection. Additionally, you will create a user-defined topic that finds all abstracts related to dynamic Web pages. This example is independent of the VAERS data example and will result in a simpler process flow diagram.

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 4.2. You must also perform the following:

1. Download the Getting Started Examples zip file under the SAS Text Miner 4.2 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 the environment in which you are viewing them.

How to Get Help for SAS Text Miner

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	11
Create a Project	11
Create a Library	12
Create a Data Source	13
Create a Diagram	15

About the Tasks That You Will Perform

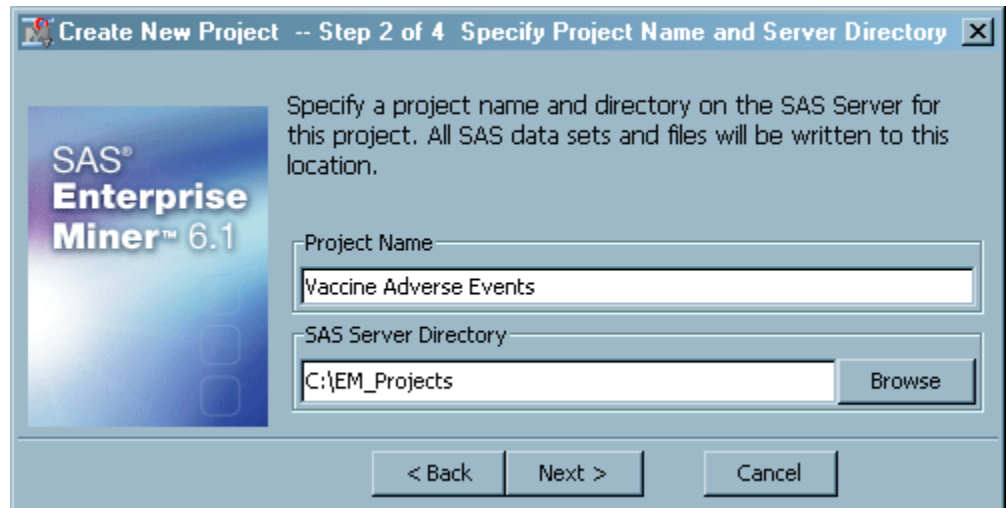
To set up your project, perform the following main tasks:

1. Create a new project where you will store all your work.
 2. Define the VAERS data as a SAS Enterprise Miner data source.
 3. Create a new process flow diagram in your project.
-

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 opens.
3. Click **Next**. The Specify Project Name and Server Directory page opens.



4. Type a name for the project, such as **Vaccine Adverse Events**, in the Project Name box.
5. In the SAS Server Directory box, type the path to the location on the server where you want to store data for your project. Alternatively, browse to a folder to use for your project.

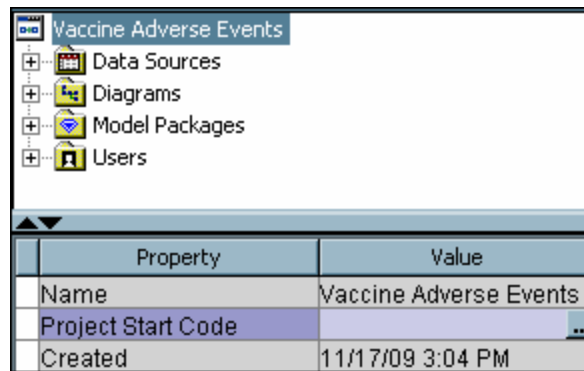
Note: The project path depends on whether you are running SAS Enterprise Miner as a complete client on your local machine or as a client/server application. If you are running SAS Enterprise Miner as a complete client, your local machine acts as its own server. Your SAS Enterprise Miner projects are 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. This example uses **C:\EM_Projects**.

6. Click **Next**. The Register the Project page opens.
7. Click **Next**. The New Project Information page opens.
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.



- Click the button for the Project Start Code property. The Project Start Code dialog box opens.
- Select the **Code** tab and enter the following code to create a SAS library:

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

Note: The location will depend on where you have stored the data for this tutorial on your system.
- Click **Run Now**.
- 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 Data Source

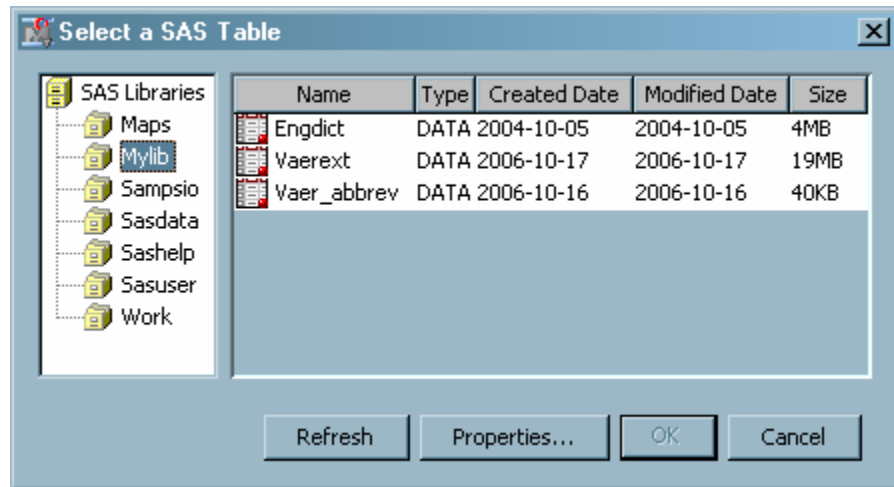
To create a data source:

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

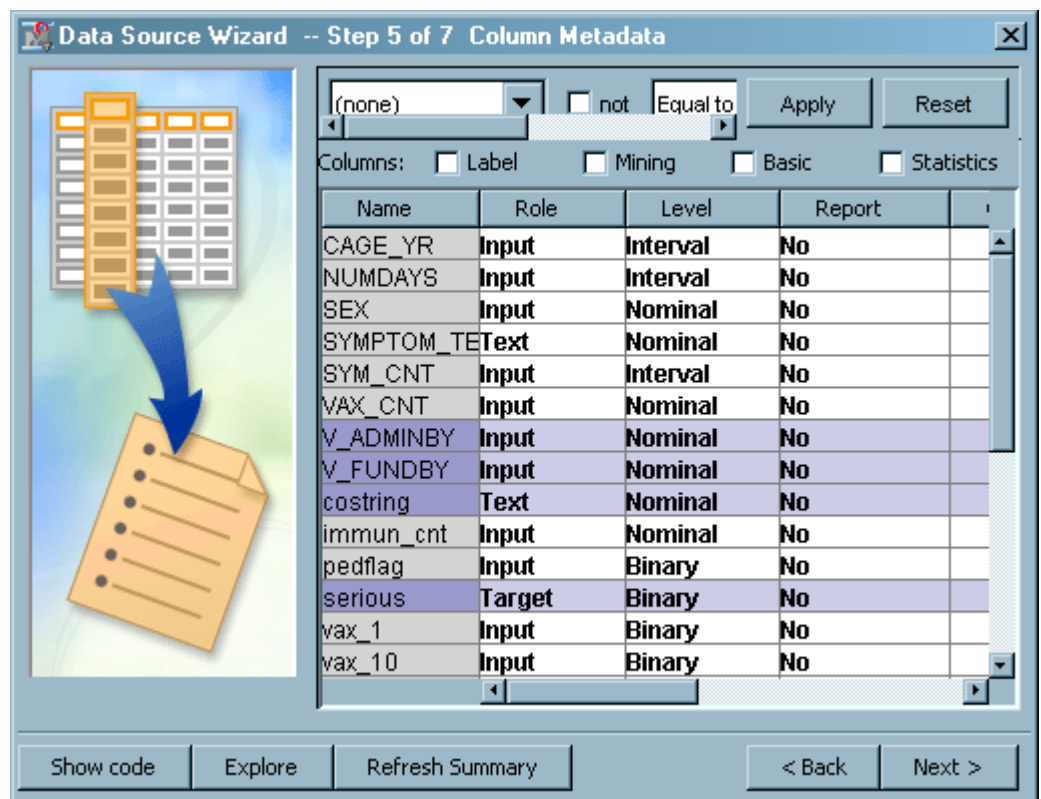


- Select **SAS Table** in the Source drop-down menu of the Metadata Source page.
- Click **Next**. The Select a SAS Table page opens.
- Click **Browse**.
- Click the SAS library named **Mylib**. The Mylib library folder contents are displayed on the Select a SAS Table dialog box.

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



6. Select the **VAEREXT** table, and then click **OK**. The two-level name MYLIB.VAEREXT is displayed in the Table box of the Select a SAS Table page.
7. Click **Next**. The Table Information page opens. The Table Properties pane displays metadata for you to review.
8. Click **Next**. The Metadata Advisor Options page opens.
9. Select **Advanced**, and then click **Next**. The Column Metadata page opens.



10. 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 **costring** to **Text**.

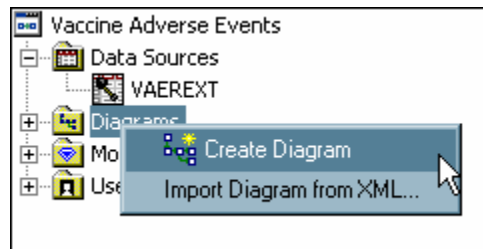
- Set the role for **serious** to **Target**.
11. Click **Next**. The Decision Configuration page opens.
 12. Click **Next**. The Create Sample page opens.
 13. Click **Next**. The Data Source Attributes page opens.
 14. Click **Next**. The Summary page opens.
 15. Click **Finish**. The VAEREXT table is added to the Data Sources folder in the Project Panel.



Create a Diagram

To create a diagram, complete the following steps:

1. Right-click the Diagram folder in the Project Panel and select **Create Diagram**. The Create New Diagram dialog box opens.



2. Type **VAERS Example** in the Diagram Name box.
3. Click **OK**. The empty VAERS Example diagram opens in the diagram workspace.

Chapter 4

Analyzing the SYMPTOM_TEXT Variable

About the Tasks That You Will Perform	17
Identify Input Data	17
Partition Input Data	18
Set Text Miner Node Properties	18
View Interactive Results	21
Examine Data Segments	24

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 data source with an Input Data node.
2. Partition the input data using the Data Partition node.
3. Set Text Miner node properties using the Properties panel, and run the Text Miner node.
4. View the results using the Interactive Results window.
5. Use the Segment Profile node to examine data segments.

Identify Input Data

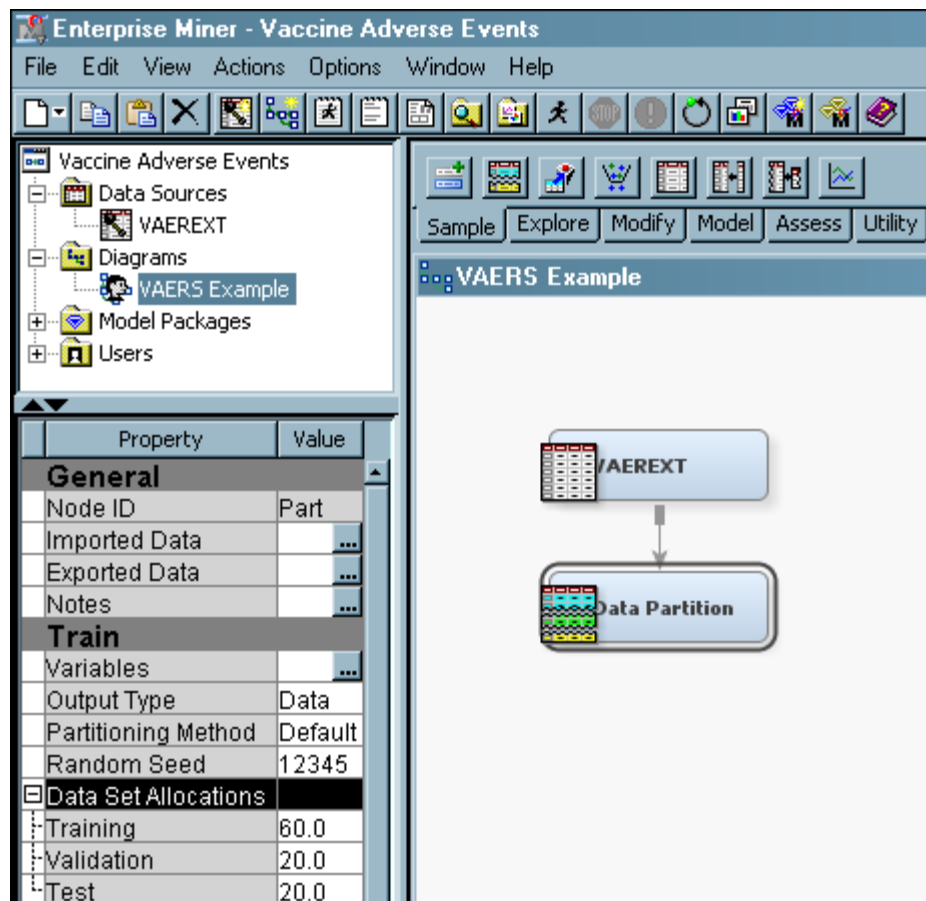
To identify input data:

1. Select the **VAEREXT** data source from the Data Sources list in the Project Panel.
2. Drag and drop VAEREXT into the diagram workspace to create an Input Data node.

Partition Input Data

To partition the input data:

1. Select the **Sample** tab from the node toolbar and drag a Data Partition node into the diagram workspace. Connect the VAEREXT Input Data node to the Data Partition node.



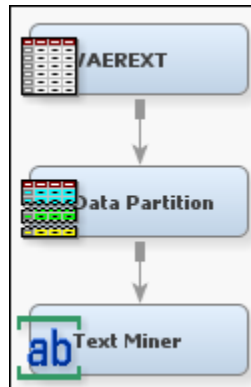
2. Select the Data Partition node to view its properties. Details about the node appear in the Properties panel. Set the Data Set Allocations properties as follows:
 - Set the **Training** property to **60.0**.
 - Set the **Validation** property to **20.0**.
 - Set the **Test** property to **20.0**.

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



Set Text Miner Node Properties

To set the Text Miner node properties:

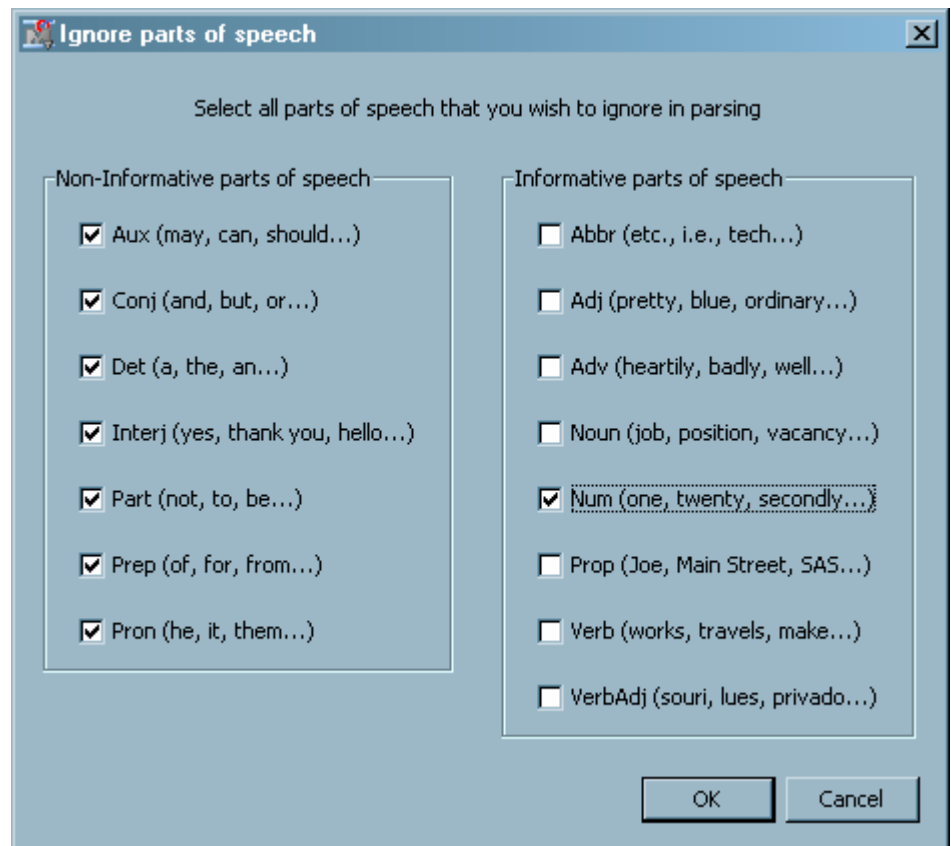
1. Select the **Text Mining** tab on the toolbar and drag and drop a Text Miner node into the diagram workspace. Connect the Data Partition node to the Text Miner node.



2. Select the Text Miner node to view its properties. Details about the node appear in the Properties panel. Set the following Parse properties:

- Set **Terms in Single Document** to **Yes** to include all terms that occur only in a single document.
- Set **Different Parts of Speech** to **No**. For the VAERS data, this setting offers a more compact set of terms.
- Click the  button for the Synonyms property. The Select a SAS Table window opens. Select **No data set to be specified**. Click **OK**.
- Click the  button for the Ignore Parts of Speech property, and select the following items, which represent parts of speech:
 - **Aux**
 - **Conj**
 - **Det**
 - **Interj**
 - **Part**
 - **Prep**
 - **Pron**
 - **Num**

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.



Click **OK**.

3. Set **Term Weight** to **Mutual Information** so that terms will be differentially weighted when they correspond to serious reactions.

Transform	
Compute SVD	Yes
SVD Resolution	Low
Max SVD Dimensions	100
Scale SVD Dimensions	No
Frequency weighting	Log
Term Weight	Mutual Information
Roll up Terms	No
No. of Rolled-up Terms	100
Drop Other Terms	No

4. Set the following Cluster properties:
 - Set **Automatically Cluster** to **Yes** to answer the question: "What are some categories of reactions that people are experiencing?" You want to categorize these adverse events.
 - Set **Descriptive Terms** to **12** to ease cluster labeling.
 - Set **Ignore Outliers** to **Yes**.


Cluster	
Automatically Cluster	Yes
Exact or Maximum Number	Maximum
Number of Clusters	40
Cluster Algorithm	EXPECTATION-MAXIMIZATION
Ignore Outliers	Yes
Hierarchy Levels	.
Descriptive Terms	12
What to Cluster	SVD Dimensions



5. Right-click the Text Miner node in the diagram workspace, and select **Run**.
6. Click **Yes** in the Confirmation dialog box when you are prompted with a question that asks whether you want to run the path.
7. Click **OK** in the Run Status dialog box that appears after the Text Miner node has finished running. The Text Miner node Parse Variable property has been populated with the SYMPTOM_TEXT variable.

Parse	
Parse Variable	SYMPTOM_TEXT

View Interactive Results

To view interactive results, complete the following steps:

1. Select the Text Miner node, and then Click the  button for the Interactive property. The Text Miner — Interactive window opens.

Train	
Variables	
Interactive	
Force Run	No

2. 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.

Terms					
	TERM	FREQ	# DOCS	KEEP	WEIGHT
+	receive	5193	3691	<input checked="" type="checkbox"/>	0.059
+	vaccine	4637	3431	<input checked="" type="checkbox"/>	0.0090
+	have	4950	3395	<input checked="" type="checkbox"/>	0.092
+	swell	4028	3331	<input checked="" type="checkbox"/>	0.223
+	day	4241	3221	<input checked="" type="checkbox"/>	0.0040
+	arm	3982	2900	<input checked="" type="checkbox"/>	0.194
+	no.	3886	2747	<input checked="" type="checkbox"/>	0.113
	pt	4471	2519	<input checked="" type="checkbox"/>	0.046
+	fever	3000	2512	<input checked="" type="checkbox"/>	0.021
+	leave	3142	2476	<input checked="" type="checkbox"/>	0.183
+	site	2688	2263	<input checked="" type="checkbox"/>	0.244
+	injection	2742	2197	<input checked="" type="checkbox"/>	0.204
+	report	3056	2177	<input checked="" type="checkbox"/>	0.045
+	patient	3730	2097	<input checked="" type="checkbox"/>	0.116
+	give	2500	2046	<input checked="" type="checkbox"/>	0.078

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

- View the documents in the Documents window. Click the Toggle Show Full Text icon



on the toolbar to see the full text contained in SYMPTOM_TEXT.

Documents		
SYMPTOM_TEXT	_DATAOBS_ ▲	COSTRING
Information has been received from an RN concerning a 64 year old white, obese female who on 11/14/01, at 11:00 AM, was vaccinated IM in the left deltoid with a dose of pneumococcal vaccine 23 polyvalent (lot 637263/1448K) . Within the 1st 24 to 36 hours, she developed a fever. She also awoke in the middle of the night with swelling and redness at the injection site and the skin was hot to the touch and tender. It was approx. the size of a 50 cents piece. It was reported that the pt temperat	1.0	ANXIETY ...
Sabin tri vaccines were not good ones. They make you taller and handicapped looking.	4.0	REACT UN...
Cellulitis at administration site.	5.0	CELLULITI...
Demyelinating disease; dizziness, blurred vision; difficulty hearing and walking.	7.0	AMBLYOPI...
Autistic mannerisms, system "shutdown". Blank stares, catatonic state.	11.0	AUTISM C...
Loss of speech and coordination.	12.0	COORDIN...
Reportedly called in after first dose to report had a rash that sounded like hives 2 days after immunization. Very itchy. Went		

- View the clusters in the Clusters window.


Clusters				
#	DESCRIPTIVE TERMS	FREQ	PERCENTAGE	RMS STD.
1	+ injection site, redness, + injection, + site, erythema, + reaction, + pain, + female, + develop, + report, + dose, + patient	926	0.0732421102...	0.0794709...
2	+ month, + find, + admit, + hospital, + seizure, + have, + do, + immunization, not, + child, + receive, + vaccine	1229	0.0972079411...	0.1528825...
3	+ fever, + have, + febrile seizure, + hospital, + minute, + seizure, + admit, febrile, + hour, + last, + child, + call	444	0.0351182472...	0.1113928...
4	left arm, + deltoid, right, + pain, + arm, + leave, erythema, + swell, pt, + give, + area, not	1664	0.1316143320...	0.1253185...
5	+ give, + rash, + hive, + call, + body, + itch, + back, benadryl, trunk, face, + start, + day	864	0.0683382108...	0.0785251...
6	+ female, male, + year, + concern, + history, + old, + patient, + month, + vaccinate, + lot, + include, + dose	1427	0.1128687811...	0.1249742...

- 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 **Filter Terms**.

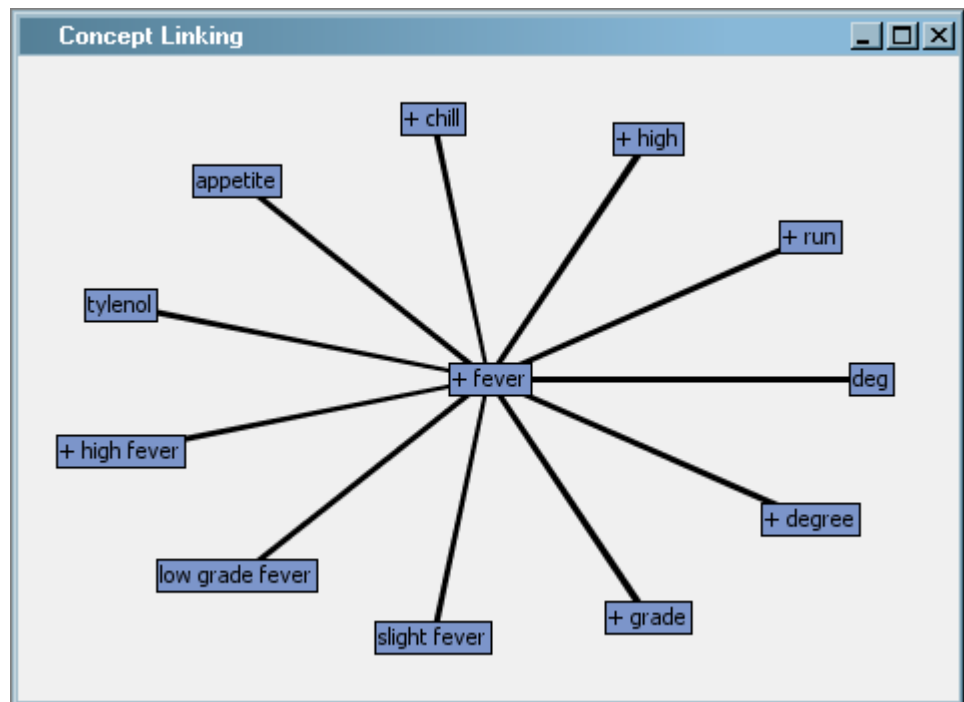
Terms					
	TERM	FREQ ▼	# DOCS	KEEP	WEIGHT
+	patient	3730	2097	<input checked="" type="checkbox"/>	0.115
+	leave	3142	2476	<input checked="" type="checkbox"/>	0.183
+	report	3056	2177	<input checked="" type="checkbox"/>	0.045
+	fever	3888	2512	<input checked="" type="checkbox"/>	0.021
+	injection			<input checked="" type="checkbox"/>	0.204
+	site			<input checked="" type="checkbox"/>	0.244
+	give			<input checked="" type="checkbox"/>	0.078
+	pain			<input checked="" type="checkbox"/>	0.063
+	develop			<input checked="" type="checkbox"/>	0.021
	not			<input checked="" type="checkbox"/>	0.045


- Note how the documents displayed and cluster frequencies change. Only those documents containing **fever** are displayed. Moreover, only the documents containing **fever** are counted. If the full text of the document is not shown, click the Toggle Show

Full Text icon  on the toolbar.

- Click the Undo icon  on the toolbar. This action removes the filter that was applied and restores the display that was shown when you opened the Text Miner — Interactive window.
- Select **fever** in the Terms window, and then right-click **fever** and select **View Concept Links**. The Concept Linking window opens. 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, right-click on any of the terms that are not in the center of the graph and select **Expand Links**.



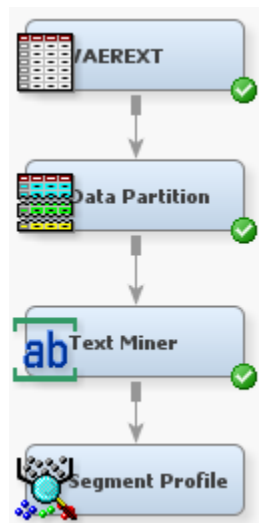
9. Look at the clusters in the Clusters window. Can you tell what they are about from the descriptive terms displayed? Do some clusters look vague or unclear?
10. Choose one of the clusters that looks vague or unclear. This is fairly subjective, but, for this example, you can use Cluster 2 as an example of a vague or unclear cluster. Right-click on Cluster 2 and select **Filter Clusters**. This action filters the results to show only those documents and terms that are relevant to Cluster 2. All the documents shown in the Documents window are contained in Cluster 2, and terms are now ordered by frequency within that cluster. Read the text of some of the documents in this cluster. Does this clarify the cluster better?
11. Click the Undo icon  on the toolbar to undo any filters.
12. Close the Text Miner — Interactive 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 derived analytically 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.

To examine data segments, complete the following steps:

1. From the **Assess** tab, drag and drop a Segment Profile node into the diagram workspace and connect the Text Miner node to the Segment Profile node.



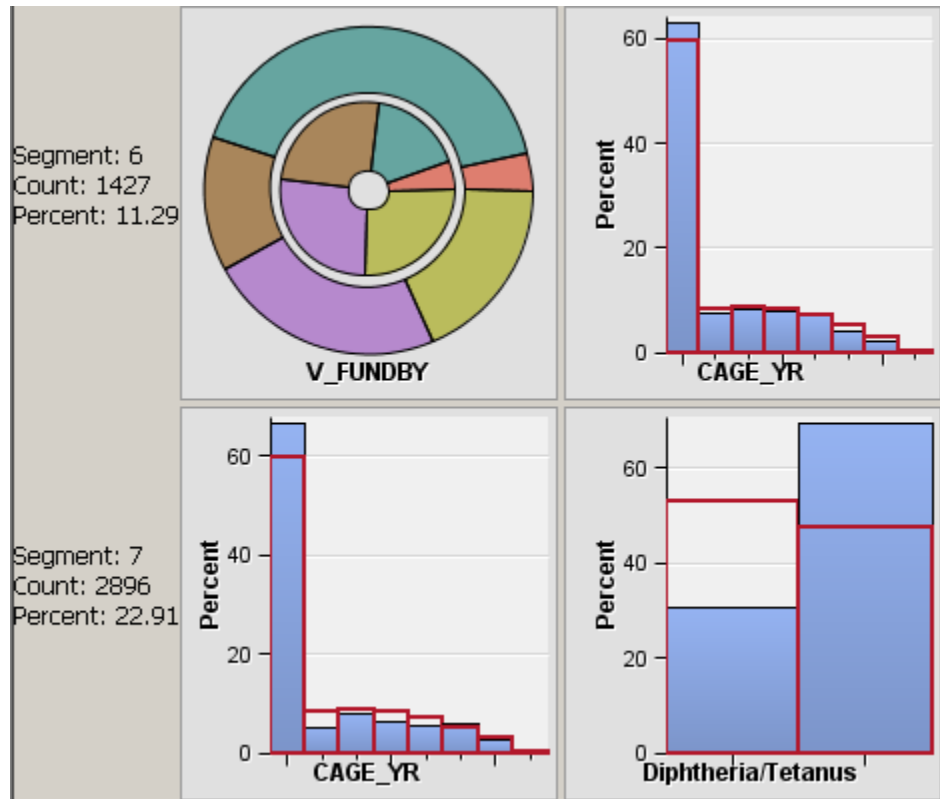
2. Select the Segment Profile node. Select the  button for the **Variables** property. The Variables — Prof window opens.
3. Select all the PROB variables and set their Use value to **No**.

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.

Name	Use	Report
CAGE_YR	Default	No
NUMDAYS	Default	No
PROB1	No	No
PROB10	No	No
PROB2	Default	No
PROB3	No	No
PROB4	Yes	No
PROB5	No	No
PROB6	No	No
PROB7	No	No
PROB8	No	No
PROB9	No	No
SEX	Default	No

4. Select all the **_SVD_** variables and set their **Use** value to **No**.
Note: You can hold down Shift and select all the **_SVD_** variables by clicking on the first **_SVD_** variable and dragging the pointer to select all **_SVD_** variables. After all **_SVD_** variables are selected, you can change the Use value of each selected **_SVD_** variable by changing the Use value of one of the **_SVD_** variables. This will change the other **_SVD_** Use values to the selected value as well.
5. Click **OK**.
6. Select the Segment Profile node in the diagram workspace. In the Properties panel, set the **Minimum Worth** property to **0.0010**.

7. Right-click the Segment Profile node, and select **Run**.
8. Click **Yes** in the Confirmation dialog box. After the node finishes running, click **Results** in the Run Status dialog box.
9. Maximize the Profile: _CLUSTER_ window. The following shows a portion of this window.

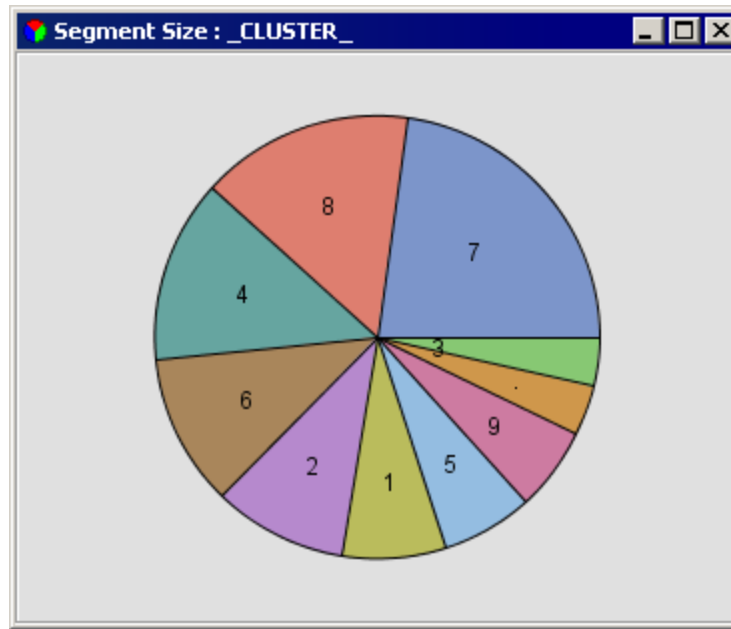


The Profile: _CLUSTER_ window displays a lattice, or grid, of plots that compare the distribution for the identified and report variables for both the segment and the population. The graphs shown in this window illustrate variables that have been identified as factors that distinguish the segment from 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.

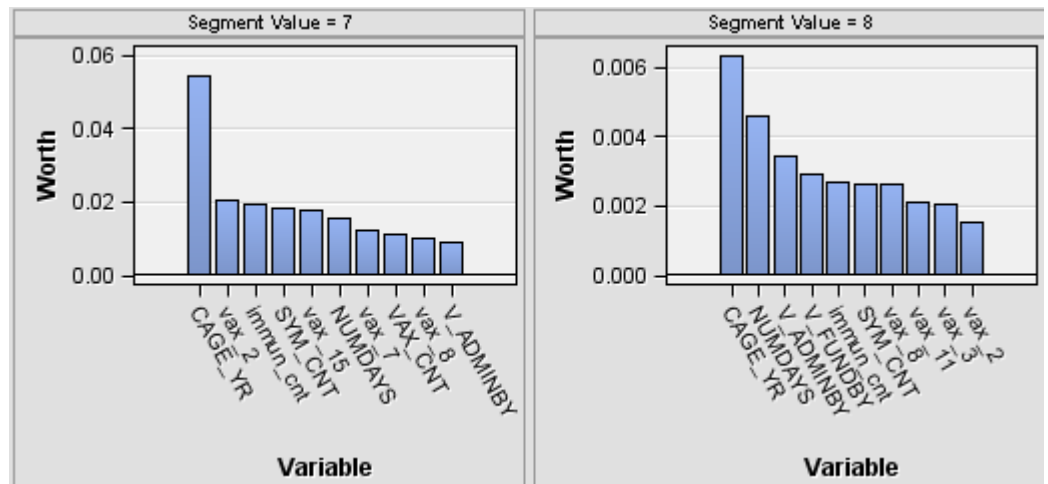
The columns are organized from left to right according to their ability to discriminate that segment from the population. Report variables, if specified, appear on the right in alphabetical order after the selected inputs. The lattice graph has the following features:

- Class variable — displays 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 — displays 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.

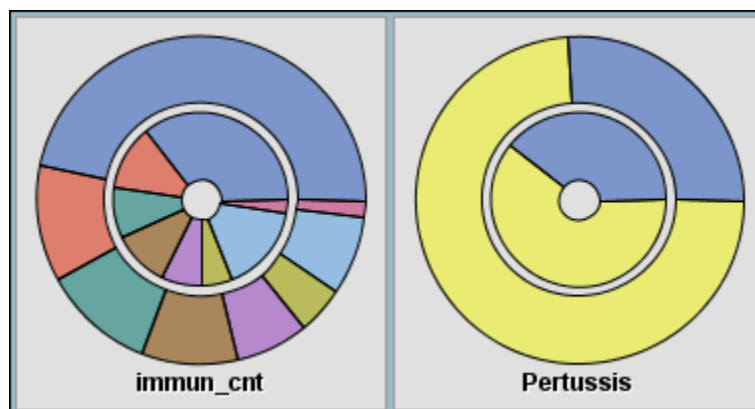
10. Maximize the Segment Size: _CLUSTER_ window. The following shows a portion of this window.



11. Maximize the Variable Worth: `_CLUSTER_` window. The following shows a portion of this window.



12. Note the strong relationships between some of the vaccinations given and the clustered categories. You can think of the "wheels" or concentric rings as follows: the inner circle represents all the adverse events, while the outer circle contains only the adverse events in that cluster.



13. Close the **Results** window.

Chapter 5

Cleaning Up Text

About the Tasks That You Will Perform	29
Use a Synonym Data Set	30
Create a New Synonym Data Set	31
Examine Results Using Merged Synonym Data Sets	34
Create a Stop List	36
Explore Results Improvements	39

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 clean the data to get better results.

The README.TXT file provided on the VAERS site contains a list of abbreviations 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 4.2 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/CAT was marked as equivalent to computerized axial tomography. For more information about the preprocessing steps, see “[Vaccine Adverse Event Reporting System Data Preprocessing](#)” on page 69.

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:

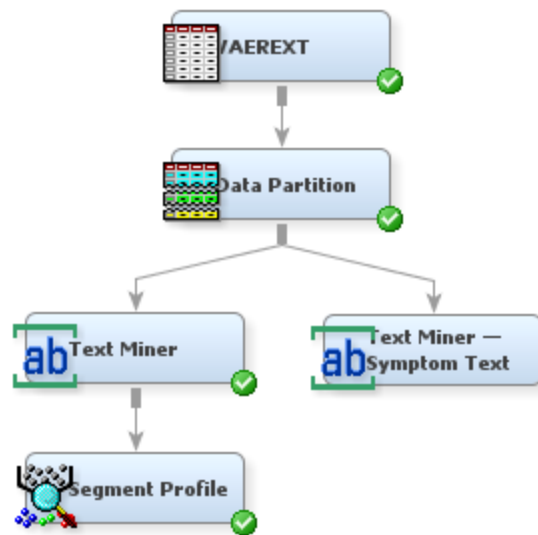
1. Use a synonym data set from the Getting Started with SAS Text Miner 4.2 zip file.
2. Create a new synonym data set using the SAS Code node and the TMSPELL procedure. The TMSPELL procedure will make a pass through all the terms, automatically identify which ones are misspellings, and create synonyms that map correctly spelled terms to the misspelled terms.
3. Examine results using merged synonym data sets.



4. Create a stop list to define which words are removed from the analysis. A **stop list** is a collection of low-information or extraneous words—previously saved as a SAS data set—that you want to remove from the text.
5. Explore whether cleaning the text improved the clustering results.

Use a Synonym Data Set

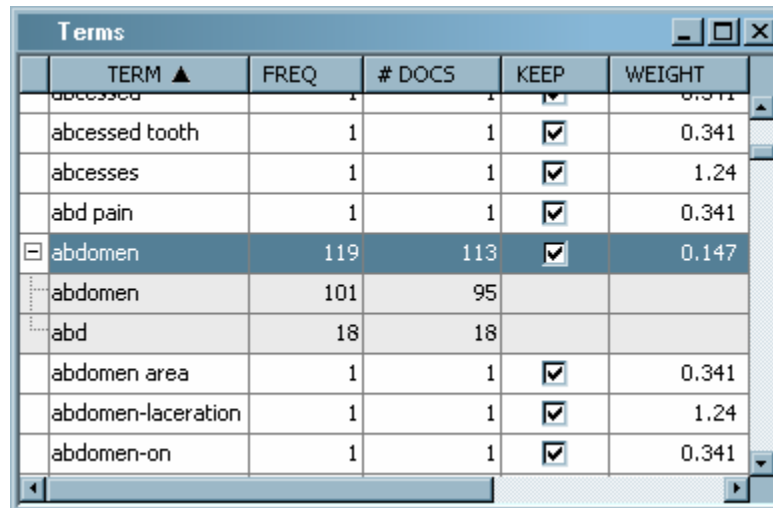
To use a synonym data set:

1. Right-click the Text Miner node in your process flow diagram and select **Copy**. For this example, it is important to copy the node instead of creating a new Text Miner node. When you copy a node, the settings you previously specified in the Text Miner node Properties panel will be used. Right-click in the empty diagram workspace and select **Paste**.
2. To distinguish this newly pasted Text Miner node from the first node, right-click it and select **Rename**. Type **Text Miner — Symptom Text** in the Node Name dialog box, and then click **OK**.
3. Connect the Data Partition node to the Text Miner — Symptom Text node.



4. Select the Text Miner — Symptom Text node in the diagram workspace. In the Properties panel, click the  button for the Synonyms property. The Select a SAS Table dialog box opens.
5. Select the **Mylib** library to view its contents. Select **VAER_ABBREV**, and then click **OK**.
6. Leave all other settings the same as in the original Text Miner node.
7. Right-click the Text Miner — Symptom Text node in the diagram workspace, and select **Run**. Click **Yes** in the Confirmation dialog box. Click **OK** in the Run Status dialog box when the node has finished running.
8. Click the  button for the Interactive property to open the Text Miner — Interactive window.

9. Click the TERM column heading to sort the Terms table.
10. Select **abdomen** under the TERM column in the Terms window. The term abdomen is one of the terms on the right side of the MYLIB.VAER_ABBREV table. In the Terms window, there should be a plus (+) sign next to **abdomen**. Click on the plus sign to expand the term and to show all synonyms and stems that are mapped to that term. A stem is the root form of a term. Make sure that the child term **abd** is included. Both **abdomen** and **abd** will be treated the same.



TERM ▲	FREQ	# DOCS	KEEP	WEIGHT
abcessed	1	1	<input checked="" type="checkbox"/>	0.341
abcessed tooth	1	1	<input checked="" type="checkbox"/>	0.341
abcesses	1	1	<input checked="" type="checkbox"/>	1.24
abd pain	1	1	<input checked="" type="checkbox"/>	0.341
abdomen	119	113	<input checked="" type="checkbox"/>	0.147
abdomen	101	95		
abd	18	18		
abdomen area	1	1	<input checked="" type="checkbox"/>	0.341
abdomen-laceration	1	1	<input checked="" type="checkbox"/>	1.24
abdomen-on	1	1	<input checked="" type="checkbox"/>	0.341

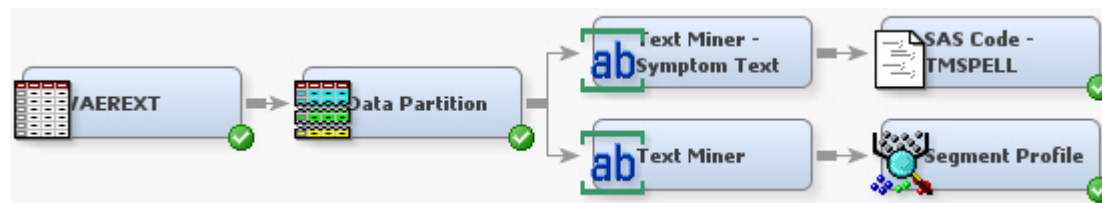
11. Close the Text Miner — Interactive window.

Create a New Synonym Data Set

You can use the SAS Text Miner TMSPELL procedure to create a new synonym data set. The TMSPELL procedure 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 and drag a SAS Code node into the diagram workspace. Connect the Text Miner — Symptom Text node to the SAS Code node. Right-click the SAS Code node, and select **Rename**. Type **SAS Code — TMSPELL** in the Node Name box. Click **OK**.




2. Select the arrow that connects the Text Miner — Symptom Text node to the SAS Code — TMSPELL node. Note the value of the Terms export **Table** property. You will use this value in the TERMDS= parameter in the next step.

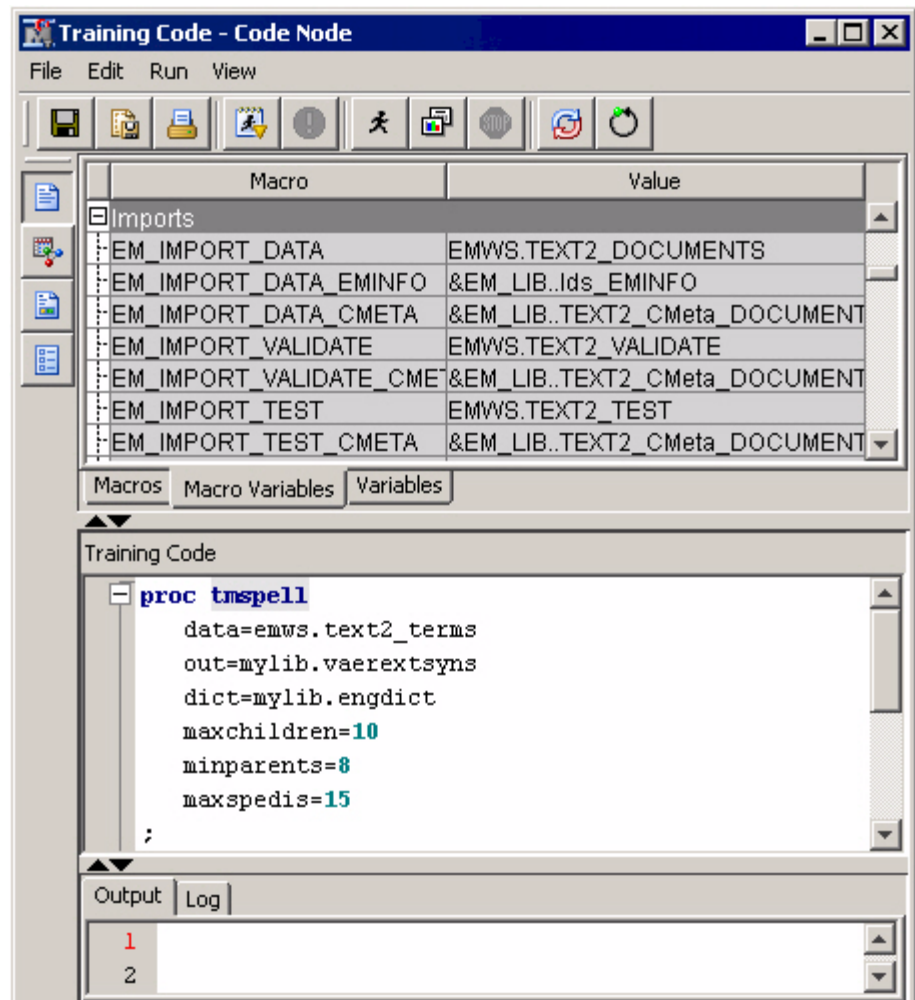
Note: The libref EMWS in the TERMS Table property is dependent upon the diagram number within your SAS Enterprise Miner project. If your diagram is the first one



created, then the libref will be EMWS, the second diagram will be EMWS1, the third will be EMWS2, and so on.

Property	Value
From	TEXT2
To	EMCODE
Table	EMWS.TEXT2_DOCUMENT ...
Variables	...
Role	Train
Table	EMWS.TEXT2_VALIDATE ...
Variables	...
Role	Validate
Table	EMWS.TEXT2_TEST ...
Variables	...
Role	Test
Table	EMWS.TEXT2_TERMS ...
Variables	...
Role	Terms
Table	EMWS.TEXT2_CLUSTER ...
Variables	...
Role	Cluster
Table	EMWS.TEXT2_OUT ...
Variables	...
Role	Transaction

3. Select the SAS Code — TMSPELL node, and click the  button for the **Code Editor** property in the Properties panel.
4. Enter the following code in the Code Editor:

```
proc tmspell
  data=emws.text2_terms
  out=mylib.vaerextsyms
  dict=mylib.engdict
  maxchildren=10
  minparents=8
  maxspedis=15
;
```



5. Click the  button to save the changes.
6. Click the  button to run the SAS Code — TMSPELL node. Click **Yes** in the Confirmation dialog box.
7. Click **OK** in the dialog box that indicates that the node has finished running.
8. Close the Training Code — Code Node window.
9. From the SAS Enterprise Miner window, select **View** ⇒ **Explorer**. The Explorer window opens.
10. Click **Mylib**, and then select **Vaerextsyms**.

Note: If the **Mylib** library is already selected and you do not see the Vaerextsyms data set, you might need to click **Get Details** or refresh the Explorer window to see the **Vaerextsyms** data set.
11. Double-click the Mylib.Vaerextsyms table to examine it.

	numdocs	term	childndocs	parent	category	minsped	dic
42	11	vasulitis	1	vasculitis		5	N
43	13	vasova...	1	vasovagal		11	N
44	59	varricel...	1	varicella...	NOUN_GR...	2	N
45	440	varricella	1	varicella		2	N
46	10	valv	1	valve		8	N
47	440	vaicella	1	varicella		6	N
48	222	vacinee	1	vaccinee		3	N

Here is a list of what the Vaerextsyms columns provide:

- Term is the misspelled word.
- Parent is a guess at the word that was meant.
- Childndocs is the number of documents that contained that term.
- # Documents 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 52 shows **abdomin** to be a misspelling of **abdominal**. Three documents contain **abdomin**, while 77 documents contain the parent, **abdominal** (this is not shown in the image). The term **abdomin** is not a legitimate English word, and an example text that contains that misspelling is **20 mins later, upper !!abdomin!!**. Note that double exclamation marks (!!) both precede and succeed the child term in the example text so you can see the term in context.

12. Examine the Vaerextsyms table to see whether you disagree with some of the choices made. For this example, however, assume that the TMSPELL macro has done a good enough job detecting misspellings.

Note: The Vaerextsyms 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.


13. Close the Mylib.Vaerextsyms table and the Explorer window.

Examine Results Using Merged Synonym Data Sets

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

1. Select the **Utility** tab and drag a SAS Code node into the diagram workspace. Connect the SAS Code — TMSPELL node to the new SAS Code node. Right-click the new


SAS Code node, select **Rename**, and type **SAS Code — Merge SL**, where “SL” stands for “Synonym List,” in the Node Name box. Click **OK**.

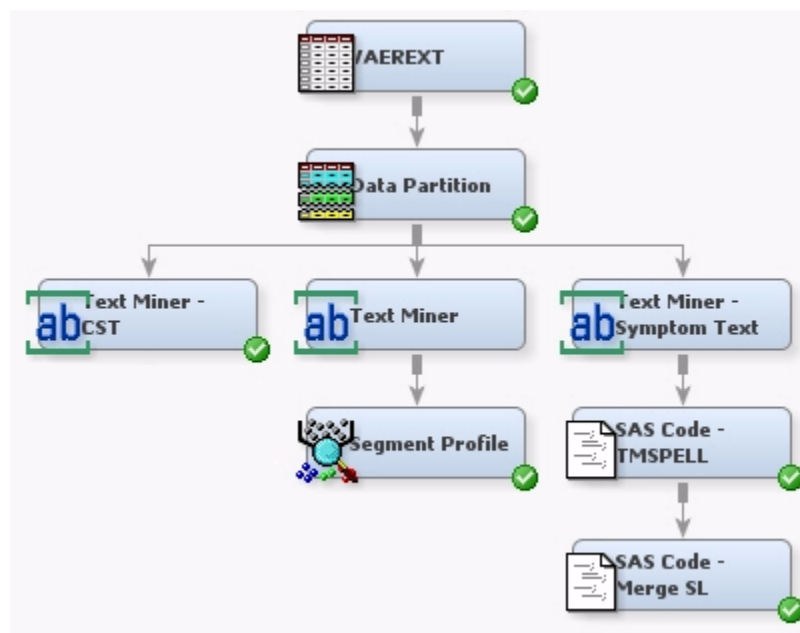
2. Select the SAS Code — Merge SL node and click the  button for the **Code Editor** property. The Code Editor opens.

3. Enter the following code in the Code Editor:



```
data mylib.vaerextsyms_new;
    set mylib.vaerextsyms mylib.vaer_abbrev;
run;
```

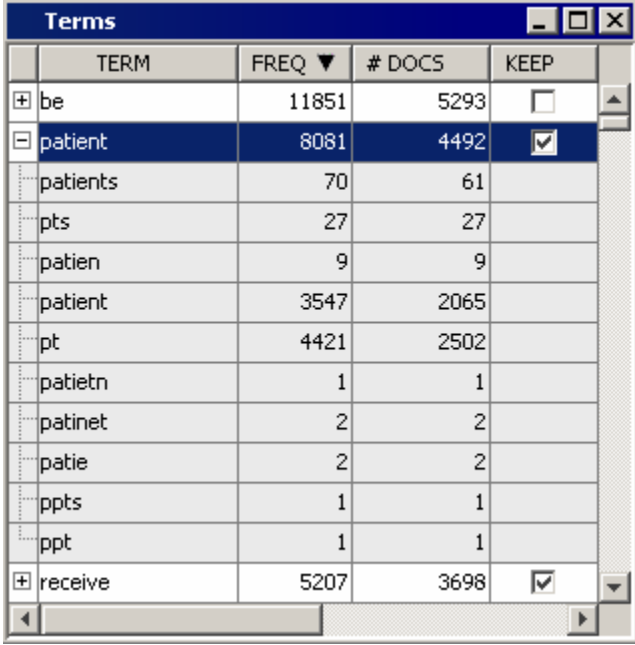
This code merges the resulting synonyms data set from the first SAS Code — TMSPELL node with the abbreviations data set.

4. Click the  button to save changes. Close the Code Editor window.
 5. Right-click the SAS Code — Merge SL node, and select **Run**. Click **Yes** in the Confirmation dialog box. Click **Results** in the Run Status dialog box when the node has finished running.
 6. 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.
7. Right-click the Text Miner — Symptom Text node and select **Copy** from the menu. Right-click an empty space in the diagram workspace and select **Paste**. It is important to copy the Text Miner — Symptom Text node instead of creating a new Text Miner node in order to keep the same property settings you previously configured for the Text Miner — Symptom Text node. Right-click the new Text Miner node, and select **Rename**. Type **Text Miner — CST**, where “CST” stands for “Cleaned Symptom Text”, in the Node Name box. Click **OK**.
 8. Connect the Data Partition node to the Text Miner — CST node.



9. Select the Text Miner — CST node. Set the following properties in the Properties panel for the Text Miner — CST node:

- Click the  button for the **Synonyms** property. Select **Mylib** to display its contents if it is not already selected. Click **Refresh**. Select **Mylib.Vaerextsyn_new** from the Select a SAS Table window. Click **OK**.
 - Set **Terms in a Single Document** to **No**.
- Right-click the Text Miner — CST node, and select **Run**. Click **Yes** in the Confirmation dialog box. Click **OK** in the Run Status dialog box when the node has finished running.
 - Click the  button for the **Interactive** property in the Text Miner — CST node Properties Panel. The Interactive Results window opens.
 - Select the plus sign (+) next to **patient** in the Terms table. Note that the misspellings **patien**, **patietn**, and **patie** are included as child terms.



	TERM	FREQ ▼	# DOCS	KEEP
+	be	11851	5293	<input type="checkbox"/>
-	patient	8081	4492	<input checked="" type="checkbox"/>
	patiens	70	61	
	pts	27	27	
	patien	9	9	
	patient	3547	2065	
	pt	4421	2502	
	patietn	1	1	
	patinet	2	2	
	patie	2	2	
	ppts	1	1	
	ppt	1	1	
+	receive	5207	3698	<input checked="" type="checkbox"/>

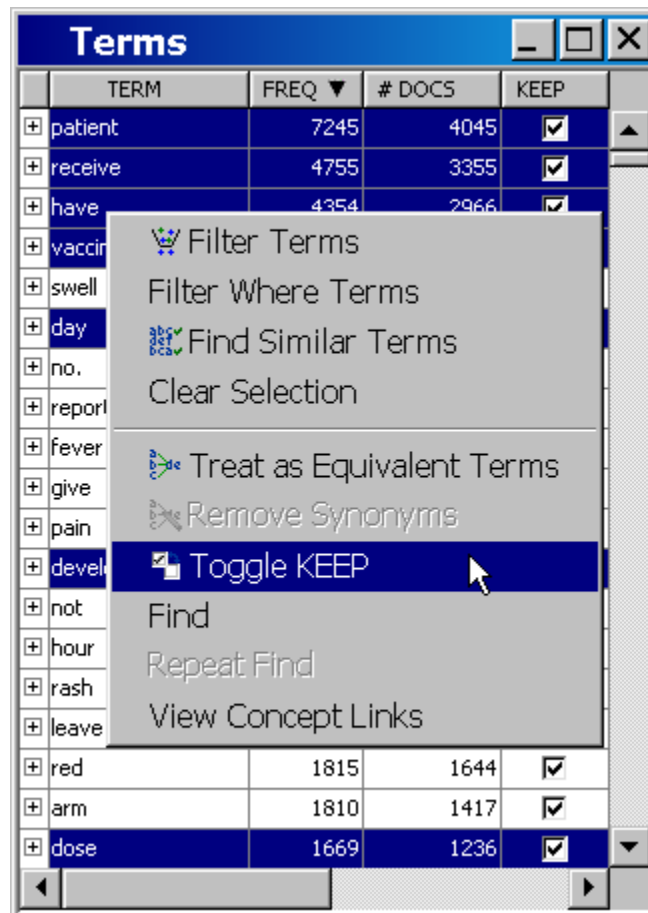
Create a Stop List

A stop list is a simple collection of low-information or extraneous words that you want to remove from the text, which has been saved as a SAS data set.

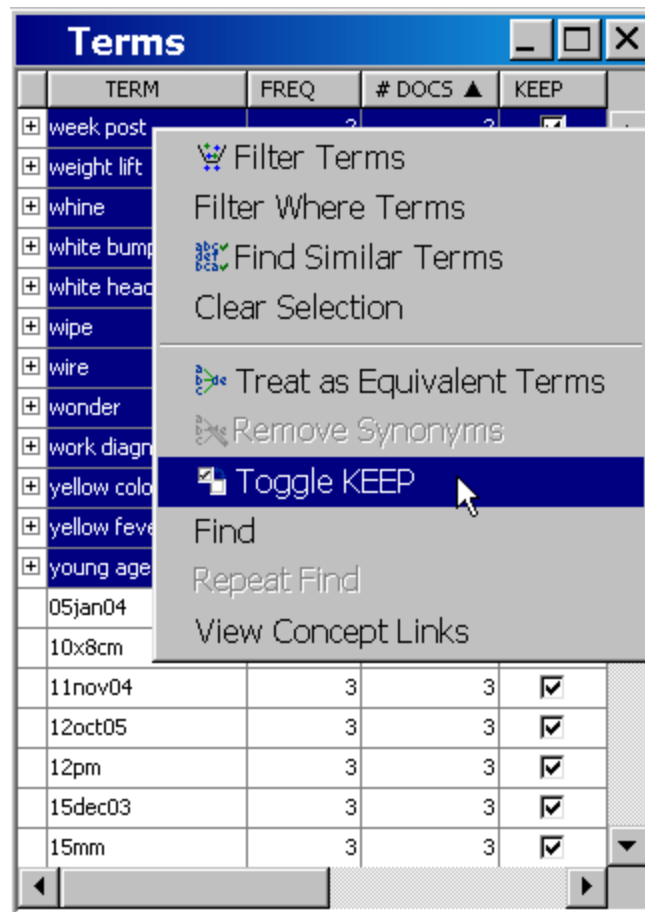
To create a stop list:

- Click the FREQ column heading to sort the Terms table by frequency. Make sure that the Freq label has an arrow that points downward to indicate that the Freq column is sorted in descending order.
- Drop some terms that have no bearing on what the adverse reaction is. Hold down the CTRL key and click on these terms: **patient**, **have**, **receive**, **vaccine**, **day**, **develop**, and **dose**. Right-click to open the menu. Select **Toggle KEEP** to uncheck the **Keep** attribute. This removes the checkmark from the Keep column for each term that you have selected.

There are several more terms you could choose to exclude. Only a few are itemized here to demonstrate the concept and process. If additional terms are dropped from the analysis, then your results will not match those later in this document.



3. Click the # DOCS column heading and ensure that the sort arrow is pointing upward. This sorts the terms by count.
4. Click and drag the mouse to select all terms with counts of 2. Right-click a selected term and select **Toggle KEEP** to drop these terms from the analysis.




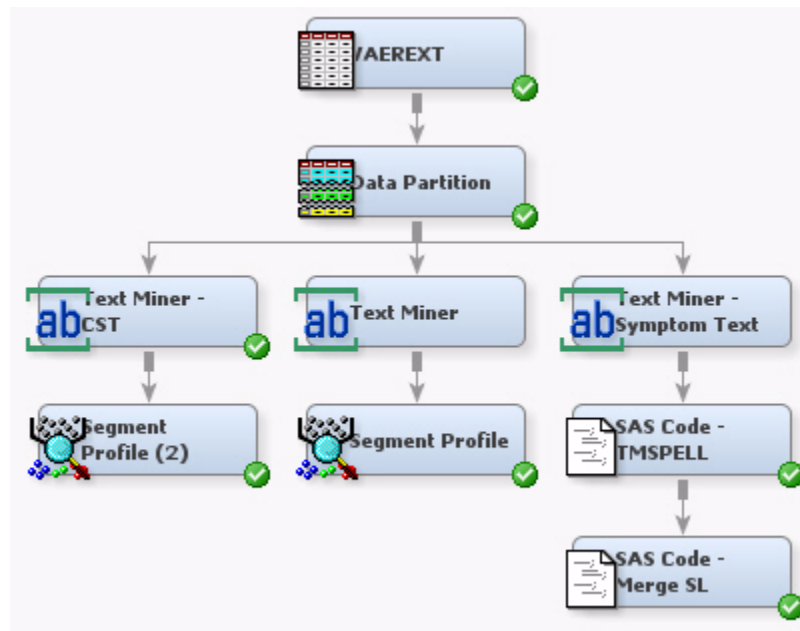
5. Select **File** ⇒ **Save Stop List**.
6. Select the **Mylib** library and type **VAEREXTSTOP** in the **Data Set Name** box.
7. Click **OK**.
8. Close the Text Miner — Interactive window.
9. Note that the **Stop List** property of the Text Miner — CST node is set to **MYLIB.VAEREXTSTOP**.


Parse	
Parse Variable	SYMPTOM_TEXT
Language	ENGLISH
Stop List	MYLIB.VAEREXTSTOP
Start List	
Stem Terms	Yes
Terms in Single Document	No
Punctuation	No
Numbers	No
Different Parts of Speech	No
Ignore Parts of Speech	
Noun Groups	Yes
Synonyms	MYLIB.VAEREXTSYNS_NEW
Find Entities	No
Types of Entities	

Explore Results Improvements

You can redo clustering to explore the improvements to results from cleaning the SYMPTOM_TEXT variable. Complete the following steps:

1. Verify that the Cluster property settings for the Text Miner — CST node are the same as in previous examples.
2. Right-click the Text Miner — CST node, and select **Run**. Click **Yes** in the Confirmation dialog box. Click **OK** in the Run Status dialog box when the node has finished running.
3. Select the  button for the Interactive property in the Text Miner — CST node property panel to open the Interactive Results window. Look at the Clusters table.
4. Compare these results to the first Text Miner node results from “[View Interactive Results](#)” on page 21. Does the clustering seem to have improved with the cleaned SYMPTOM_TEXT data?
5. Close the Interactive Results window.
6. Right-click the Segment Profile node, and select **Copy** from the menu. Right-click on an empty space in the diagram workspace, and select **Paste** from the menu.
7. Connect the Text Miner — CST node to the Segment Profile (2) node.



8. Click the  for the **Variables** property for the Segment Profile (2) node to open the Variables — Prof2 window. Make sure that the PROB variables and the _SVD_ variables have a Use value of **No**.
9. Click **OK** to save the variables settings and close the Variables — Prof2 window.
10. Right-click the Segment Profile (2) node and select **Run**. Click **Yes** in the Confirmation dialog box.

11. Click **Results** in the Run Status dialog box to open the Results window when the node has finished running. Note the significant relationships in the table. Do the relationships appear clearer with the cleaned text than they did with the uncleaned text?
12. Close the Results window.

Chapter 6

Predictive Modeling with Text Variables

About the Tasks That You Will Perform	41
Use the COSTRING Variable to Model	41
Use the SYMPTOM_TEXT Variable to Model	45
Compare the Models	47
Additional Exercises	48

About the Tasks That You Will Perform

Long before text mining, researchers have needed to analyze text. In the field of drug trials, the need was acute enough that coding systems were developed to automatically identify keywords that could be analyzed to understand adverse events. The COSTART coding system was one such attempt. COSTART terms consist of one to three tokens: a symptom, an optional body part, and an optional subpart. One initial task is to find what factors influence whether a reaction becomes serious and how well these factors are captured by the COSTART terms. One way of doing this is to use SAS Text Miner to see how well the COSTART terms predict the seriousness of the adverse event. This chapter explores an example of predictive modeling in SAS Text Miner.

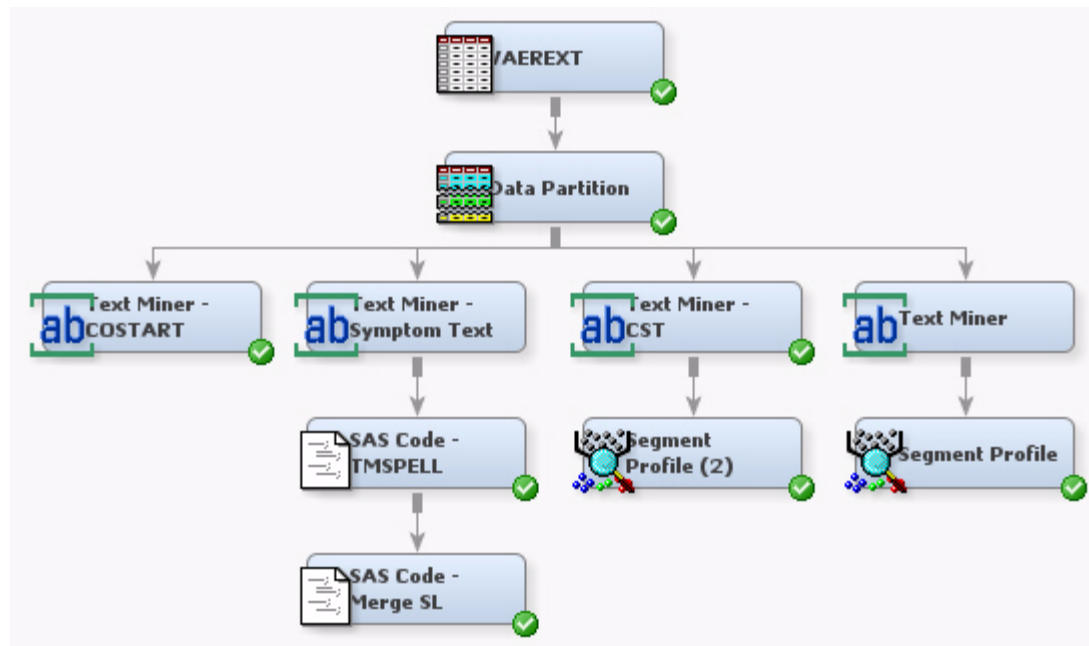
To analyze texts with predictive models, you will perform the following tasks:


1. Use the COSTRING variable and the Decision Tree node to create a model.
2. Use the SYMPTOM_TEXT variable and the Decision Tree node to create a model.
3. Compare the models using the Model Comparison node.

Use the COSTRING Variable to Model

To use the COSTRING variable to create a model:

1. Select the **Text Mining** tab on the toolbar and drag and drop a Text Miner node into the diagram workspace. Connect the Data Partition node to the Text Miner node.
2. Right-click the new Text Miner node and select **Rename**. Type **Text Miner – COSTART** in the Node Name box, and click **OK**.




3. Select the VAEREXT node in the diagram workspace. Click the  button for the **Variables** property in the Properties panel for the VAEREXT node.



Recall that there were two text variables, COSTRING and SYMPTOM_TEXT, from the initial data source. By default, SAS Text Miner will use the longer text variable, SYMPTOM_TEXT. In this chapter, you want to mine the COSTRING variable.

Click **OK** to close the Variables window.

4. Select the Text Miner — COSTART node. Set the following properties in the Properties panel for the Text Miner — COSTART node:

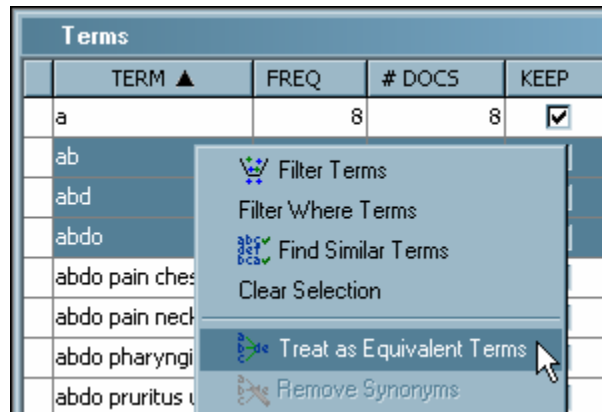
- Click the  button for the **Variables** property. In the Variables window, set the **Use** value for the SYMPTOM_TEXT variable to **No**, the **Use** value for the **costring** variable to **Yes**, and the **Use** value for the **serious** variable to **Yes**. Click **OK** to save your changes.

Name	Use	Role	Level
SYMPTOM_TEXT	No	Text	Nominal
costring	Yes	Text	Nominal
serious	Yes	Target	Binary

- Click the  button to the right of the **Stop List** property. Select the **No data set to be specified** check box in the Select a SAS Table dialog box. This removes the entry for the stop list so that no stop list is used. Click **OK**.
 - Set **Different Parts of Speech** to **No**.
5. Right-click the Text Miner — COSTART node, and select **Run**. Click **Yes** in the Confirmation dialog box. Click **OK** in the Run Status dialog box when the node has finished running.
 6. In the Properties panel, make sure that the **Parse Variable** property of the Text Miner — COSTART Terms node is set to **costring**.
 7. Click the  button for the **Interactive** property to open the Interactive Results window. One problem with COSTART is that it does not always use the same keyword


to describe the same term or equivalent terms. For example, **abdomen** is shown in COSTART as **ab** and as **abdo**. Sometimes there are modifiers that you do not need. You could run the TMSPELL procedure, but because these are abbreviations, the procedure probably will not find all of the correct spellings. You need to manually clean some terms.

8. Sort the terms in the Terms window by clicking on the Term column heading. Select **ab**, **abd**, and **abdo** from the TERM column. Right-click and select **Treat as Equivalent Terms**.



Select **abdo** from the Create Equivalent Terms dialog box. Click **OK**.

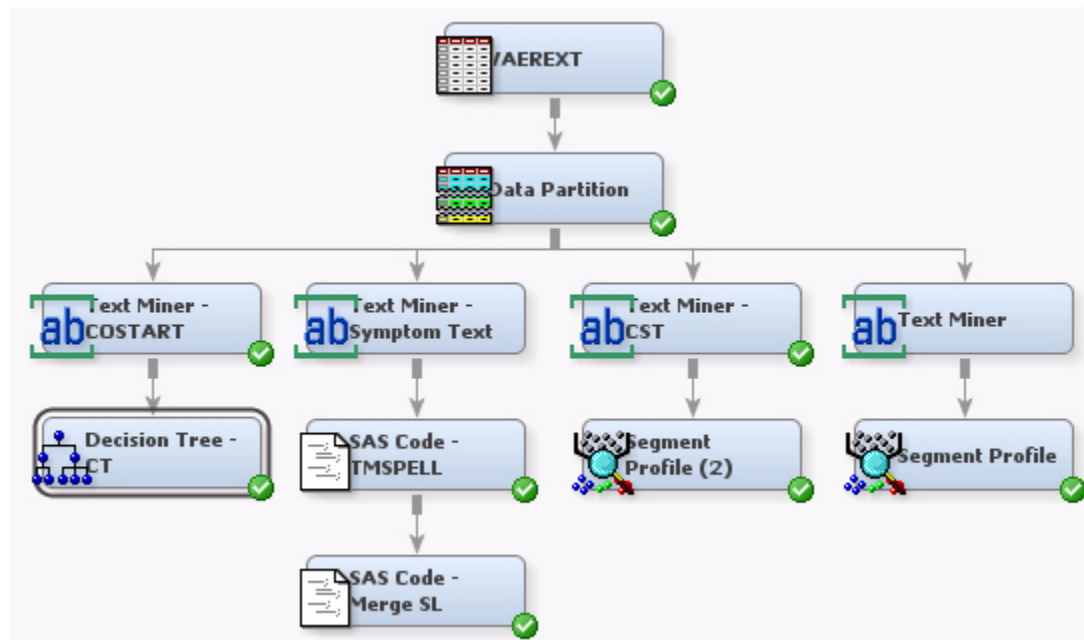
Look through the data set and create synonyms by holding the CTRL or Shift keys and clicking the terms that you consider to be the same. Then, right-click on these selected terms and select **Treat as Equivalent Terms**.

9. Repeat this process as many times as you need. It might be helpful to filter the terms so that you can view the full text of COSTART before combining terms.
10. Select **File** ⇒ **Save Synonyms** from the Interactive Results window menu. Select **Mylib** in the drop-down menu for the library field, and type **COSTARTSYNS** in the Data Set Name field. Click **OK**.
11. Close the Text Miner — Interactive window.
12. Note that the **Synonyms** property in the Properties panel has been set to the new MYLIB.COSTARTSYNS synonym data set.
13. COSTART terms should represent keywords, so you want to create variables for each keyword. Set the following **Transform** properties in the Properties panel:
 - Set **Compute SVD** to **No**.
 - Set **Term Weight** to **Mutual Information**.
 - Set **Roll up Terms** to **Yes**.
 - Set **No. of Rolled-up terms** to **400**.
 - Set **Drop Other Terms** to **Yes**.
14. Right-click the Text Miner — COSTART node, and select **Run**. Click **Yes** in the Confirmation dialog box. Click **OK** in the Run Status dialog box when the node has finished running.
15. Click the  button for the Interactive property to open the Text Miner — Interactive window and view the Terms window.
16. Sort the TERM column until the arrow on the column heading is pointing up.

Note: Terms with a plus (+) sign indicate the synonyms that you have specified. Click the plus (+) sign to expand the child terms underneath the respective parent term.

Terms					
	TERM ▲	FREQ	# DOCS	KEEP	WEIGHT
	a	8	8	<input checked="" type="checkbox"/>	0.8
+	abdo	186	183	<input checked="" type="checkbox"/>	0.421
+	abd	2	2		
+	abdo	163	161		
+	ab	21	21		
	abdo pain chest	3	3	<input type="checkbox"/>	0.682

17. Scroll down until you see terms that do not have a checkmark beneath the Keep column. A separate variable will not be created for these terms. They were not considered significant enough (based on rolling up only 400 variables) to create a separate variable. Recall that you set the **Roll up Terms** property to **Yes** and the **No. of Rolled-up Terms** property to **400**. When you roll up terms, the terms are sorted in descending order of the value of the term weight times the square root of the number of documents. The top 400 highest-ranked terms are then used as variables in the document collection.
18. Close the Text Miner — Interactive window.
19. From the **Model** tab, drag and drop a Decision Tree node into the diagram workspace. Connect the Text Miner — COSTART node to the Decision Tree node. Right-click the Decision Tree node, and select **Rename**. Type **Decision Tree — CT**, where “CT” stands for “COSTART Terms.” Click **OK**.



20. Right-click the Decision Tree — CT node and select **Run**. Click **Yes** in the Confirmation dialog box. Recall that when you created the VAEREXT data set, you set **serious** as the target variable.
21. Click **Results** in the Run Status dialog box after the node has finished running.
22. Select **View** ⇒ **Assessment** ⇒ **Classification Chart: serious** from the Results window menu to view the Classification Chart.

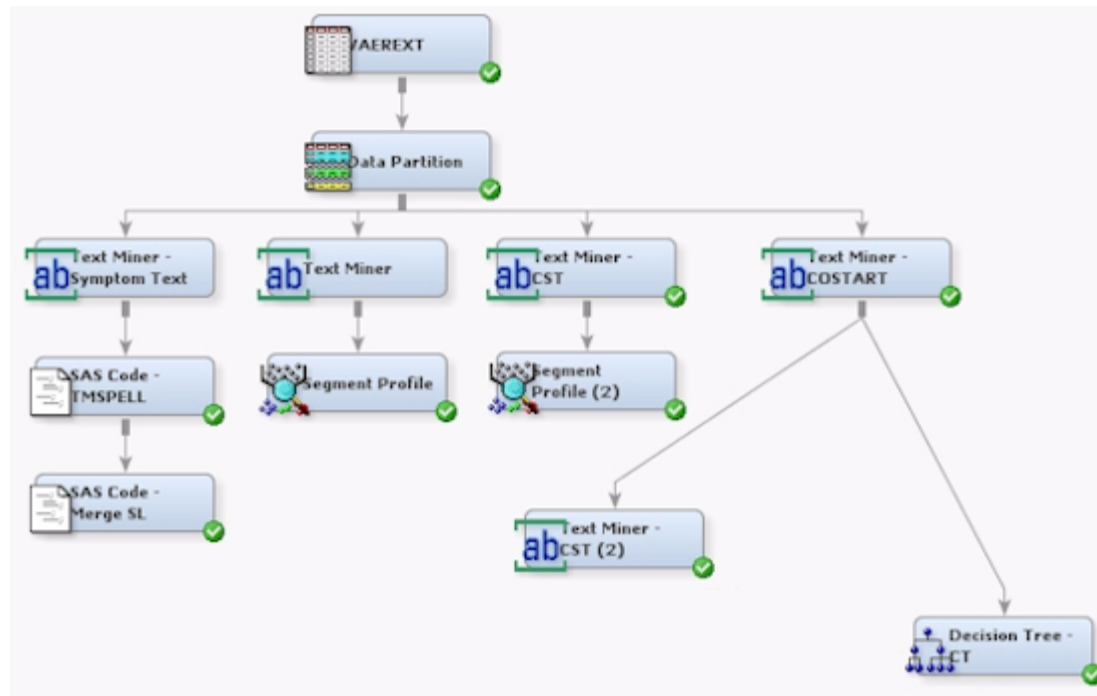
Note: Blue indicates correct classification, and red indicates incorrect classification.

23. Close the Results window.


Use the SYMPTOM_TEXT Variable to Model

To use the SYMPTOM_TEXT variable to create a model, complete the following steps:

1. Right-click the Text Miner — CST node, and select **Copy** from the menu. Right-click an empty space in the diagram workspace and select **Paste**. Connect the Text Miner — COSTART node to the Text Miner — CST (2) node.




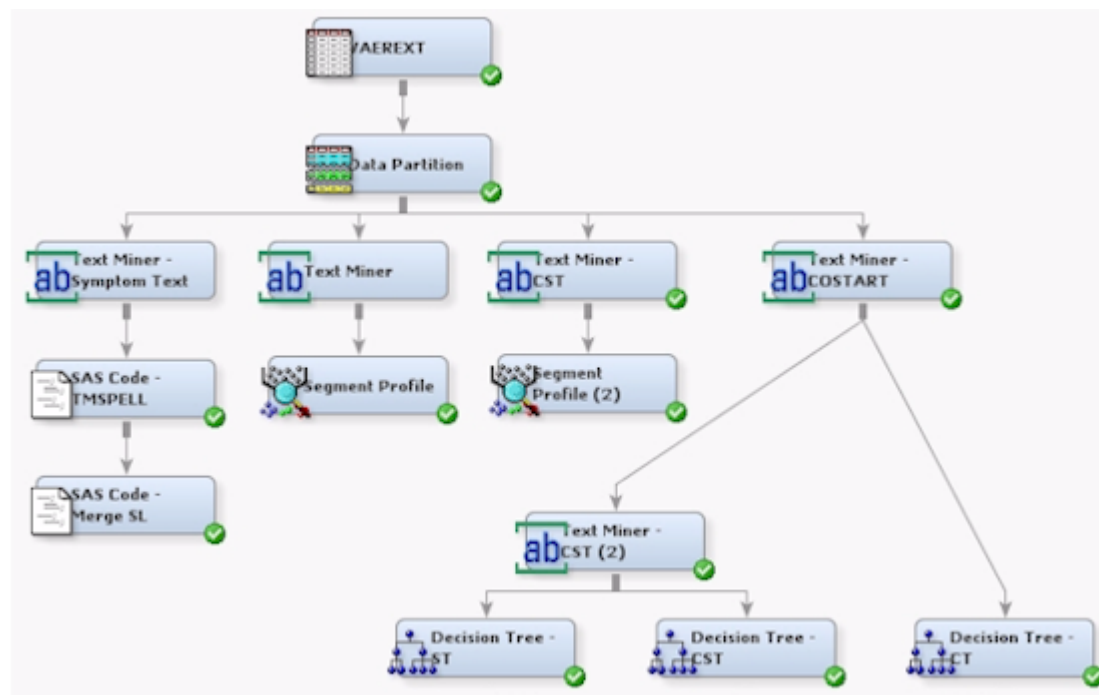
This second Text Miner — CST node will be used to analyze the SYMPTOM_TEXT variable. SYMPTOM_TEXT will be the default parse variable because it is the longest text field in the data set. You need to specify COSTRING as a parse variable as well.

2. Select the second Text Miner — CST (2) node. Click the  button for the **Variables** property in the Properties panel.
3. In the Variables window, ensure that the following values are set:
 - Set the **Use** value of **SYMPTOM_TEXT** to **Yes**.
 - Set the **Use** value of **costring** to **Yes**.
 - Set the **Use** value of **serious** to **Yes**.

Click **OK**.

4. Ensure that the following properties are set in the Properties panel:
 - Set **Compute SVD** to **Yes**.
 - Set **SVD Resolution** to **Low**.
 - Set **Term Weight** to **Mutual Information**.

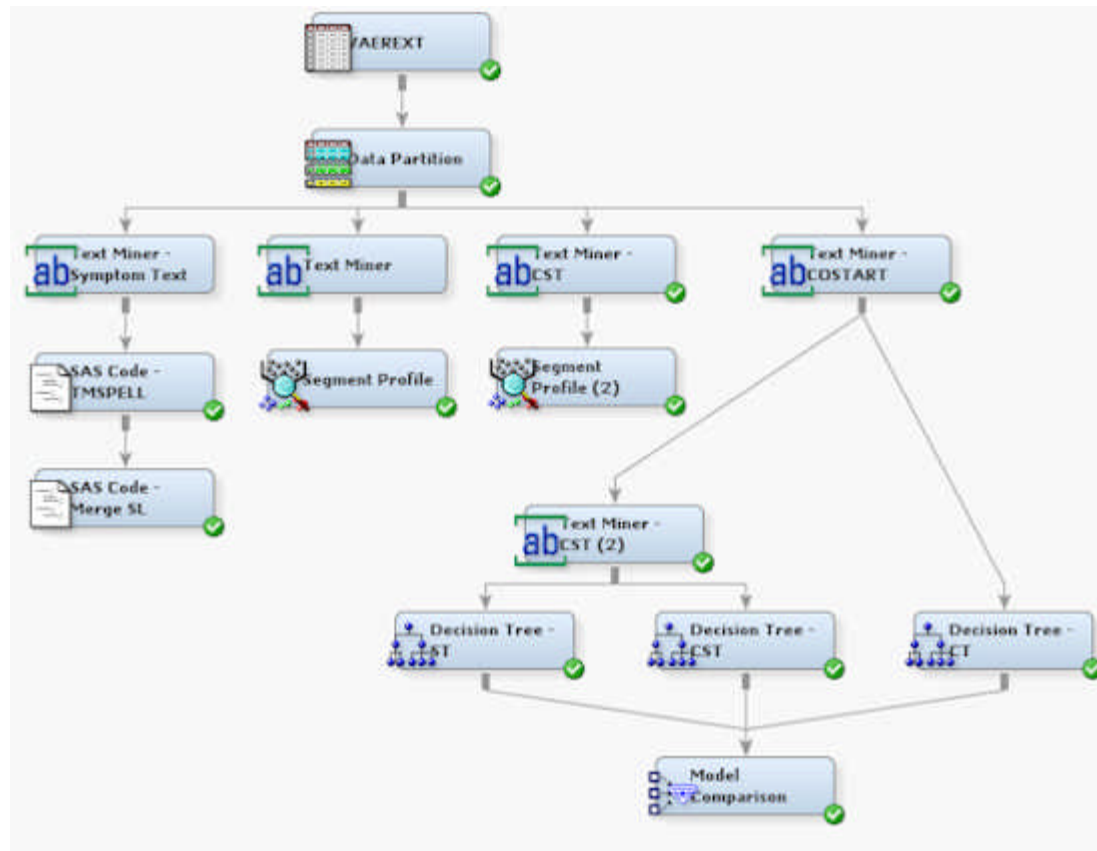
5. Right-click the new Text Miner node, and select **Run**. Click **Yes** in the Confirmation dialog box. Click **OK** in the Run Status dialog box.
6. From the **Model** tab, drag and drop a Decision Tree node into the diagram workspace. Connect the Text Miner — CST (2) node to the Decision Tree node. You will use the decision tree to see whether text mining the original text can do a better job of predicting serious events than just mining the COSTART terms.
7. Right-click the new Decision Tree node and select **Rename**. Type **Decision Tree — ST**, where “ST” stands for “Symptom Text,” in the Node Name text box. Click **OK**.
8. Click the  button for the **Variables** property in the Decision Tree — ST properties panel. The Variable window opens.
9. Click and scroll to select all of the **_ROLL_** variables, and then set the **_ROLL_ Use** values to **No**.
10. Click **OK** to save your changes.
11. Right-click the Decision Tree — ST node, and select **Run**. Click **Yes** in the Confirmation dialog box. Click **OK** in the Run Status dialog box when the node has finished running.
12. From the **Model** tab, drag and drop a Decision Tree node and connect it to the Text Miner — CST (2) node.
13. Right-click the new Decision Tree node, and select **Rename**. Type **Decision Tree — CST**, where “CST” stands for “COSTART and Symptom Text,” in the Node Name text box. Click **OK**. This node will let you see how well you can predict serious events with all the information available to you. Use the default settings for the node.



Compare the Models

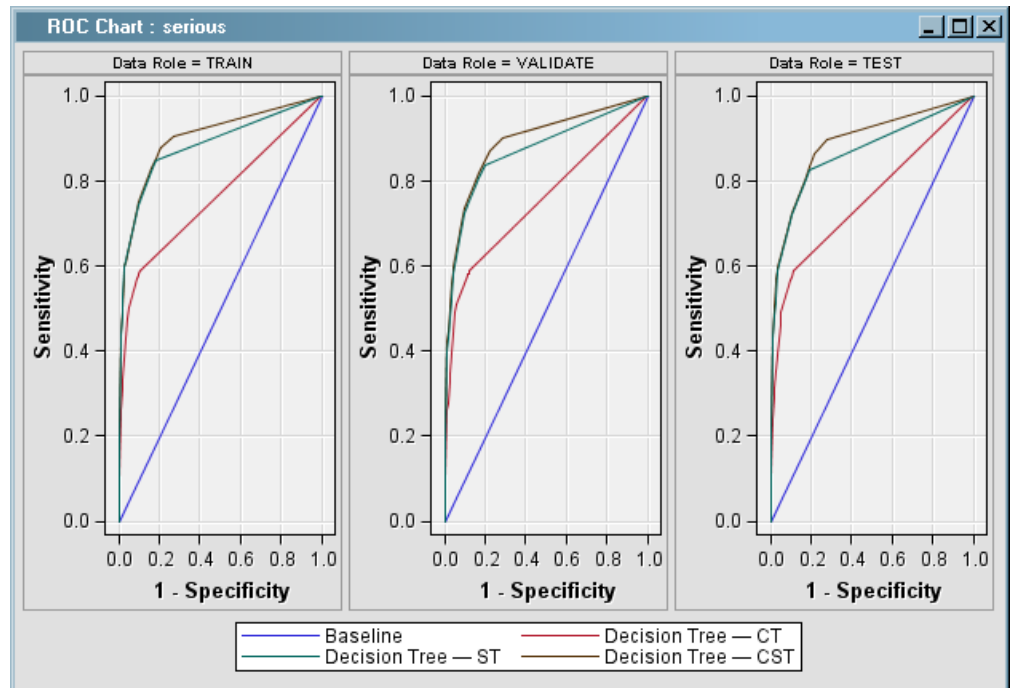
To compare the models:

1. From the **Assess** tab, drag and drop a Model Comparison node into the diagram workspace. Connect all three Decision Tree nodes to it. The Model Comparison node enables you to compare the performance of the three different models. Your diagram should look something like the following:



2. Right-click the Model Comparison node, and select **Run**. Click **Yes** in the Confirmation dialog box. Click **Results** in the Run Status dialog box when the Model Comparison node has finished running.
3. Maximize the ROC Chart.

The greater the area under the curve, the better the model. In the following image, the red line shows the results of the model using COSTART terms, the green line shows the results of the SYMPTOM_TEXT terms, and the brown line shows the results of the combined COSTART and SYMPTOM_TEXT terms. The worst model uses only the COSTART terms, while the best model uses a combination of COSTART and SYMPTOM_TEXT. Apparently, text mining can add information not contained in the COSTART terms. The text mining model provides better results than the keyword-based model. Combining the models offers the best results.



4. Select **View** ⇒ **Assessment** ⇒ **Classification Chart** from the menu at the top of the Results window to view the Classification Chart.

Note: Blue indicates correct classification, and red indicates incorrect classification.

5. Close the Results window. It would be useful to see which variables in the combined model are most important for predicting serious events.
6. Right-click on the Decision Tree — CST node and select **Results** to view the results of the combined Decision Tree models.
7. Click the Output window to maximize it. Scroll through the output to the **Variable Importance** results.

Note: The SVD terms are more important than the individual terms in predicting a serious adverse event.

8. Minimize the Output window, and then maximize the window that contains the decision tree. Browse the decision tree results.

Additional Exercises

You have looked at predicting the seriousness of adverse events. To explore additional exercises, complete the following steps:

1. You might want to look at the types of adverse events that occur. Try the following:
 - See if you can use the COSTART analysis to predict the clusters that you obtained from analyzing the SYMPTOM_TEXT variable. You can do this with the Cluster node. To combine variables together, you might want to try a Decision Tree node.
 - The original data contains other variables, such as medications and lab tests. You know that the type of adverse event is affected by drug interactions. Using the original data, see if you can text mine the medications field to roll up variables for

the medications that patients are currently taking. Then use these variables to try to predict the clusters that you obtained for the SYMPTOM_TEXT variable.

2. If you have access to a MedDRA program, run the text through that and perform the same tasks with the MedDRA results that you did with the COSTART terms in this book.

Now that you have completed this example, you are ready to learn about the three new nodes in Text Miner 4.2: the Text Parsing, Text Filter, and Text Topic nodes.

Chapter 7

Using the Text Parsing Node

About the New Nodes	51
About the Text Parsing Node	51
About the Tasks That You Will Perform	52
Creating a Project	52
Creating a Data Source	52
Creating a Diagram	53
Using the Text Parsing Node	53

About the New Nodes

In Text Miner 4.2, there are three new text mining nodes. These nodes are Text Parsing, Text Filter, and Text Topic. They were created to provide more options to the way that you perform text mining. These nodes are designed to stand alone in the text mining process. The following example provides an overview of these three nodes using a sample data set. For more information about these nodes, see the Text Miner help documentation.

About the Text Parsing Node

The Text Parsing node enables you to use advanced natural language processing software to represent each document as a collection of terms. A term can represent a single word, either with or without its part of speech; multi-word phrases; and custom or built-in entities. Additionally, you can choose to have certain terms represented by their synonyms or stems. The Text Parsing node gives you control over exactly what types of terms to include in your analysis. This node is the first step of any text mining analysis. It can be used with diverse textual data such as e-mail messages, news articles, Web pages, research papers, and surveys.

About the Tasks That You Will Perform

The Abstract data table contains title and abstract data from a series of SAS Users Group International (now called SAS Global Forums) meetings from 1998 through 2001. In this example you will parse the data with particular settings and view the reports that are generated for the terms that are identified. This chapter will explain how to perform the following steps:

1. Create a new project where you will store all your work.
2. Define the ABSTRACT data as a SAS Text Miner data source.
3. Create a new process flow diagram.
4. Run the Text Parsing node.

Creating 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 window opens.
3. Click **New Project**. The specify Project Name and Server Directory page opens.
4. Type a name for the project, such as **Abstract Data Patterns** in the Project Name dialog box.
5. In the SAS Server Directory dialog box, type the path to the location on the server where you want to store data for your project. Alternatively, browse to a folder to use for your project.
6. Click **Next**. The Register the Project page opens.
7. Click **Next**. The New Project Information page opens.
8. Click **Finish** to create your project

Creating a Data Source

To create a data source:

1. Right-click the Data Sources folder in the Project panel and select **Create Data Source** to open the Data Source wizard.



2. Select **SAS Table** in the Source drop-down menu of the Metadata Source window.
3. Click **Next**. The Select a SAS Table window opens.
4. Click **Browse**.
5. Click the SAS library named **Sampsio**. The Sampsio library folder contents are displayed on the Select a SAS Table dialog box.
6. Select the **Abstract** table, and then click **OK**. The two-level name **SAMPSIO.ABSTRACT** is displayed in the Table box of the Select a SAS Table page. Click **Next**.
7. The Table Information page opens. The Table Properties panel displays metadata for you to review. Click **Next**.
8. The Metadata Advisor Options page opens. Click **Next**.
9. The Column Metadata page opens.
10. The default variables for both **TEXT** and **TITLE** should be set to **Text** by default. If this is not the case, then you need to change their roles to **Text** using the drop-down list. Click **Next**.
11. The Create Sample page opens. Click **Next**.
12. The Data Source Attributes page opens. Click **Next**.
13. The Summary page opens. Click **Finish** and the **ABSTRACT** table is added to the Data Sources folder in the Project panel.

Creating a Diagram

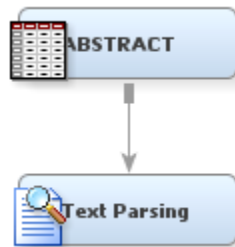
To create a diagram, complete the following steps:

1. Right-click the Diagram folder in the Project Panel and select **Create Diagram**. The Create New Diagram dialog box opens.
2. Type **Abstract Data** in the Diagram Name box.
3. Click **OK**. The empty Abstract Data diagram opens in the diagram workspace.
4. Drag and drop the **ABSTRACT** data source from the Data Sources list into the diagram workspace.

Using the Text Parsing Node

1. Select the **Text Mining** tab in the Enterprise Miner Toolbar, and drag a Text Parsing node into the diagram workspace.

2. Connect the Abstract data source to the Text Parsing node.



3. Select the Text Parsing node to highlight it and then select the ellipsis for the **Ignore Parts of Speech** property. Select all parts of speech except for **Noun**, and click **OK**.

TIP To do this step, hold the Control key and click every entry but **Noun** so that they become highlighted.

4. Right-click the Text Parsing node and select **Run**. Click **Yes** in the Confirmation dialog box.
5. When the node has finished running, select **Results** in the Run Status dialog box.
6. In the Results window, find the Terms table and click anywhere inside it to make it active. The Terms table presents terms that have been parsed by the Text Parsing node, the term's role, the number of documents that it appears in, whether the term was kept or rejected, and other attributes. A term might be dropped from analysis if it appears on a stop list or was ignored for another reason.

Term	Role	Attribute	Freq	# Docs	Keep
+ datum	Noun	Alpha	2746	785Y	
+ system	Noun	Alpha	1082	549Y	
software	Noun	Alpha	857	485Y	
+ paper	Noun	Alpha	524	422Y	
+ application	Noun	Alpha	835	389Y	
+ user	Noun	Alpha	642	379Y	
information	Noun	Alpha	537	297Y	
+ use	Noun	Alpha	345	262N	
+ tool	Noun	Alpha	299	216Y	
+ example	Noun	Alpha	247	208N	
+ analysis	Noun	Alpha	377	205Y	
+ presentati	Noun	Alpha	250	202Y	

7. Close the Results window.

Chapter 8

Using the Text Filter Node

About the Text Filter Node	55
About the Tasks That You Will Perform	55
Using The Text Filter Node	55

About the Text Filter Node

The Text Filter node enables you to focus on the terms and documents that are most likely to enhance your model. For most of your text mining projects, you will want to follow the Text Parsing node with the Text Filter node. This way, you can eliminate extraneous information caused by the presence of noise terms and other terms that are not pertinent to your analysis. If your model would be improved by focusing on a subset of the collection, then the Text Filter node can remove documents that do not fit your criteria. The end result of the Text Filter node is a compact, yet information rich, representation of your collection. For example, you can put a Text Filter node after a Text Topic node and filter for documents that contain specific topics.

About the Tasks That You Will Perform

In this section you will filter the parsed data to eliminate low frequency words and other terms that contain little information value. To accomplish this task, you will remove any misspelled words and omit terms appearing in fewer than ten documents. Also, you will interactively remove frequently occurring words that contain little information value.

Using The Text Filter Node

Note: This example assumes you completed the example in [“Using the Text Parsing Node” on page 53](#). It builds off the process flow diagram created there.

Now that you have parsed the terms in the text of the abstracts, you want to filter out the terms that have little to no information value. This will create more relevant topics in the next step.

1. Select the **Text Mining** tab in the Enterprise Miner Toolbar, and drag a Text Filter node into the diagram workspace.
2. Connect the Text Parsing node to the Text Filter node.



3. Right-click the Text Parsing node, and select **Results**. Now find the Terms table and click anywhere inside the Terms table to make it the active window. Scroll through the list of terms and note that the majority of terms occur in fewer than 10 documents.
4. Click the Text Filter node in the diagram workspace and notice the **Minimum Number of Documents** property in the **Term Filters** section of the Train properties. Change this property from the default value to **10**. This ensures that only terms that occur in at least ten documents are included in your analysis.
5. To enable spell checking, set the **Check Spelling** option in the **Spelling** section of the Train properties to **Yes**.

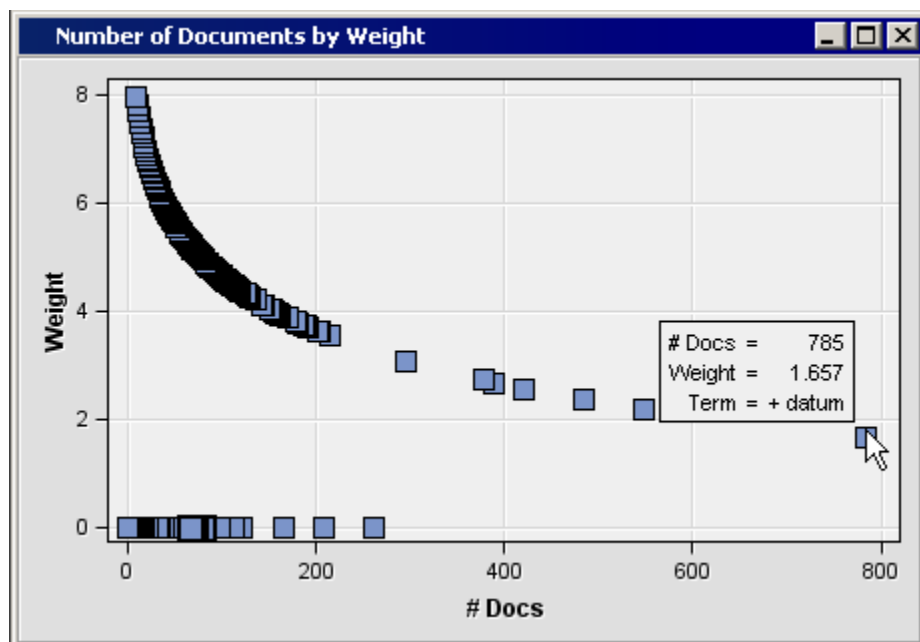
Because you are dealing with professionally written documents, you might think that there is no reason to perform this operation. However, this spell checking might suggest terms that should be treated as synonyms. The algorithm used to find misspellings frequently identifies terms that are slight variants of one another in spelling and meaning. Checking for spelling should not remove a large number of terms from this data set but might help when you deal with a different collection of documents. If you are dealing with informal document sets such as e-mails or customer comments, then spell checking will prove even more beneficial.

6. To run the Text Filter Node, right-click on the node and select **Run**. When you have successfully run the Text Filter node, click the **Results** button in the Run Status window.
7. Find the Terms window within the Results window and click anywhere inside the Terms window to make it the active window. Select the **#Docs** column to sort terms by document frequency. Scroll down and notice that terms that appear in fewer than 10 documents are dropped by the Text Filter node. Terms in 10 or more documents might have been dropped for other reasons.

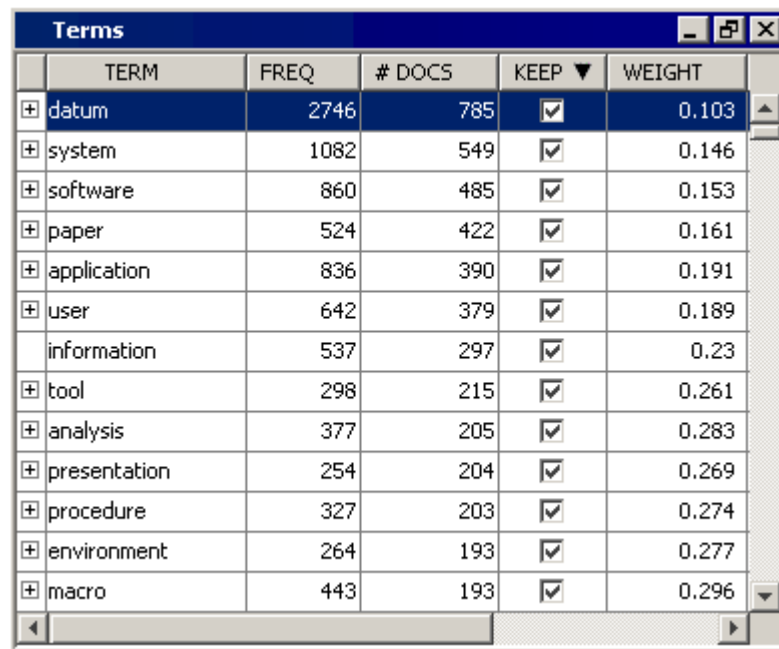
Term	Role	Attribute	Status	# Docs ▲	Weight
+ datum 30...	Noun Group	Alpha	Drop	9	0.000
+ datum tab...	Noun Group	Alpha	Drop	9	0.000
+ datum vis...	Noun Group	Alpha	Drop	9	0.000
+ doe	Noun	Alpha	Drop	9	0.000
+ external d...	Noun Group	Alpha	Drop	9	0.000
+ lot	Noun	Alpha	Drop	9	0.000
popularity	Noun	Alpha	Keep	10	0.677
+ offering	Noun	Alpha	Keep	10	0.677
background	Noun	Alpha	Keep	10	0.677
+ conversion	Noun	Alpha	Keep	10	0.714
legacy	Noun	Alpha	Keep	10	0.681
+ plan	Noun	Alpha	Keep	10	0.699
+ domain	Noun	Alpha	Keep	10	0.681
find	Noun	Alpha	Keep	10	0.681

- Now find and maximize the Number of Documents by Weight window. This window shows a scatter plot of the terms with the number of documents on the x-axis and term weight on the y-axis. In the image below, notice that there is a row of points that all have a weight of zero. These points represent the terms that were dropped from analysis. When you hold the mouse over a point on the graph, a tooltip appears that notes the term represented, the number of documents that it appears in, and the assigned weight value.

Notice that the point representing the term **datum** appears separated from the rest of the points. Because the term **datum** appears significantly more often than any other term, you will drop this term from analysis in the following steps. There are six more terms that appear separated from the rest of the terms. These terms are **system**, **software**, **paper**, **application**, **user**, and **information**. You might choose to drop these terms from your analysis; but if you do, then you will get different results from those presented in the next chapter.



9. Close the Results window.
10. Now find the **Results** section of the Train properties and click the ellipsis button next to the **Spell-Checking Results** property. The EMWS.TextFilter_spellDS window lists the misspelled terms, and their assigned parent term, that were detected by the spell checker. Other information, such as the term role, term number, and number of documents, is also displayed in this window.
11. Select the Text Filter node, and then click the ellipsis for the **Filter Viewer** property to open the Interactive Filter Viewer. In the Interactive Filter Viewer, you can refine the parsed and filtered data that exists after the Text Filter node has run. The refinement is achieved by filtering documents based on the results of a search expression or modifying the keep or synonym status of a term. The following image shows the Terms table of the Interactive Filter Viewer.



	TERM	FREQ	# DOCS	KEEP ▼	WEIGHT
+	datum	2746	785	<input checked="" type="checkbox"/>	0.103
+	system	1082	549	<input checked="" type="checkbox"/>	0.146
+	software	860	485	<input checked="" type="checkbox"/>	0.153
+	paper	524	422	<input checked="" type="checkbox"/>	0.161
+	application	836	390	<input checked="" type="checkbox"/>	0.191
+	user	642	379	<input checked="" type="checkbox"/>	0.189
	information	537	297	<input checked="" type="checkbox"/>	0.23
+	tool	298	215	<input checked="" type="checkbox"/>	0.261
+	analysis	377	205	<input checked="" type="checkbox"/>	0.283
+	presentation	254	204	<input checked="" type="checkbox"/>	0.269
+	procedure	327	203	<input checked="" type="checkbox"/>	0.274
+	environment	264	193	<input checked="" type="checkbox"/>	0.277
+	macro	443	193	<input checked="" type="checkbox"/>	0.296

12. For this example, you are going to exclude the single term **datum** because it occurs more than twice as often as the next frequent word. This will give you more relevant topics when you run the [Text Topic node on page 61](#).

To exclude the term **datum** from the analysis, click the check box in the **KEEP** column of the Terms window. The check box should be unchecked as in the diagram below.

Terms					
	TERM	FREQ	# DOCS	KEEP ▼	WEIGHT
+	datum	2746	785	<input type="checkbox"/>	0.103
+	system	1082	549	<input checked="" type="checkbox"/>	0.146
+	software	860	485	<input checked="" type="checkbox"/>	0.153
+	paper	524	422	<input checked="" type="checkbox"/>	0.161
+	application	836	390	<input checked="" type="checkbox"/>	0.191
+	user	642	379	<input checked="" type="checkbox"/>	0.189
	information	537	297	<input checked="" type="checkbox"/>	0.23
+	tool	298	215	<input checked="" type="checkbox"/>	0.261
+	analysis	377	205	<input checked="" type="checkbox"/>	0.283
+	presentation	254	204	<input checked="" type="checkbox"/>	0.269
+	procedure	327	203	<input checked="" type="checkbox"/>	0.274
+	environment	264	193	<input checked="" type="checkbox"/>	0.277
+	macro	443	193	<input checked="" type="checkbox"/>	0.296

13. Close the Interactive Filter Viewer. When you do this, the Save window will ask you to save your changes. Make sure that you click **Yes** or you will have to redo steps 10 and 11.

Chapter 9

Using the Text Topic Node

About the Text Topic Node	61
About the Tasks That You Will Perform	61
Using the Text Topic Node	61

About the Text Topic Node

The Text Topic node enables you to define, discover, and modify sets of topics contained in your collection. A topic is a collection of terms that are strongly associated with a subset of your collection. Each document can contain zero, one, or many topics. Also, terms that are used to describe and define one topic can be used in other topics. There are no rules to determine how many topics you should create. The number of topics that you create is your preference and will vary from problem to problem.

The Text Topic node must be preceded by a Text Parsing node. If it is also preceded by a Text Filter node, then it will use the term weights set in that node. Otherwise, the Text Topic node will create term weights with the Log frequency weighting and Entropy term weighting settings.

About the Tasks That You Will Perform

In this section you will use the Text Topic node to create a set of topics. These topics will include a user-created topic and topics that are created by the Text Topic node. Then, you will view all of these topics together with the documents in which they are contained.

Using the Text Topic Node

Note: This example assumes you completed the example in [“Using The Text Filter Node” on page 55](#), and builds off the process flow diagram created there.

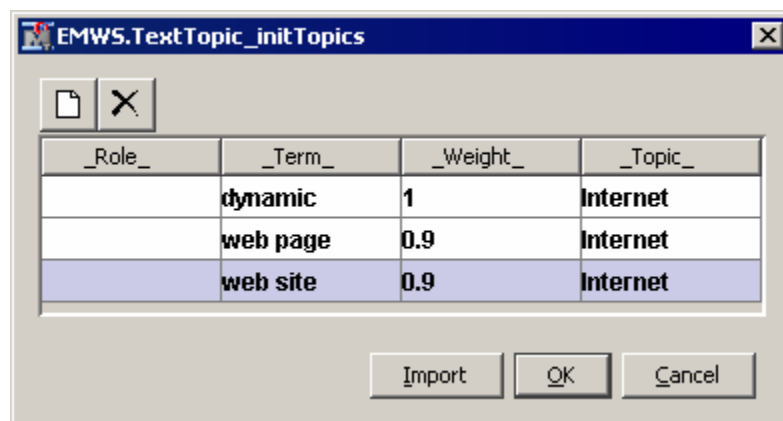
1. Select the **Text Mining** tab in the Enterprise Miner Toolbar, and drag a Text Topic node into the diagram workspace.
2. Connect the Text Filter node to the Text Topic node.



3. Before running the Text Topic node, you are going to create a topic that identifies abstracts that discuss dynamic Web sites. To do this, click on the ellipsis button next to the User Topics property of the Text Topic node. This opens the User Topics window where you can create your own topics.

In the User Topics window, there are four columns, labeled **_Role_**, **_Term_**, **_Weight_**, and **_Topic_**. For this example, you can leave the **_Role_** column empty. To add a row, click the **Add** button at the top of the window. To delete a term, select that row and click the **Delete** button at the top of the window.

To create the topic about dynamic Web sites, enter the three terms, with their corresponding weights and topics, as shown in the image below. The weight of a term is relative and can be any value between 0 and 1. More important terms should have a higher weighting than less important terms.



After you have entered the three terms shown above, click **OK** to save your changes. You are now ready to run the Text Topic node.

4. Right-click the Text Topic node and select **Run** to run the Text Topic node with all other settings at their default values. Click **Yes** in the Confirmation dialog box. When the node finishes running, select **Results** in the Run Status dialog box.
5. From the Results window, expand the Number of Terms by Topic chart. The default settings created twenty-five multi-term topics in addition to the one you defined above. Close the Results window.
6. Select the Text Topic node, and then click the ellipsis button for the **Topic Viewer** property to open the Interactive Topic Viewer window. At the top of the **Topics** list

should be the topic **Internet**, which you created. Notice that the **Category** of this topic is given as **User** while the rest of the topics are given as **Multiple**.

You can view all of the documents in your topic by right-clicking anywhere in the first row and selecting **Select Current Topic**, if it is not already selected. The middle pane shows you that only the terms *dynamic*, *web page*, and *web site* are in this topic. It also shows how many documents each term is in and how frequently each term appeared. The bottom pane contains the observations from the Abstract data set that belong to your topic. You can read each of these by right-clicking anywhere in the bottom pane and selecting **Toggle Show Full Text** from the menu. Close the Interactive Topic Viewer.

7. One problem that you might have noticed is that many of the topics appear to be closely related because they have many terms in common. This suggests that some topics can be merged together. In the **Topics** list of the Interactive Filter Viewer, there are four topics (topics 2, 3, 5, and 10) that contain both **user** and **application** as descriptive terms. These topics can be merged to form one, user-created topic.

To merge the topics, double-click the first topic you want to rename to make it the active topic. You can rename the topic by deleting all of the text in the **Topic** field and replacing it with **User Applications**. Repeat this process for the other three topics that contain both **user** and **application**. When you rename a topic, the **Category** of the topic changes to **User**. Close the Interactive Filter Viewer and save the changes you made.

8. Right-click the Text Topic node and select **Run** to rerun the Text Topic node. Click **Yes** in the Confirmation dialog box and click **OK** in the Run Status dialog box when the node is finished running.
9. Now, click the ellipsis next to the **Topic Viewer** property to see the new topics that have been created. As you can see, there are now two user-defined topics, **Internet** and **User Applications**. However, the Text Topic node still created 25 topics. If you are not satisfied with these topics, you can continue to merge multiple topics together until the topics are distinct enough for your needs.

Chapter 10

Tips for Text Mining

Processing a Large Collection of Documents	65
Dealing with Long Documents	65
Processing Documents from an Unsupported Language or Encoding	66

Processing a Large Collection of Documents

Using the text mining 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.
- When using the Text Miner node, set some of the Parse properties to **No**, such as **Find Entities**, **Noun Groups**, and **Terms in Single Document**.
- When using the Text Parsing node, set some of the Detect properties to **No**, such as **Find Entities** and **Noun Groups**.
- In the Text Miner node, 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.
- Use the Ignore properties of the Text Parsing node to limit parsing to high information words. You can do this by ignoring all parts of speech other than nouns, proper nouns, noun groups, and verbs.
- You can also use the Parse properties in the Text Miner node to ignore 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:
 - **Stem Terms**
 - **Different Parts of Speech**
 - **Noun Groups**
 - **Find Entities**
3. Run the Text Miner node.

Many of the terms might have characters that do not display correctly, but the Interactive Results window should function and you should be able to create stop lists, start lists, and synonym lists.

Chapter 11

Next Steps: A Quick Look at Additional Features

The %TMFILTER Macro	67
---------------------------	----

The %TMFILTER Macro

The %TMFILTER macro is a SAS macro that enables you to convert files into SAS data sets. You can use the macro to perform the following tasks:

- Read documents contained in many different formats (such as PDF and Microsoft Word), convert the files to HTML, and create a corresponding SAS data set that can be used as input for the Text Miner node.
- Retrieve Web pages starting from a specified URL and create a SAS data set that can be used as input for the Text Miner node.
- With the language options, separate your collection by language.

Note: The %TMFILTER macro runs only on Windows operating environments.

See **Using the %TMFILTER Macro** in the Text Miner node documentation in SAS Text Miner for more information.

Appendix 1

Vaccine Adverse Event Reporting System Data Preprocessing

The VAERS data for 2002-2006 is read into a SAS data set using a SAS program called `Vaers_Import.sas`. This SAS program creates a table called `VAERALL`. `Vaers_Import.sas` is included in the *Getting Started with Text Miner 4.2* zip file.

```
proc import out= dmtm9.vaers2006
  datafile= "d:\vaers files\2006vaersdata.csv"
  dbms=csv replace;
  getnames=yes;
  datarow=2;
run;
proc import out= dmtm9.vaers2005
  datafile= "d:\vaers files\2005vaersdata.csv"
  dbms=csv replace;
  getnames=yes;
  datarow=2;
run;
proc import out= dmtm9.vaers2004
  datafile= "d:\vaers files\2004vaersdata.csv"
  dbms=csv replace;
  getnames=yes;
  datarow=2;
run;
proc import out= dmtm9.vaers2003
  datafile= "d:\vaers files\2003vaersdata.csv"
  dbms=csv replace;
  getnames=yes;
  datarow=2;
run;
proc import out= dmtm9.vaers2002
  datafile= "d:\vaers files\2002vaersdata.csv"
  dbms=csv replace;
  getnames=yes;
  datarow=2;
run;
proc import out= dmtm9.vaervax2006
  datafile= "d:\vaers files\2006vaersvax.csv"
  dbms=csv replace;
  getnames=yes;
  datarow=2;
run;
proc import out= dmtm9.vaervax2005
  datafile= "d:\vaers files\2005vaersvax.csv"
  dbms=csv replace;
  getnames=yes;
```

```

        datarow=2;
run;
proc import out= dmtm9.vaervax2004
    datafile= "d:\vaers files\2004vaersvax.csv"
    dbms=csv replace;
    getnames=yes;
    datarow=2;
run;
proc import out= dmtm9.vaervax2003
    datafile= "d:\vaers files\2003vaersvax.csv"
    dbms=csv replace;
    getnames=yes;
    datarow=2;
run;
proc import out= dmtm9.vaervax2002
    datafile= "d:\vaers files\2002vaersvax.csv"
    dbms=csv replace;
    getnames=yes;
    datarow=2;
run;
data dmtm9.vaerall;
    set dmtm9.vaers2002(drop=datedied hospdays)
        dmtm9.vaers2003(drop=datedied hospdays)
        dmtm9.vaers2004(drop=datedied hospdays)
        dmtm9.vaers2005(drop=datedied hospdays)
        dmtm9.vaers2006(drop=datedied hospdays);
run;
data dmtm9.vaervaxall;
    set dmtm9.vaervax2002
        dmtm9.vaervax2003
        dmtm9.vaervax2004
        dmtm9.vaervax2005
        dmtm9.vaervax2006;
run;

```

The data is then further processed to come up with the extract used in the example:

- The separate COSTART terms are appended into a single COSTRING field for each adverse event.
- Additional indicator variables are created for each of the vaccinations received. In the case of DTP, both the Pertussis and Diphtheria/Tetanus variables would be flagged.

The SAS code Vaerssetup.sas used to generate the resulting table, VAEREXT, is in the Getting Started with Text Miner 4.2 zip file.

```

libname dmtm9 'd:\emdata\dmtm9';
/*----- TJW Modification: within DATA step -----*/
%macro FixJunk(TextVar=);
    &TextVar = tranwrd(&TextVar,'n_t ', " not ");
    &TextVar = tranwrd(&TextVar,'N_T ', " NOT ");
    &TextVar = tranwrd(&TextVar,"n't ", " not ");
    &TextVar = tranwrd(&TextVar,"N'T ", " NOT ");
    &TextVar = tranwrd(&TextVar,',' , " ");
    &TextVar = tranwrd(&TextVar,') ', " ) ");
    &TextVar = tranwrd(&TextVar,'( ', " ( ");
    &TextVar = tranwrd(&TextVar,'] ', " ] ");
    &TextVar = tranwrd(&TextVar,[' ', " [ ");

```

```

&TextVar = tranwrd(&TextVar,'}', " } ");
&TextVar = tranwrd(&TextVar,'{', " { ");
&TextVar = tranwrd(&TextVar,'*', " * ");
&TextVar = tranwrd(&TextVar,',', " , ");
&TextVar = tranwrd(&TextVar,' w/', " with ");
*&TextVar = tranwrd(&TextVar,'/', " / ");
&TextVar = tranwrd(&TextVar,'\\', " \ ");
&TextVar = tranwrd(&TextVar,'~', " ~ ");
&TextVar = tranwrd(&TextVar,'!', " ! ");
&TextVar = tranwrd(&TextVar,'"s", " " ");
&TextVar = tranwrd(&TextVar,'_', " _ ");
&TextVar = tranwrd(&TextVar,'&', " and ");
&TextVar = tranwrd(&TextVar, '.', " . ");
&TextVar = tranwrd(&TextVar,'<=', " less than or equal ");
&TextVar = tranwrd(&TextVar,'>=', " greater than or equal ");
&TextVar = tranwrd(&TextVar,'<', " less than ");
&TextVar = tranwrd(&TextVar,'>', " greater than ");
&TextVar = tranwrd(&TextVar,'=', " equals ");
&TextVar = trim(left(compbl()));
%mend FixJunk;

data dmtm9.vaerext(keep=cage_yr sex symptom_text serious numdays pedflag sym_cnt
                  vax_1-vax_16 vax_cnt immun_cnt costring v_adminby v_fundby);
length cotermin $ 25 costring $255;
array syms{20} $ 25 sym01-sym20;
array vaxs{8} $ vax1-vax8;
array nvax{16} vax_1-vax_16;

set dmtm9.vaerall;

/* Only include adverse events that occurred within 90 days of vaccination */
if numdays <= 90;
if cage_yr = . then cage_yr = 0;
if cage_mo = . then cage_mo = 0;
if vax_date ne .;

/* Serious events are ones that required an overnight hospital stay or caused */
/* disability, death, or a life-threatening event */
if l_threat='Y' or died='Y' or hospital='Y' or x_stay='Y' or disable='Y'
then serious='Y';
else serious='N';

/* Determine age of vaccine recipient -- year + month, mark all those under */
/* 9 as pediatric */
cage_yr = cage_yr+cage_mo;
if cage_yr <=9 then pedflag='Y'; else pedflag='N';
if died= ' ' then died='N';
if er_visit = ' ' then er_visit='N';
if recovd = ' ' then recovd='U';

/* Since serious adverse events are rare (approx 8%) oversample serious events*/
if serious='N' and uniform(0) < .7 then delete;

/* Create flag variables for illnesses frequently inoculated against, also*/
/* count up number of immunizations given at one time to a patient as immun_cnt*/
label vax_1='Anthrax'

```

```

vax_2='Diphtheria/Tetanus'
vax_3='Flu'
vax_4='Hepatitis A'
vax_5='Hepatitis B'
vax_6='HIB (Haemophilus)'
vax_7='Polio (IPV,OPV)'
vax_8='Measles,Mumps,Rubella'
vax_9='Meningococcal'
vax_10='Pneumo (7-valent)'
vax_11='Pneumo (23-valent)'
vax_12='Rabies'
vax_13='Smallpox'
vax_14='Typhoid'
vax_15='Pertussis'
vax_16='Varicella'
;

do i=1 to 16;
  nvax{i}=0;
end;

immun_cnt=0;
do i=1 to min(vax_cnt,8);
  select (vaxs{i});
    when ('6VAX-F') do; vax_2=1; vax_5=1; vax_6=1; vax_7=1;
      immun_cnt=immun_cnt+5; end;
    when ('ANTH') do; vax_1=1; immun_cnt=immun_cnt+1; end;
    when ('DPP') do; vax_2=1; vax_15=1; vax_7=1; immun_cnt=immun_cnt+4; end;
    when ('DT','DTP','TD','TTP') do; vax_2=1; immun_cnt=immun_cnt+2; end;
    when ('DTAP','DTP','TDAP') do;
      vax_2=1; vax_15=1; immun_cnt=immun_cnt+3; end;
    when ('DTAPH','DTPHIB') do;
      vax_2=1; vax_15=1; vax_6=1; immun_cnt=immun_cnt+4; end;
    when ('DTAPHE') do;
      vax_2=1; vax_15=1; vax_5=1; vax_7=1; immun_cnt=immun_cnt+5; end;
    when ('FLU','FLUN') do; vax_3=1; immun_cnt=immun_cnt+1; end;
    when ('HBHEPB') do; vax_6=1; vax_5=1; immun_cnt=immun_cnt+2; end;
    when ('HBPV','HBVC','HIBV') do; vax_6=1; immun_cnt=immun_cnt+1; end;
    when ('HEP') do; vax_5=1; immun_cnt=immun_cnt+1; end;
    when ('HEPA') do; vax_4=1; immun_cnt=immun_cnt+1; end;
    when ('HEPAB') do; vax_4=1; vax_5=1; immun_cnt=immun_cnt+2; end;
    when ('IPV','OPV') do; vax_7=1; immun_cnt=immun_cnt+1; end;
    when ('MEA','MER','MM','MMR','MU','MUR','RUB') do;
      vax_8=1; immun_cnt=immun_cnt+3; end;
    when ('MMRV') do; vax_8=1; vax_16=1; end;
    when ('MEN','MNC','MNQ') do; vax_9=1; end;
    when ('PNC') do; vax_10=1; immun_cnt=immun_cnt+1; end;
    when ('PPV') do; vax_11=1; immun_cnt=immun_cnt+1; end;
    when ('RAB','RABA') do; vax_12=1; immun_cnt=immun_cnt+1; end;
    when ('SMALL') do; vax_13=1; immun_cnt=immun_cnt+1; end;
    when ('TYP') do; vax_14=1; immun_cnt=immun_cnt+1; end;
    when ('VARCEL') do; vax_16=1; immun_cnt=immun_cnt+1; end;
    otherwise;
  end;
end;

end;

```

```

        if immun_cnt > 0;

        /* Create a field, costring, with all the constart terms concatenated */
        /* together in one string */
        costring = '';

        do i=1 to min(sym_cnt,20);
            coterm = syms{i};
            costring=trim(costring) || ' ' || trim(coterm);
        end;

        /* Fix punctuation issues */
        %FixJunk(textvar=symptom_text);

run;

proc freq;
    tables pedflag immun_cnt vax_cnt v_adminby v_fundby sex serious vax_1-vax_16;
run;

```


Index

A

accessibility features [4](#)
Automatically Cluster property [20](#)

C

Classification Chart [48](#)
cleaning data [29](#)
 creating a stop list [36](#)
 creating a synonym data set [31](#)
 examining results using merged synonym data sets [34](#)
 exploring result improvements [39](#)
 using a synonym data set [30](#)
Cluster properties [20](#)
clusters [22](#)
Coding Symbols for Thesaurus of Adverse Reaction Terms [6](#)
compatibility [4](#)
concept linking [23](#)
converting files into data sets [67](#)
COSTART coding system [41](#)
COSTART node [41](#)
COSTRING variable [41](#)

D

data cleaning
 See [cleaning data](#)
Data Partition node [18](#)
data segments [24](#)
data sets
 converting files into [67](#)
 importing [29](#)
 merged synonym data sets [34](#)
 synonym data sets [30, 31](#)
data source
 creating for projects [13](#)
descriptive mining [1](#)
Descriptive Terms property [20](#)
diagrams

 creating [15](#)

Different Parts of Speech property [19](#)
document analysis [3](#)
document requirements [1](#)
documents
 from unsupported language or encoding [66](#)
 large collection of [65](#)
 long [65](#)

E

encoding
 unsupported [66](#)

F

file preprocessing [3](#)
files
 converting into data sets [67](#)

H

Help [9](#)

I

Ignore Outliers property [20](#)
Ignore Parts of Speech property [19](#)
importing data sets [29](#)
input data
 identifying [17](#)
 partitioning [18](#)
interactive results
 viewing [21](#)

L

languages
 unsupported [66](#)
large collection of documents [65](#)

library
 create 12
 long documents 65

M

macros
 %TMFILTER 67
 MedDRA program 49
 merged synonym data sets 34
 misspelled terms 31
 Model Comparison node 47
 modeling
 See [predictive modeling](#)

P

Parse properties 19
 Parse Variable 21
 partitioning input data 18
 path for projects 12
 predictive mining 1
 predictive modeling 1
 comparing models 47
 COSTRING variable for 41
 exercises 48
 SYMPTOM_TEXT variable for 45
 projects
 creating 11
 creating data source 13
 creating diagrams 15
 path for 12
 setting up 11

R

results
 examining with merged synonym data sets 34
 exploring result improvements 39
 viewing 21
 ROC Chart 47

S

SAS Enterprise Miner 6.1 2
 SAS Text Miner 4.1 2
 Help 9
 SAS Text Miner 4.2
 accessibility features 4
 Section 508 standards 4
 Segment Profile node 25
 segments 24

stems 31
 stop lists
 about 30
 creating 36
 defined 36
 SYMPTOM_TEXT variable analysis
 examining data segments 24
 identifying input data 17
 partitioning input data 18
 setting node properties 18
 viewing interactive results 21
 SYMPTOM_TEXT variable for modeling 45
 synonym data sets
 creating 31
 merged 34
 Synonyms property 19

T

Term Weight property 20
 Terms in a Single Document property 19
 Terms window 21
 text cleaning
 See [cleaning data](#)
 Text Miner node
 about 2
 properties 18
 text mining
 descriptive mining 1
 document requirements for 1
 general order for 3
 large collection of documents 65
 long documents 65
 predictive mining 1
 process 3
 tips for 65
 unsupported language or encoding 66
 text parsing 3
 tips for text mining 65
 Transform properties 20
 transformation 3

U

unsupported languages or encoding 66

V

Vaccine Adverse Event Reporting System 5
 VAERS data preprocessing 69

Your Turn

We welcome your feedback.

- If you have comments about this book, please send them to **`yourturn@sas.com`**. Include the full title and page numbers (if applicable).
- If you have comments about the software, please send them to **`suggest@sas.com`**.

