



Getting Started with SAS[®] Text Miner 3.1

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2007. *Getting Started with SAS® Text Miner 3.1*. Cary, NC: SAS Institute Inc.

Getting Started with SAS® Text Miner 3.1

Copyright © 2002–2007, SAS Institute Inc., Cary, NC, USA

ISBN 978–1–59047–695–6

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

For a hard-copy 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.

March 2007

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/pubs 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

Chapter 1	△ Introduction to Text Mining and SAS Text Miner	1
What Is Text Mining?		1
What Is SAS Text Miner?		2
The Text Mining Process		3
Accessibility Features of SAS Text Miner 3.1		3
Chapter 2	△ Learning by Example: Text Mining Using SAS Text Miner 3.1	5
About the Scenario in This Book		5
Prerequisites for This Scenario		10
How to Get Help for SAS Text Miner 3.1		10
Chapter 3	△ Setting Up Your Project	11
About the Tasks that You Will Perform		11
Create a Project		11
Create a Data Source		14
Create a Diagram		18
Chapter 4	△ Analyzing the SYMPTOM_TEXT Variable	19
About the Tasks that You Will Perform		19
Identify Input Data		19
Partition the Input Data		20
Set Text Miner Node Properties		20
View Interactive Results		23
Examine Data Segments		28
Chapter 5	△ Cleaning Up Text	33
About the Tasks that You Will Perform		33
Use a Synonym Data Set		35
Create a New Synonym Data Set		37
Examine Results Using Merged Synonym Data Sets		40
Create a Stop List		43
Explore Result Improvements		46
Chapter 6	△ Predictive Modeling with Text Variables	51
About the Tasks that You Will Perform		51
Use the COSTRING Variable to Model		51
Use the SYMPTOM_TEXT Variable to Model		58
Compare the Models		61
Additional Exercises		64
Chapter 7	△ Next Steps: A Quick Look at Additional Features	65
The %TMFILTER Macro		65
The %TMPUNC Macro		65

Tips for Text Mining 66

Appendix 1 △ Recommended Reading 69

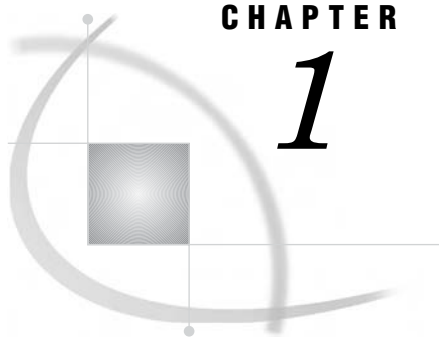
Recommended Reading 69

Appendix 2 △ Vaccine Adverse Event Reporting System Data Preprocessing 71

VAERS Data Preprocessing 71

Glossary 77

Index 81



CHAPTER

1

Introduction to Text Mining and SAS Text Miner

<i>What Is Text Mining?</i>	1
<i>What Is SAS Text Miner?</i>	2
<i>The Text Mining Process</i>	3
<i>Accessibility Features of SAS Text Miner 3.1</i>	3

What Is Text Mining?

Text mining helps you understand what textual documents tell you without having to read every word. Text mining uncovers the underlying themes or concepts that are contained in large document collections. Text mining applications fall into two areas: exploring the textual data for its content and then using the information to improve the existing processes. Both are important and can be referred to as *descriptive mining* and *predictive mining*.

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

Predictive mining involves classifying the documents into categories and using the information that is implicit in the text for decision making. You might want to identify the customers who ask standard questions so that they receive an automated answer. Or you might want to predict whether a customer is likely to buy again, or even if you should spend more effort in keeping him or her as a customer.

Predictive modeling involves examining past data to predict future results. You might have a data set that contains information about past buying behaviors, along with comments that the customers made. You can then build a predictive model that can be used to score new customers: to analyze new customers based on the data from past customers. For example, if you are a researcher for a pharmaceutical company, you know that hand-coding adverse reactions from doctors' reports in a clinical study is a laborious, error-prone job. Instead, you could create a model by using all your historical textual data, noting which doctors' reports correspond to which adverse reactions. When the model is constructed, processing the textual data can be done automatically by scoring new records that come in. You would just have to examine the "hard-to-classify" examples, and let the computer handle all the rest.

Both of these aspects of text mining share some of the same requirements. Namely, text documents that human beings can easily understand must first be represented in a form that can be mined by the software. The raw documents need processing before the patterns and relationships that they contain can be discovered. Although the human

mind comprehends chapters, paragraphs, and sentences, computers require structured (quantitative or qualitative) data. As a result, an unstructured document must be converted into a structured form before it can be mined.

What Is SAS Text Miner?

SAS Text Miner is an add-on for the SAS Enterprise Miner environment. Enterprise Miner provides a rich set of data mining tools that facilitate the prediction aspect of text mining. The integration of SAS Text Miner within SAS Enterprise Miner combines textual data with traditional data mining variables. A Text Miner node can be embedded into a SAS Enterprise Miner process flow diagram. SAS Text Miner supports various sources of textual data: local text files, text as observations in SAS data sets or external databases, and files on the Web. The Text Miner node encompasses the parsing and exploration aspects of text mining and sets up the data for predictive mining and further exploration using other Enterprise Miner nodes. This enables you to analyze the new structured information that you have acquired from the text however you want, combining it with other structured data as desired.

The node is highly customizable and allows a variety of parsing options. It is possible to parse documents for detailed information about the terms, phrases, and other entities in the collection. You can also cluster the documents into meaningful groups and report the concepts that you discover in the clusters. All this is done in an environment that enables you to interact with the collection. Sorting, searching, filtering (subsetting), and finding similar terms or documents all enhance the exploration process.

The Text Miner node's extensive parsing capabilities include

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

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

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

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

The Text Mining Process

Whether you intend to use textual data for descriptive purposes, predictive purposes, or both, the same processing steps take place, as shown in Table 1.1.

Table 1.1 General Order for Text Mining

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

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

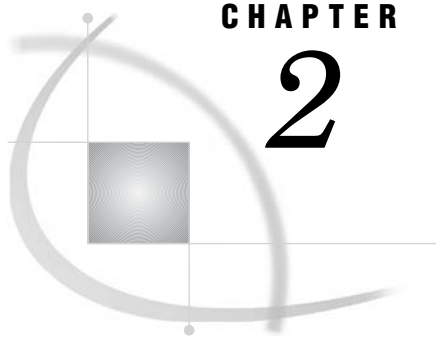
You might not need to include all of these steps in your analysis, and it might be necessary to try a different combination of text-parsing options before you are satisfied with the results.

Accessibility Features of SAS Text Miner 3.1

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

Section 508 Accessibility Criteria	Support Status	Explanation
When software is designed to run on a system that has a keyboard, product functions shall be executable from a keyboard where the function itself or the result of performing a function can be discerned textually.	Supported with exceptions.	<p>The software supports keyboard equivalents for all user actions with the exceptions noted below:</p> <ul style="list-style-type: none"> <input type="checkbox"/> 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: <ul style="list-style-type: none"> <input type="checkbox"/> Primary window - Shift+F10+spacebar <input type="checkbox"/> Secondary window - Shift+F10+down key. <input type="checkbox"/> The Explore action in the data source popup menu cannot be invoked directly from the keyboard, but there is an alternative way to invoke the data source explorer using the View ► Table menu.
Color coding shall not be used as the only means of conveying information, indicating an action, prompting a response, or distinguishing a visual element.	Supported with exception.	Node run or failure indication relies on color, but there is always the corresponding message in the bottom right panel of the main window.

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



CHAPTER

2

Learning by Example: Text Mining Using SAS Text Miner 3.1

<i>About the Scenario in This Book</i>	5
<i>Prerequisites for This Scenario</i>	10
<i>How to Get Help for SAS Text Miner 3.1</i>	10

About the Scenario in This Book

This book is intended for SAS Text Miner 3.1 users. Each topic in this book builds on the previous topic, so you must work through the chapters in sequence.

This book uses an extended example that is intended to familiarize you with the many features of Text Miner. Several key components of the Text Miner process flow diagram are covered. In this step-by-step example, you learn to do basic tasks in Text Miner: you 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 example is designed to be used in conjunction with SAS Text Miner 3.1 software.

The Vaccine Adverse Event Reporting System (VAERS) data is publicly available from the U.S. Department of Health and Human Services (HHS). Due to the Freedom of Information Act, 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 1991. 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. Please see the Guide to Interpreting Case Report Information Obtained from the Vaccine Adverse Event Reporting System (VAERS) available at <http://vaers.hhs.gov/info.htm>.

See the following in the Getting Started with SAS Text Miner 3.1 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 10 for information on where to download the Getting Started with SAS Text Miner 3.1 zip file. Δ

The following figure shows the first 14 columns in the table of VAERS data for 2005. Included is a unique identifier, the state of residence, the recipient’s age, and an unstructured text string SYMPTOM_TEXT containing the reported problem.

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 recently 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 is still used. The FDA uses 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.

Note that, from the figure here, VAERS_ID 231844 has SYMPTOM_TEXT of 101 fever, stiff neck, cold—the program has automatically extracted the COSTART terms **FEVER, NECK RIG, PHARYNG, RHINITIS** from this text.

A	B	C	D	E	F	G	H	I	J	K	L	M	N
VAERS_ID	RECVDAT	STATE	AGE_YR	SAGE_YR	CAGE_MC	SEX	RPT_DATE	SYMPTOM_TEXT	SYM_CNT	SYM01	SYM02	SYM03	SYM04
231786	1/1/2005	MA	63	63		F	1/1/2005	HOARSENESS, DIZZINES	2	DIZZINESS	VOICE ALTERAT		
231787	1/2/2005	MD	30	30		F	1/2/2005	fever - 101 degrees for 2 d	5	ASTHENIA	CHILLS	COUGH IN	FEVER
231788	1/2/2005	VA	18	18		F	1/2/2005	Sunburn-like rash on upper	5	ASTHENIA	MALAISE	PHARYNG	RASH
231789	1/2/2005	PA	1.3	1	0.3	M	1/2/2005	Within 7-10 days of reciev	8	AUTISM	DIARRHEA	FEVER	INFECT VI
231790	1/2/2005	CA	16	15		M	1/2/2005	Immediately saw sparkles	16	CHILLS	CRAMPS	DIZZINES	FEVER
231791	1/2/2005		20	19		M	1/2/2005	CHEST PAIN, HEADACH	4	ECG ABN	HEADACH	LAB TEST	PAIN CHE
231829	1/3/2005	DC	45	45		M	12/28/2004	Developed proteinuria whic	9	ALBUMIN	COAGUL	LEMB PULI	HEPATOM
231830	1/3/2005	TN	90	89		F	12/22/2004	From initial information rec	5	APNEA	DIZZINES	KIDNEY F	PNEUMON
231838	1/3/2005	CA	1.1	1	0.1	F	12/27/2004	Patient had febrile seizure	4	CHILLS	FEBRILE	FEVER	LEUKOCY
231839	1/3/2005	LA	59	58		F	12/17/2004	02/15/2004 Received flu a	1	FEVER			
231840	1/3/2005	LA	64	64		F	12/17/2004	Noticed right of 12/15/04 l	2	EDEMA IN	HYSN IN	JECT SITE	
231841	1/3/2005	CA	46	45		F	11/6/2004	Starting on 10/25/04 the fr	15	ANOREXIA	ARTHRAL	ASTHENIA	COORDIN
231842	1/3/2005	LA	67	66		F	12/15/2004	Received Pneumovax vacc	4	CHILLS	EDEMA IN	HYSN INJ	INJECT SI
231843	1/3/2005	NJ	32	32		M	9/10/2004	Developed alopecia in mid	1	ALOPECIA			
231844	1/3/2005	MA	9	9		F	12/23/2004	101 fever, stiff neck, cold	4	FEVER	NECK RIG	PHARYNG	RHINITIS
231845	1/3/2005		20	19		M	11/10/2004	Rash on chest and arms,	1	RASH			
231846	1/3/2005	TX	3	3		M	12/23/2004	Redness, swelling of arm,	2	EDEMA IN	HYSN IN	JECT SITE	
231847	1/3/2005		36	36		M	12/12/2004	Right shoulder, bicep, wris	2	ASTHENIA	PARESTHESIA		
231848	1/3/2005	DC	20	19		U	12/28/2004	New onset migraine after	1	MIGRAINE			
231849	1/3/2005		24	24		M	12/29/2004	Developed peeling palms f	2	DERM EX	PARESTHESIA		
231850	1/3/2005	ND	13	12		M	9/30/2004	Fine red rash on Abdomen	4	PAIN	PRURITUS	RASH	VASODILA
231851	1/3/2005	WI	36	35		M	9/29/2004	Probable cellulitis second	1	CELLULITIS			
231852	1/3/2005	DE	72	72		F	12/15/2004	Received PPV23 approx a	5	CELLULITIS	ECCHYM	EDEMA IN	HYSN INJ
231853	1/3/2005	OH	1	1	0	U	12/2/2004	Rash, pink maculopapular	3	PRURITUS	RASH MA	RASH VESIC	BULL
231854	1/3/2005	VA	1.1	1	0.1	F	12/15/2004	Pt developed a severe itch	5	ASTHENIA	FEVER	HERPES	PAIN
231855	1/3/2005	LA	15	15		F	12/22/2004	Small red bumps lesions i	3	RASH	ULCER SI	VASODILAT	
231856	1/3/2005	WV	1	1	0	F	12/28/2004	Chicken pox	1	INFECT VIRAL			
231857	1/3/2005	WV	1	1	0	F	12/28/2004	Chicken pox	1	INFECT VIRAL			
231858	1/3/2005	AL	1.5	1	0.5	M	12/30/2004	12/08/2004 Prevnar given.	4	ERYTHEM	FEVER	ULCER SI	VEIN VAR
231859	1/3/2005	LA	4	4		F	12/23/2004	Erythromatic Rash 6 x 4 c	2	RASH	VASODILAT		
231860	1/3/2005	LA	1.3	1	0.3	M	12/20/2004	Expired immunization give	1	MED ERROR			
231861	1/3/2005	LA	5	5		F	12/16/2004	PN 12/15/2004 late PM pr	5	ABSCESS	EDEMA IN	HYSN INJ	INJECT SI
231862	1/3/2005	LA	1.4	1	0.4	F	12/20/2004	Expired immunization give	1	MED ERROR			

The following figure shows other columns in the table, including a variety of flags that indicate the seriousness of the event (life-threatening illness, emergency room or doctor visit, hospitalized, disability, 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 indicating where the shots were given, who funded them, what medications the patient was taking, and so on.

	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC
1	L	THREATEN_VIST	HOSPITAL	HOSPDAY	STAY	DISABLE	REC	OV	VAX_DATE	NUMDAYS	LAB_DAT	VAX_CHT	VAX1	VAX2	VAX3	VAX4	VAX5	VAX6	VAX7	VAX8	V_ADMIN	FUNCB	OTHER_N
2									12/1/2004	1			1 FLU								UNK	UNK	
3							N		12/30/2004	1			1 FLUN								UNK	UNK	
4							Y		12/27/2004	1			1 FLUN								UNK	UNK	
5								Y	3/8/2002	7	Measles K	2	MMR	RNC							UNK	UNK	
6	Y	Y	Y		3			Y	10/13/2004	0	Medical Rv	1	FLU								UNK	UNK	None
7								U	12/22/2004	10	ABNCRMP	1	SMALL								UNK	UNK	NONE
8								N	5/24/2000	15	BP 102/60	1	ANTH								MIL	MIL	NONE
9		Y	Y			63		N	10/11/2004	0	EKG, CT s	1	FLU								PVT	UNK	Liptor, Pl
10			Y			1		Y	12/21/2004	0	Information	3	FLU	MMR	VARCEL						PVT	PVT	
11								U	12/15/2004	0		2	FLU	PPV							PUB	UNK	Vi D, Cel
12								U	12/15/2004	0		3	FLU	PPV	TIGI						PUB	UNK	Serish and
13		Y						N	10/12/2004	13	None and	1	FLU								OTH	PVT	Ectozel 4mg
14		Y						U	12/14/2004	1	NONE	2	FLU	PPV							PUB	PUB	UNK
15								N	1/13/2004	34	Test for dit	3	FLU	HEP	TYD						MIL	MIL	
16								Y	12/2/2004	8		1	FLUN								PVT	PVT	NONE
17								Y	10/15/2004	25	NONE	2	FLU	SMALL							MIL	MIL	NONE
18								U	12/21/2004	1	NONE	1	FLU								PVT	PVT	Zitrosman
19		Y						U	9/5/2004	88	Pending E	1	ANTH								MIL	MIL	Proscam
20								N	12/11/2003	0	NONE	1	ANTH								MIL	MIL	NONE
21								N	6/30/2004	0		1	ANTH								MIL	MIL	NONE
22								Y	9/22/2004	3	NONE	1	HEP								OTH	PUB	Benadryl
23		Y						Y	12/1/2004	11		1	SMALL								MIL	MIL	
24		Y						U	12/10/2004	0		1	PPV								MIL	UNK	
25								U	2/26/1999	2098		1	VARCEL								PVT	PVT	
26		Y						Y	10/4/2001	1126	Doctor visi	2	IPV	VARCEL							PVT	PVT	NONE
27		Y						Y	12/21/2004	0	Sent to all	3	HEP	TD	VARCEL						PVT	PUB	NONE
28		Y						Y	7/23/1999	1962		1	UNK								PVT	PVT	
29		Y						Y	4/23/1998	0		1	UNK								PVT	PVT	
30		Y						Y	12/8/2004	1	CBG	1	PNC								PVT	UNK	
31		Y						U	12/20/2004	2		3	DTAP	IPV	MMR						PVT	UNK	
32		Y						Y	12/16/2004	1	IVA	1	DTAPH								PVT	OTH	NONE

The README file taken from the VAERS Web site decodes the vaccine abbreviations. Note that some vaccinations contain multiple vaccines (e.g., 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 PERTUSSIS
DTAPH  =  TETRAMUNE
DTIPV  =  DT-IPV COMBINED DT AND IPV VACCINE
DTOX   =  DIPHTHERIA TOXOID
DTP    =  DIPHTHERIA AND TETANUS TOXOIDS AND PERTUSSIS VACCINE
DTPH   =  DIPHTHERIA, TETANUS TOX, PERTUSSIS,& HAEMOPHILUS
DTPHIP =  COMBINATION of DTP, IPV and ACT-HIB
DTIPV  =  DTP-IPV COMBINED DTP AND IPV VACCINE
FLU    =  INFLUENZA VIRUS VACCINE
HBHEPB =  COMVAX
HBPV   =  HAEMOPHILUS B POLYSACCHARIDE VACCINE
HEPB   =  HEPATITIS B VIRUS VACCINE
HEPA   =  HEPATITIS A
HIBV   =  HAEMOPHILUS B CONJUGATE VACCINE
IPV    =  POLIOVIRUS VACCINE INACTIVATED
JEV    =  JAPANESE ENCEPHALITIS VIRUS VACCINE
M      =  MEASLES VIRUS VACCINE, LIVE
M-RV   =  RUBEOVAX-LIVE MEASLES DISCONTINUED DEC. 1971
MEN    =  MENINGOCOCCAL POLYSACCHARIDE VACCINE
MM     =  MEASLES AND MUMPS VIRUS VACCINE, LIVE
MMR    =  MEASLES, MUMPS AND RUBELLA VIRUS VACCINE, LIVE
MR     =  MEASLES AND RUBELLA VIRUS VACCINE, LIVE
MU     =  MUMPS VIRUS VACCINE, LIVE
MUR    =  MUMPS AND RUBELLA VIRUS VACCINE, LIVE
OPV    =  POLIOVIRUS VACCINE TRIVALENT, LIVE, ORAL
P      =  PERTUSSIS VACCINE
PLAGUE =  PLAGUE VACCINE
PPV    =  PNEUMOCOCCAL VACCINE, POLYVALENT
R      =  RUBELLA VIRUS VACCINE, LIVE
RAB    =  RABIES VIRUS VACCINE
SMALL  =  SMALLPOX VACCINE
TD     =  TETANUS AND DIPHTHERIA TOXOIDS, ADULT
TTOX   =  TETANUS TOXOID
TYP    =  TYPHOID VACCINE
VARCEL =  VARIVAX-VARICELLA VIRUS LIVE
YF     =  YELLOW FEVER VACCINE
CEE    =  Central European Encephalitis
DPPIV  =  Diphtheria/Pertussis/Inactivated polio virus
DPP    =  Diphtheria/Pertussis/Polio (oral [live] or inactivated
DTAPIP =  DTaP+IPV (Quadrace1®)
DTPHEP =  Diphtheria/tetanus/pertussis/hepatitis B
DTPIHI =  Diphtheria/Tetanus/whole pertussis-Inactivated Polio Vi
DTPO   =  Combined DTP and oral Polio Vaccines (Foreign Vaccine)
HBCV   =  Haemophilus B Conjugate vaccine (Foreign Vaccine)
MBR    =  Varicella (verivac®)
MGC    =  Meningitis Conjugate
PENT   =  Diphtheria/Tetanus/whole pertussis-Inactivated Polio
SEV    =  Spring/summer encephalitis vaccination

```

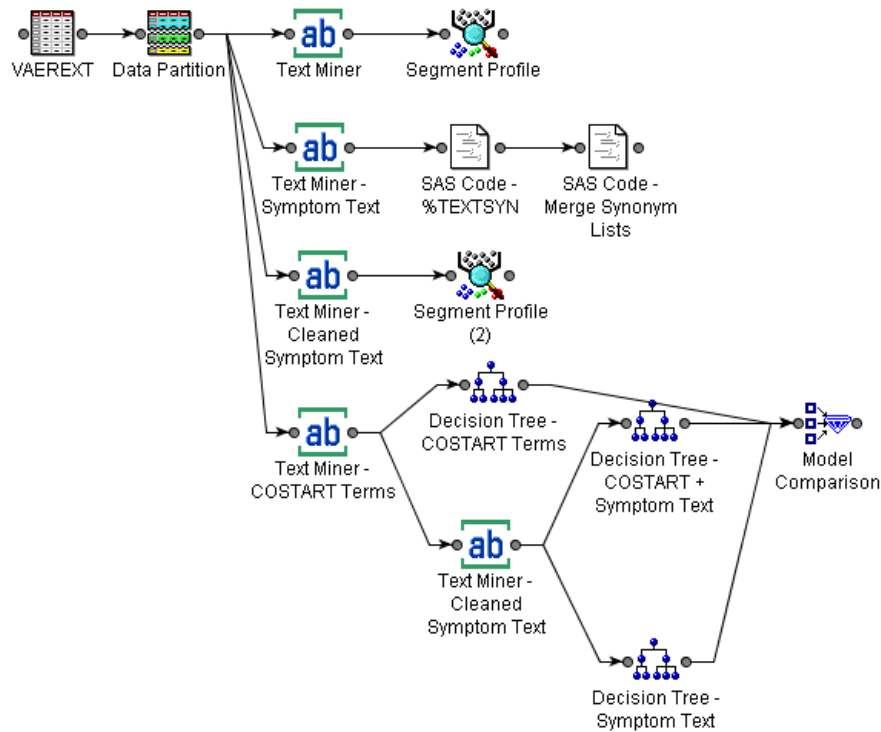
In addition to explaining more about text mining and the kind of business problems it can address, this example demonstrates some of the key capabilities of SAS Text Miner 3.1.

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 due to one or more of the vaccinations they are given, or they might be induced by an improper procedure from the administering lab (e.g., a non-sanitized needle). Some of them will be totally unrelated. For example, perhaps someone happened to get a cold just after receiving a

flu vaccine and reported it. You particularly want to look at serious reactions that required a hospital stay or caused a lifetime disability or death. You want to 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 they received the vaccine, or other pertinent information?
- What factors influence whether a reaction becomes serious?
- How well are these 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:



Prerequisites for This Scenario

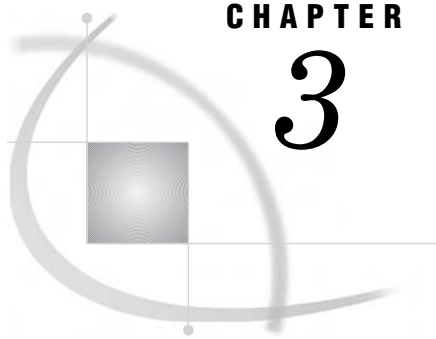
By the time that you are ready to perform the tasks in this book, administrators at your site should have installed and configured all necessary components of SAS Text Miner 3.1. For details about the SAS Text Miner 3.1 postinstallation guide, see: <http://support.sas.com/documentation/onlinedoc/txtminer/postinstall131.pdf>. You must also perform the following:

- Download the Getting Started with SAS Text Miner 3.1 data file in .zip format from the following URL: <http://support.sas.com/documentation/onlinedoc/txtminer>.
- Unzip this file into any folder in your file system.
- Create a folder called **vaersdata** on your C:\ drive.
- Copy the following files into **C:\vaersdata**:
 - Vaerext.Sas7bdat
 - Vaer_Abbrev.Sas7bdat
 - Engdict.Sas7bdat

How to Get Help for SAS Text Miner 3.1

Use any of the following methods to get Help for the SAS Text Miner 3.1:

- From the menu bar, select **Help ► Contents**
- Press F1 in most application windows.



CHAPTER

3

Setting Up Your Project

<i>About the Tasks that You Will Perform</i>	11
<i>Create a Project</i>	11
<i>Create a Data Source</i>	14
<i>Create a Diagram</i>	18

About the Tasks that You Will Perform

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

- 1 You will create a new project where you will store all your work.
- 2 After the project is created, you will define the VAERS data as an Enterprise Miner data source.
- 3 You will create a new process flow diagram in your project.

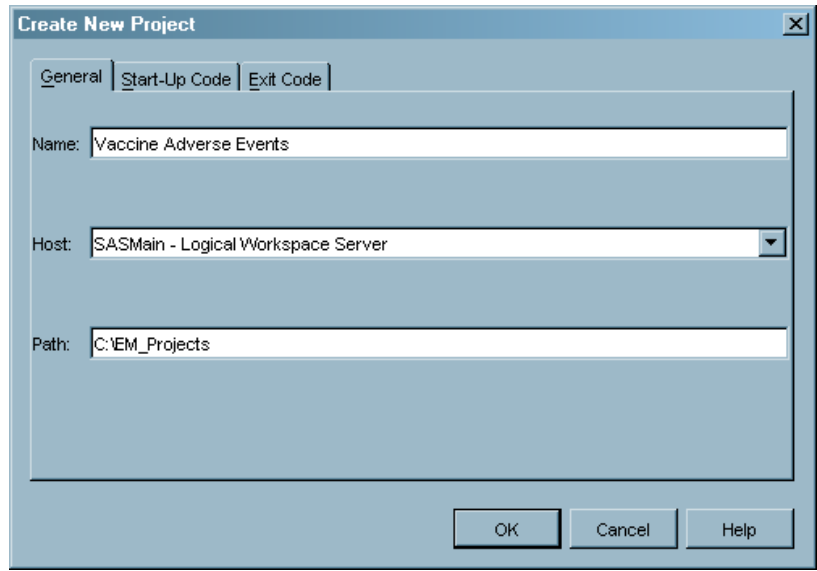
Create a Project

To create a project, complete the following steps:

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



- 3 The Create New Project window opens. In the **Name** box, type a name for the project, such as **Vaccine Adverse Events**.



- 4 In the **Path** box, type the path to the location on the server where you want to store the data that is associated with the example project. Your project path depends on whether you are running Enterprise Miner as a complete client on your local machine or as a client/server application.

If you are running Enterprise Miner as a complete client, your local machine acts as its own server. Your Enterprise Miner projects are stored on your local machine in a location that you specify, such as **C:\EM_Projects**.

If you are running Enterprise Miner as a client/server application, all projects are stored on the Enterprise Miner server.

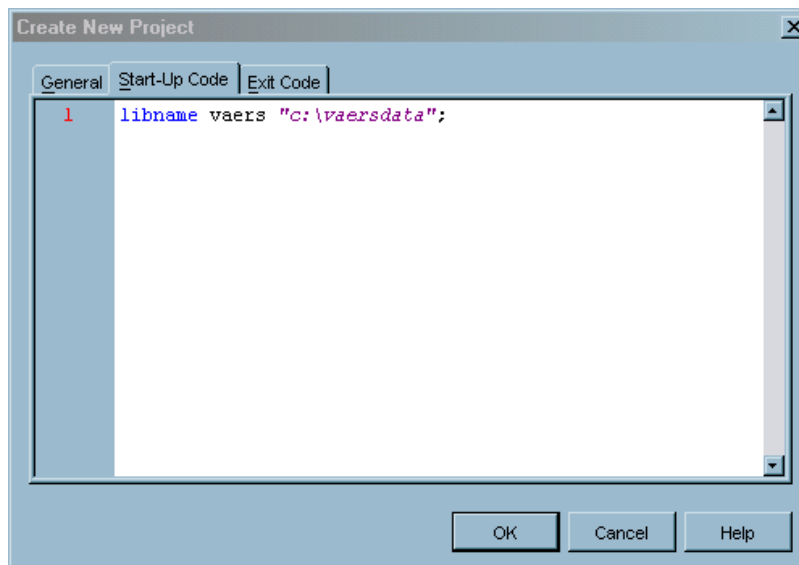
If you see a default path in the **Path** box, you can accept the default project path, or you can specify your own project path. This example uses **C:\EM_Projects**.

- 5 Click the **Start-Up Code** tab and enter the following code to create a SAS library:

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

Note: The location will depend on where you have stored the data for this tutorial on your system. The example uses the local path specification,

C:\Vaersdata. △



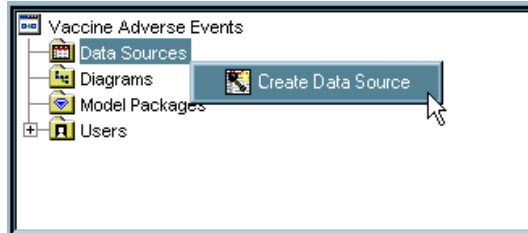
- 6 Click **OK**. The new project is created and opens automatically.

Note: Example results might differ from your results. Enterprise and Text Miner nodes and their statistical methods might incrementally change between releases. Your process flow diagram results might differ slightly from the results that are shown in this example. However, the overall scope of the analysis will be the same. △

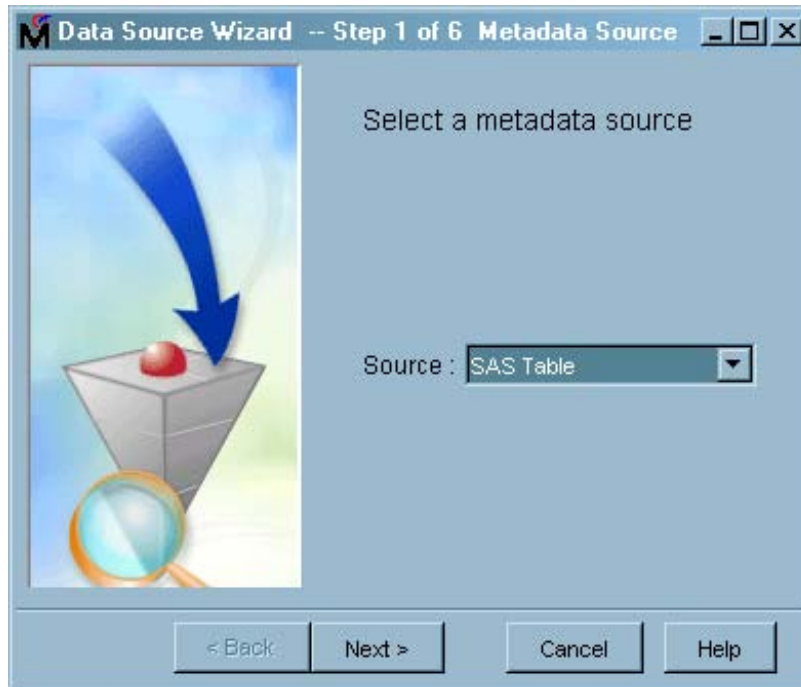
Create a Data Source

To create a data source, complete the following steps:

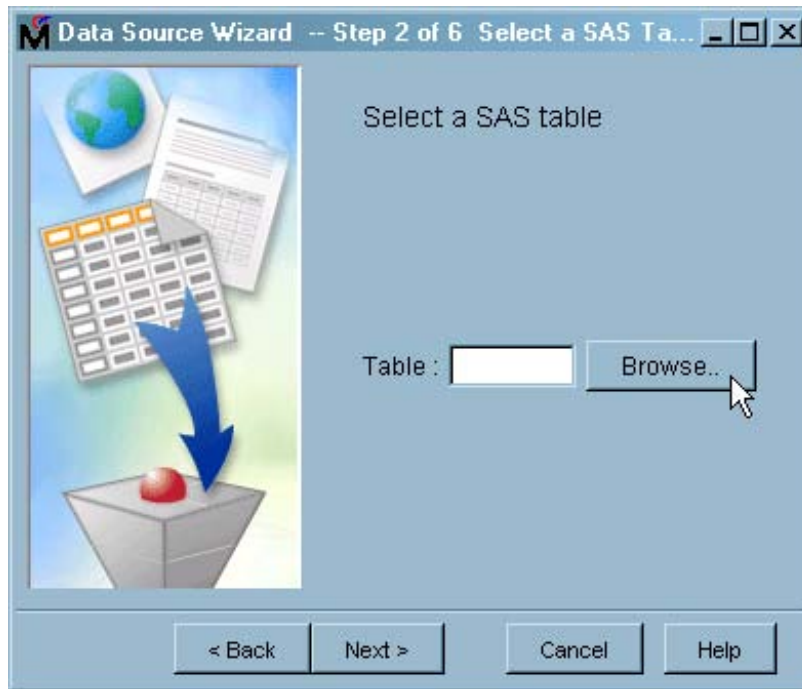
- 1 Right-click the **Data Sources** folder in the Project panel and select **Create Data Source** to open the Data Source wizard.



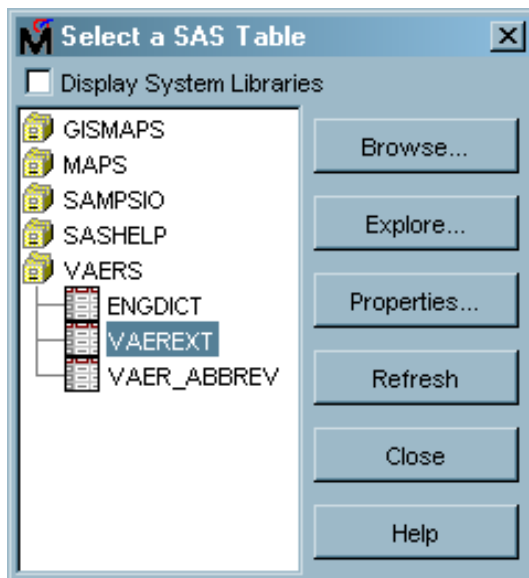
- 2 In the **Source** box of the Data Source Wizard — Metadata Source window, select **SAS Table** to tell SAS Enterprise Miner that the data is formatted as a SAS table.



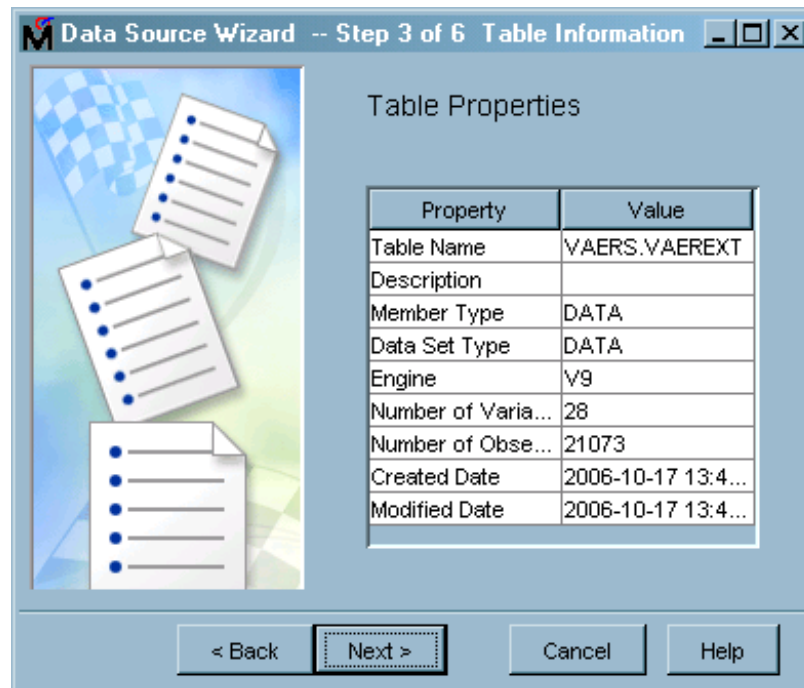
- 3 Click **Next**. The Data Source Wizard — Select a SAS Table window opens.
- 4 Click **Browse** in the Data Source Wizard — Select a SAS Table window. The Select a SAS Table window opens.



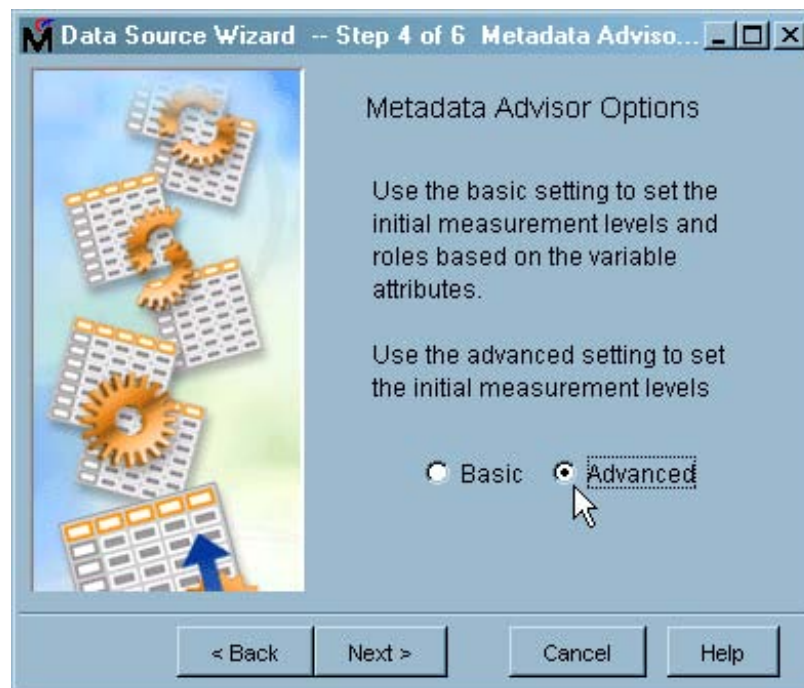
- 5 Double-click the SAS library named **VAERS**. The **VAERS** library folder expands to show all the data sets that are in the library. Select the **VAEREXT** table, and click **OK**. The two-level name **VAERS.VAEREXT** appears in the **Table** box of the Select a SAS Table window.



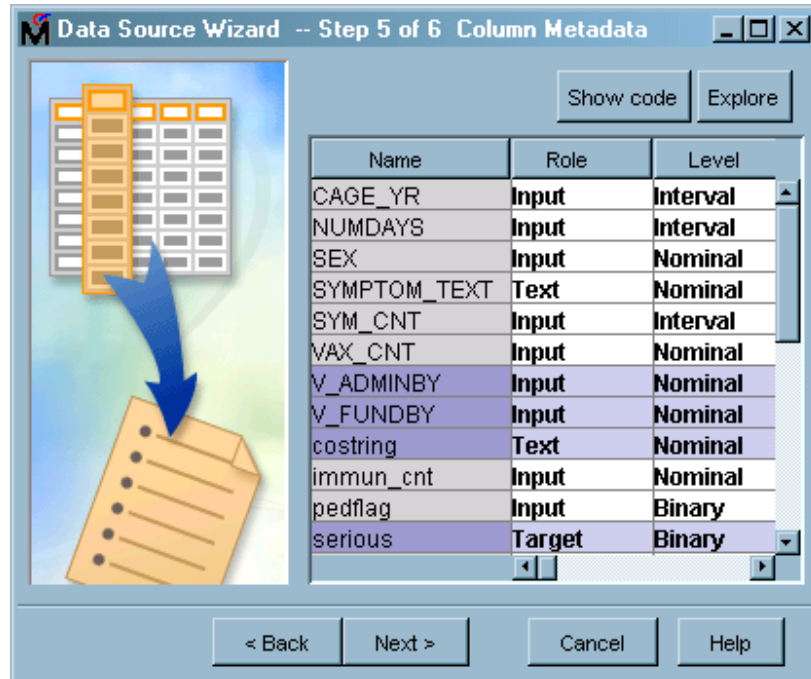
- 6 Click **next**. The Table Information window opens. Examine the metadata in the Table Properties section.



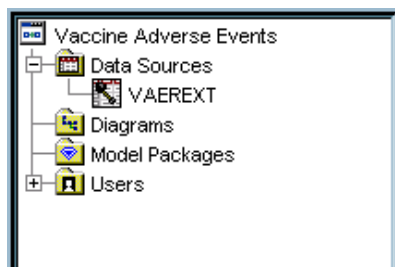
- 7 After you finish examining the table metadata, click **next**. The Data Source Wizard — Metadata Advisor Options window opens.
- 8 Select **Advanced** and click **next**.



- 9 The Data Source Wizard — Column Metadata window opens. Redefine these variable roles:
- 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**.



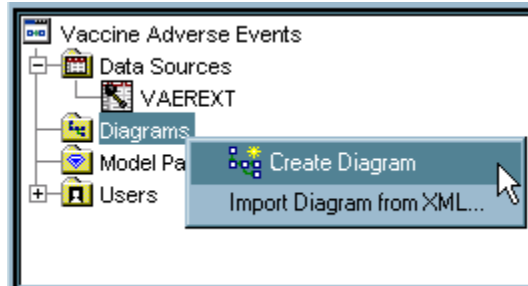
- 10 Click **Next**. The Decision Configuration window opens. Leave the default selection set to **No**.
- 11 Click **Next**. The Data Source Attributes window opens.
- 12 Click **Finish**. The VAEREXT table is added to the **Data Sources** folder of the Project panel. You might need to expand the **Data Sources** folder to see the new table.



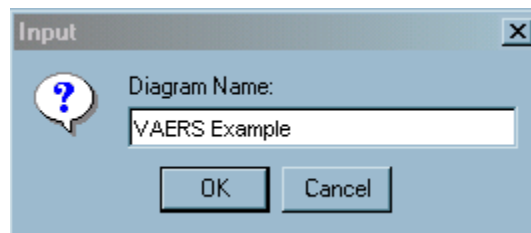
Create a Diagram

To create a diagram, complete the following steps:

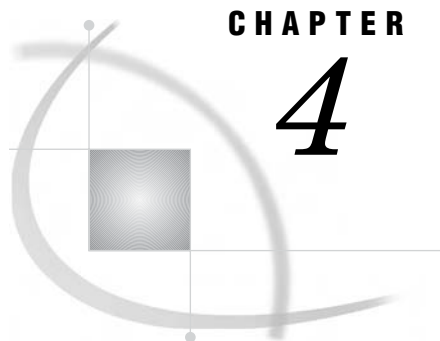
- 1 Right-click the **Diagrams** folder of the Project panel and select **Create Diagram**.



- 2 Enter **VAERS Example** in the **Diagram Name** box and click **OK**.



The empty **VAERS Example** diagram opens in the diagram workspace.



CHAPTER

4

Analyzing the SYMPTOM_TEXT Variable

<i>About the Tasks that You Will Perform</i>	19
<i>Identify Input Data</i>	19
<i>Partition the Input Data</i>	20
<i>Set Text Miner Node Properties</i>	20
<i>View Interactive Results</i>	23
<i>Examine Data Segments</i>	28

About the Tasks that You Will Perform

The SYMPTOM_TEXT variable contains the text of the adverse event as it was reported. This is the variable that you will analyze using SAS Text Miner. You will perform the following tasks:

- 1 Use the Input Data node to identify the VAERS data source.
- 2 After adding the VAERS data source to your process flow diagram, you will use the Data Partition node to partition the input data.
- 3 Set some Text Miner node properties using the Properties panel.
- 4 View the results using the Interactive Results window.
- 5 Use the Segment Profile node to examine data segments.

Identify Input Data

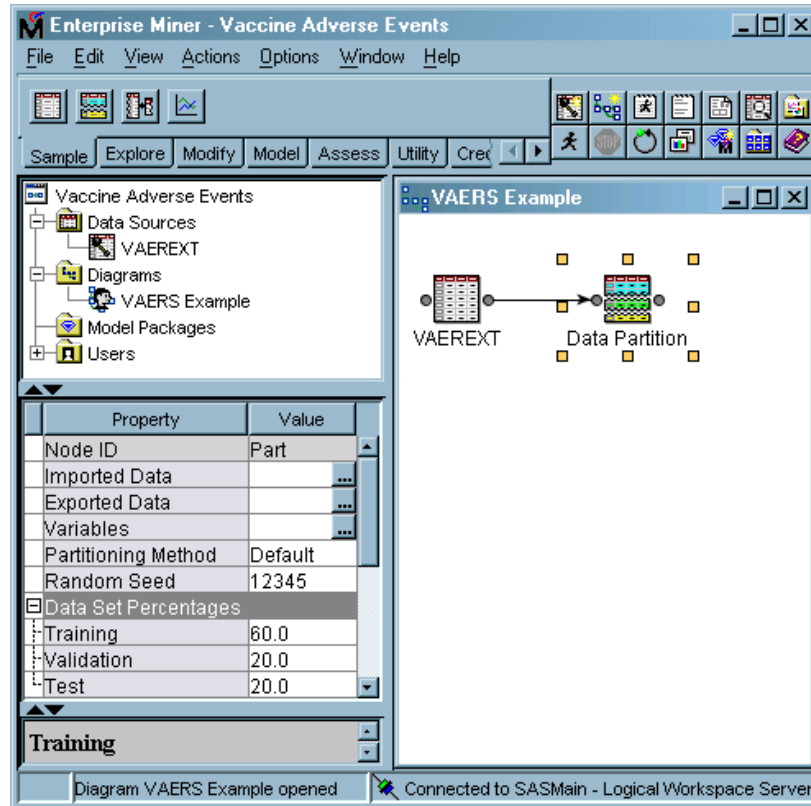
To identify input data, complete the following step:

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

Partition the Input Data

To partition the input data, complete the following step:

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



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

This will ensure adequate data when you build prediction models.

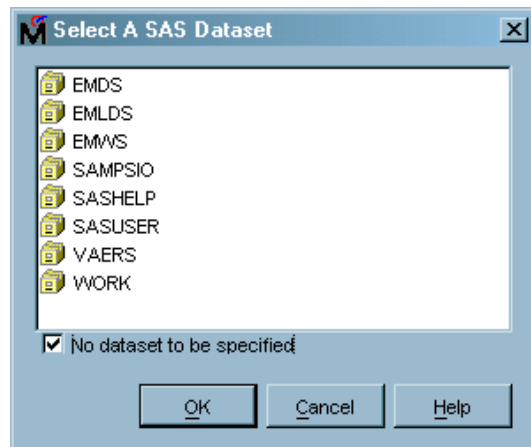
Set Text Miner Node Properties

To set the Text Miner node properties, complete the following steps:

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

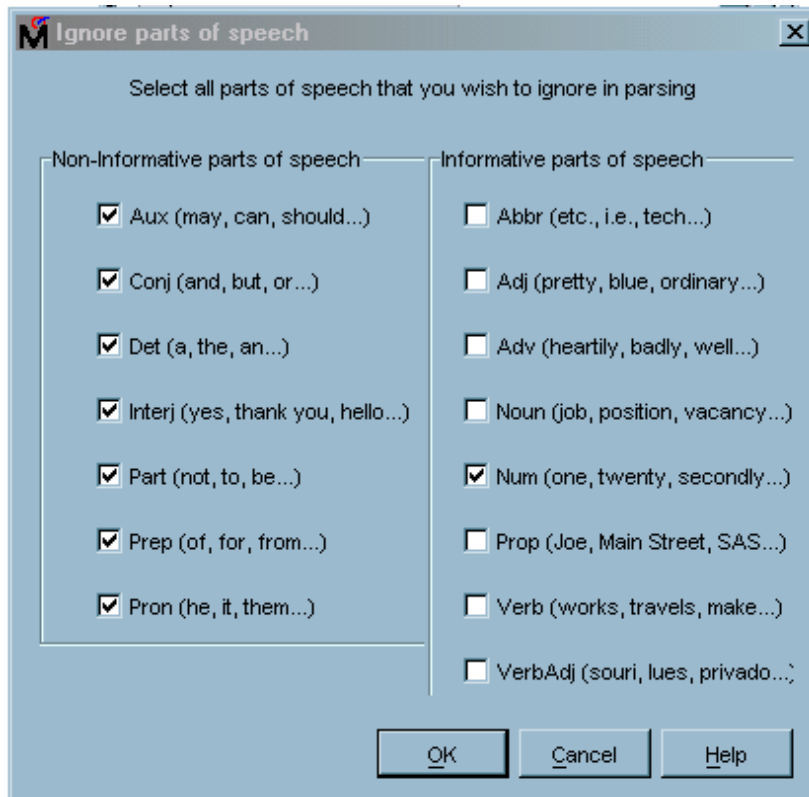


- 2 Select the Text Miner node to view its properties. Details about the node appear in the Properties panel. Set the following Parse properties:
 - Set **Terms in Single Document** to **Yes**. This will include all terms that occur only in a single document.
 - Set **Different Parts of Speech** to **No**. For the VAERS data, this setting offers a more compact set of terms.
 - Click the ellipsis button to the right of the Synonyms property. The Select A SAS Dataset window opens. Select **No dataset to be specified** and click **OK**.



- Click the ellipsis button to the right of 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 these parts of speech that you select in this dialog box are ignored during parsing. The selections indicated here ensure that the analysis ignores low-content words such as prepositions and determiners.



3 Set the following Transform properties:

- 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**. This eases 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 window when you are prompted with **Do you want to run this path?**.
- 7 After the Text Miner node has run, make sure that **Parse Variable** in the Property panel 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 Click the ellipsis button to the right of the Interactive property to open the Interactive Results window. The Text Miner — Interactive window opens.

Property	Value
Node ID	TEXT
Imported Data	...
Exported Data	...
Variables	...
Interactive	...
Rerun	No

- View the terms in the Terms window. The terms are sorted in decreasing order of occurrence.

TERM	Freq	# Docs	Keep	WEIGHT	Role	
+	have	6339	4154	<input checked="" type="checkbox"/>	0.084	
+	receive	5191	3690	<input checked="" type="checkbox"/>	0.059	
+	vaccine	4628	3430	<input checked="" type="checkbox"/>	0.011	
+	swell	3945	3271	<input checked="" type="checkbox"/>	0.221	
+	day	3896	2999	<input checked="" type="checkbox"/>	0.011	
+	arm	3956	2880	<input checked="" type="checkbox"/>	0.195	
+	fever	2951	2477	<input checked="" type="checkbox"/>	0.022	
+	leave	3104	2460	<input checked="" type="checkbox"/>	0.183	
	pt	4339	2455	<input checked="" type="checkbox"/>	0.045	
+	site	2681	2258	<input checked="" type="checkbox"/>	0.244	
+	injection	2742	2197	<input checked="" type="checkbox"/>	0.204	
+	report	3057	2178	<input checked="" type="checkbox"/>	0.045	
+	patient	3728	2096	<input checked="" type="checkbox"/>	0.118	
+	develop	2296	2045	<input checked="" type="checkbox"/>	0.022	
+	give	2493	2039	<input checked="" type="checkbox"/>	0.078	

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

at the far right of the toolbar to see the full text contained in SYMPTOM_TEXT.

The screenshot shows the 'Text Miner - Interactive' application window. The menu bar includes 'File', 'Edit', 'Tools', 'View', and 'Window'. The toolbar contains several icons, with a tooltip 'Toggle Show Full Text' pointing to the icon on the far right. Below the toolbar, the 'Documents' window is open, displaying the text for 'SYMPTOM_TEXT'.

SYMPTOM_TEXT

Information has been received from an RN concerning a 64 year old white, obese Sabin tri vaccines were not good ones. They make you taller and handicapped loo Cellulitus at administration site.

Demyelinating disease; dizziness, blurred vision; difficulty hearing and walking.

Autistic mannerisms, system "shutdown". Blank stares, catatonic state.

Loss of speech and coordination.

Reportedly called in after first dose to report had a rash that sounded like hives 2 Rash immediately following 03/21/1997 hep vaccine with arrested head growth an Large firm, red region at site of injection. Skin is itchy but not painful. Estimate 3 inc This report is concerning a 4 month old female who on 29-NOV-2001 was vaccin Five to ten minute seizure, generalized. No meds given for seizure. On ampicillin a Fever for 3 days. Emesis (recurrent) . Empiric antibiotics.

Fever, fussiness, questionable hematuria. 07/14: Right otitis media. 07/15: Persiste Temp 104. 5 within 24 hours of immunizations. Pt had septic URI admitted for 48 ho


4 View the clusters in the Clusters window.

#	Descriptive Terms	Freq	Percentage	RMS Std.
1	+ admit, + month, er, + hospitalize, + seizure, + discharge, + hospital, + diagnosis, + state, + day, + have, + patient	1029	0.081388910...	0.1433842...
2	er, febrile, + fever, + seizure, + hospital, + minute, + admit, + last, + have, + hour, + call, + child	575	0.045479712...	0.1189353...
3	+ vaccination, male, + include, + old, + vaccine, + receive, + history, + female, information, + virus, + year, concomitant	1710	0.135252709...	0.1298290...
4	+ week, + headache, + vomit, + fever, + cry, + hour, + symptom, + feel, + shoot, + start, + have, + day	2229	0.176303092...	0.1443532...
5	+ call, er, + itch, + body, face, benadryl, + give, + see, + start, + rash, + fever, + state	1666	0.131772522...	0.1061495...
6	+ find, dead, + bed, + feel, + have, + mother, + child, er, + patient, pt, + state, + shoot	135	0.010677845...	0.0868399...
7	+ arm, + deltoid, right, + leave, + pain, + female, + injection, + year, information, + dose, + site, + develop	1502	0.118800917...	0.1178260...
8	+ swell, + deltoid, right, redness, + thigh, + injection site, + injection, + site, + red, + arm, erythema, + touch	3495	0.276437554...	0.1140865...

5 Select a term that is related to an adverse reaction that you want to investigate further. In particular, select **fever** under the TERM column of the Terms window. Right-click on the term and select **Filter Terms**.

TERM	Freq	# Docs	Keep	WEIGHT	Role
have	6339	4154	<input checked="" type="checkbox"/>	0.084	
receive	5191	3690	<input checked="" type="checkbox"/>	0.059	
vaccine	4628	3430	<input checked="" type="checkbox"/>	0.011	
swell	3945	3271	<input checked="" type="checkbox"/>	0.221	
day	3896	2999	<input checked="" type="checkbox"/>	0.011	
arm	3956	2880	<input checked="" type="checkbox"/>	0.195	
fever			<input checked="" type="checkbox"/>	0.022	
leave			<input checked="" type="checkbox"/>	0.183	
pt			<input checked="" type="checkbox"/>	0.045	
site			<input checked="" type="checkbox"/>	0.244	
injection			<input checked="" type="checkbox"/>	0.204	
report			<input checked="" type="checkbox"/>	0.045	
patient			<input checked="" type="checkbox"/>	0.118	
develop			<input checked="" type="checkbox"/>	0.022	
give			<input checked="" type="checkbox"/>	0.078	

- Filter Terms
- Filter Where Terms
- Filter (Keep the selected records)
- Find Similar Terms
- Clear Selection
- Treat as Equivalent Terms
- Remove Synonyms
- Toggle KEEP
- Find
- Repeat Find
- View Concept Links

- 6 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 right side of the toolbar.

Text Miner - Interactive

File Edit Tools View Window

Documents


SYMPTOM_TEXT ▲	COSTRING	Anthrax	Diphtheri...
Fever +100 redness of skin, slightly swollen, no pain. Pt is ok. Come by on 10/7/04 all ok.	EDEMA FEVER VAS...	0.0	0.0
Fever - temperature to 104. 4 degrees, decreased with Tylenol and Motrin.	FEVER	0.0	0.0
Fever /Headache started 24 hours after vaccine: then nasal congestion. Cough started 7-8 days after vaccine, presents today with wheezing. No fever. No n/o wheeze, pneumonia, bronchitis. Treated with Albuteral Pulmicort.	ASTHMA COUGH IN...	1.0	0.0
Fever 01/10/2005, ER 01/12/2005. Sepsis requiring hospitalization for 2 days.	COUGH INC DIARRH...	0.0	1.0
Fever 08/01/2003 without any other symptoms.	FEVER	0.0	1.0
Fever 100 - 101, local edema left thigh measured 13" acute redness 15 cm x 14 cm. Tender to touch. PT very fussy, mom administering Tylenol infant gets 4 hrs and pushing PO fluids.	AGITATION EDEMA I...	0.0	1.0

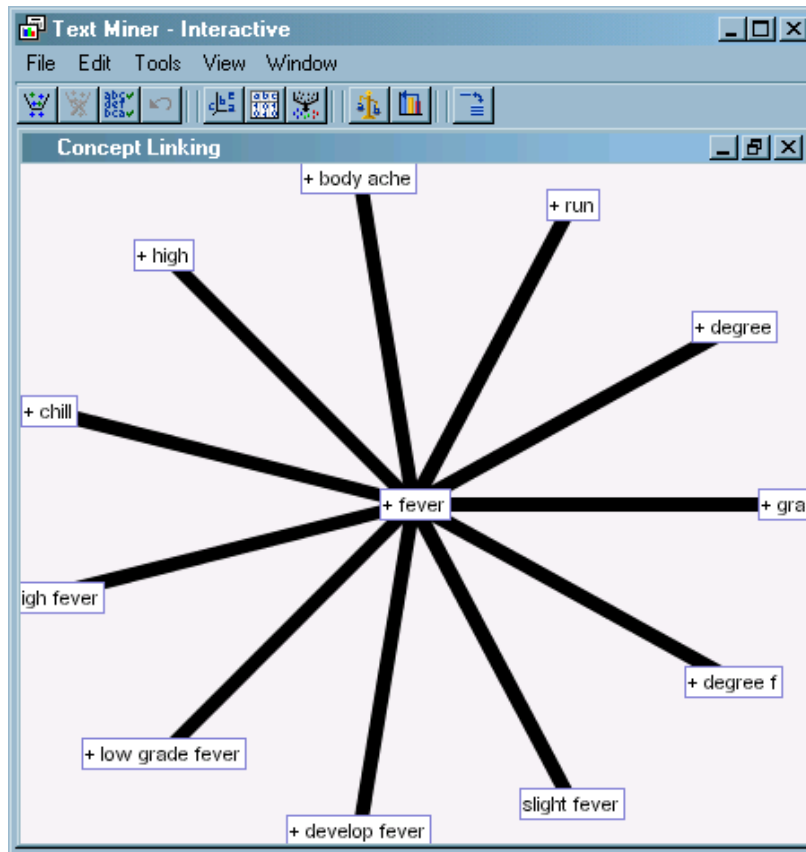
Terms

TERM	Freq	# Docs	Keep ▼	WEIGHT
fever	2951	2477	<input checked="" type="checkbox"/>	0.02

Clusters

#	Descriptive Terms	Freq ▲	Percen
2	er, febrile, + fever, + seizure, + hospital, + minute, + admit, + last, + have, + hour, + call, + child	173	0.0698
3	+ vaccination, male, + include, + old, + vaccine, + receive, + history, + female, information, + virus, + year, concomitant	239	0.0964
7	+ arm, + deltoid, right, + leave, + pain, + female, + injection, + year, information, + dose, + site, + develop	249	0.1005
1	+ admit, + month, er, + hospitalize, + seizure, + discharge, + hospital, + diagnosis, + state, + day, + have, + patient	277	0.1118


- 7 Click the Undo icon  on the toolbar. This removes the filter that was applied and restores the display that was shown when you opened the Interactive Results window.
- 8 Right-click on **fever** in the Terms window, 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 strongly associated with the term **fever**. In order 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?

# ▲	Descriptive Terms	Freq	Percentage	RMS Std.
1	+ admit, + month, er, + hospitalize, + seizure, + discharge, + hospital, + diagnosis, + state, + day, + have, + patient	1029	0.081388910...	0.143384
2	er, febrile, + fever, + seizure, + hospital, + minute, + admit, + last, + have, + hour, + call, + child	575	0.045479712...	0.118935
3	+ vaccination, male, + include, + old, + vaccine, + receive, + history, + female, + formation, + virus, + year, concomitant	129829		
4	+ week, + headache, + vomit, + fever, + cry, + hour, + symptom, + feel, + shock, + start, + have, + day	144353		

- Filter Clusters
- Filter (Keep the selected records)
- Find Similar Clusters
- Clear Selection
- View Hierarchical Clusters

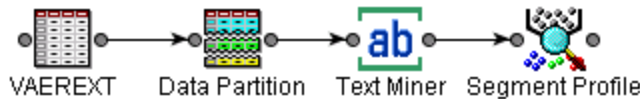
- 11 Click the Undo icon  on the toolbar to undo any filters.
- 12 Close the Interactive Results window.

Examine Data Segments

In this section, you will examine segmented or clustered data using the Segment Profile node. A *segment* is a cluster number 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:

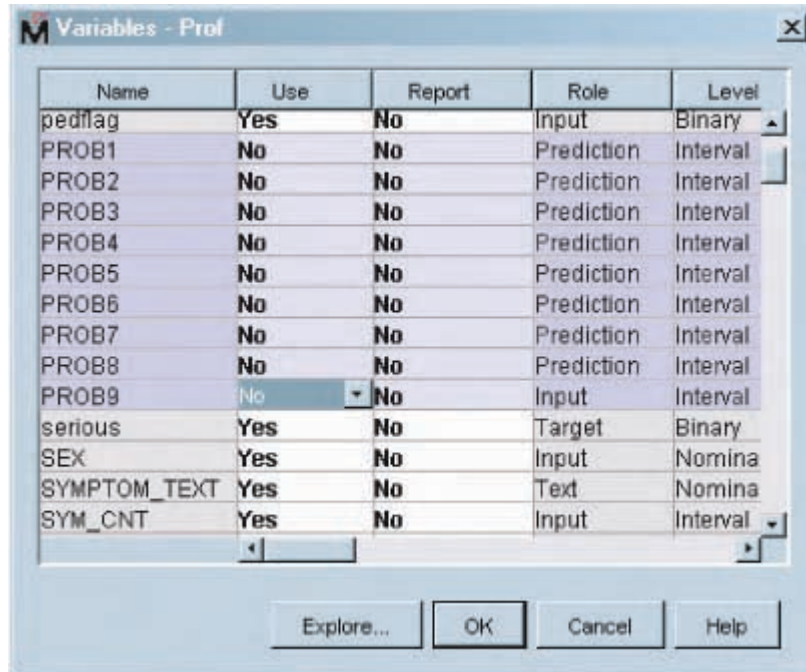
- 1 From the **Assess** tab, drag and drop the Segment Profile node into the diagram workspace and connect it to the Text Miner node.



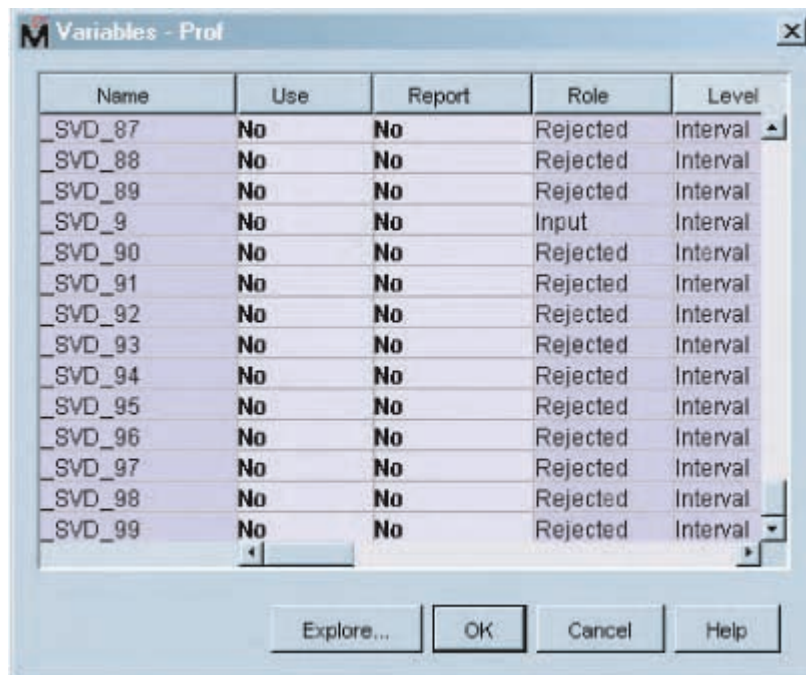
- 2 Select the Segment Profile node. On the Properties panel, select the ellipsis button to the right of the **Variables** property. The Variables — Prof window opens.

Property	Value
Node ID	Prof
Imported Data	...
Exported Data	...
Variables	...

- 3 Select all the PROB variables and set their **Use** to **No**.



- 4 Select all the _SVD_variables and also set their **Use** to **No**.

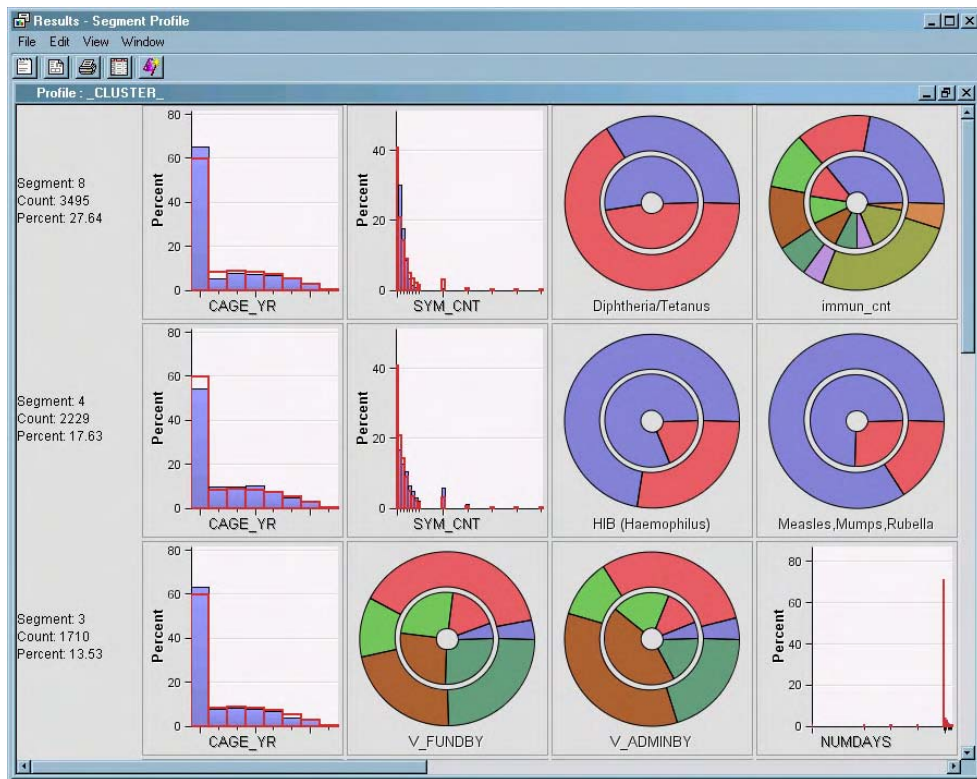


- 5 Click **ok**.

- 6 Select the Segment Profile node in the diagram workspace. In the Properties panel, set **Minimum Worth** to **0.0010**.

Property	Value
Input Variables	
Number of Inputs	10
Minimum Worth	0.0010
Print Worth Statistics	Yes

- 7 Run the Segment Profile node.
 8 After the node finishes running, click **Results** in the Run Status window.
 9 Maximize the Profile: `_CLUSTER_` window.



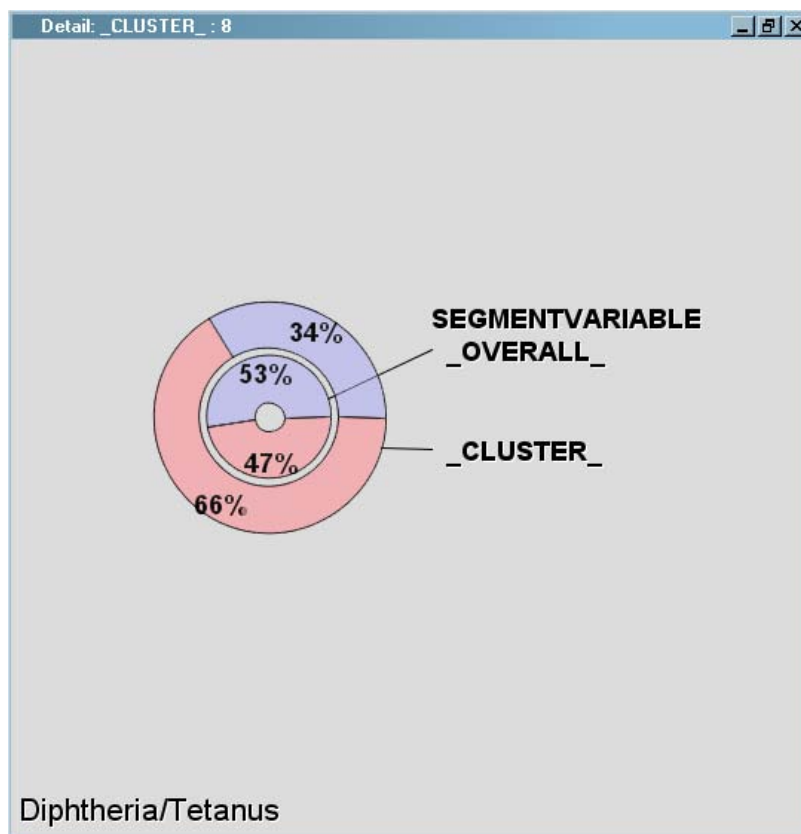
The Results — Segment Profile window displays a lattice, or grid, of plots comparing the distribution for the identified and report variables for both the segment and the population. The graphs shown in this window are variables that have been identified as factors that distinguish the segment from the population it represents. Each row represents a single segment. The far-left margin identifies the segment, its count, and the percentage of the total population. By default, the rows are sorted in ascending size order from top to bottom.

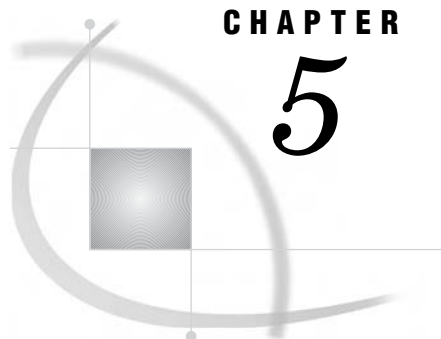
The columns are organized from left to right according to their ability to discriminate that segment from the population. Report variables, if specified, appear on the right in alphabetical order after the selected inputs. The lattice graph has the following features:

- 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 In the example, note the strong relationships between some of the vaccinations given and the clustered categories (e.g., Diphtheria/Tetanus and Segment 8 or HIB and Segment 4). 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 ones in that cluster. Position your mouse over an area, and it displays the value associated with that color. Double-click the top cluster labeled Diphtheria/Tetanus. The cluster occurs with Diphtheria/Tetanus about 66 percent of the time as compared to only about 47 percent of the time in the total population.





CHAPTER

5

Cleaning Up Text

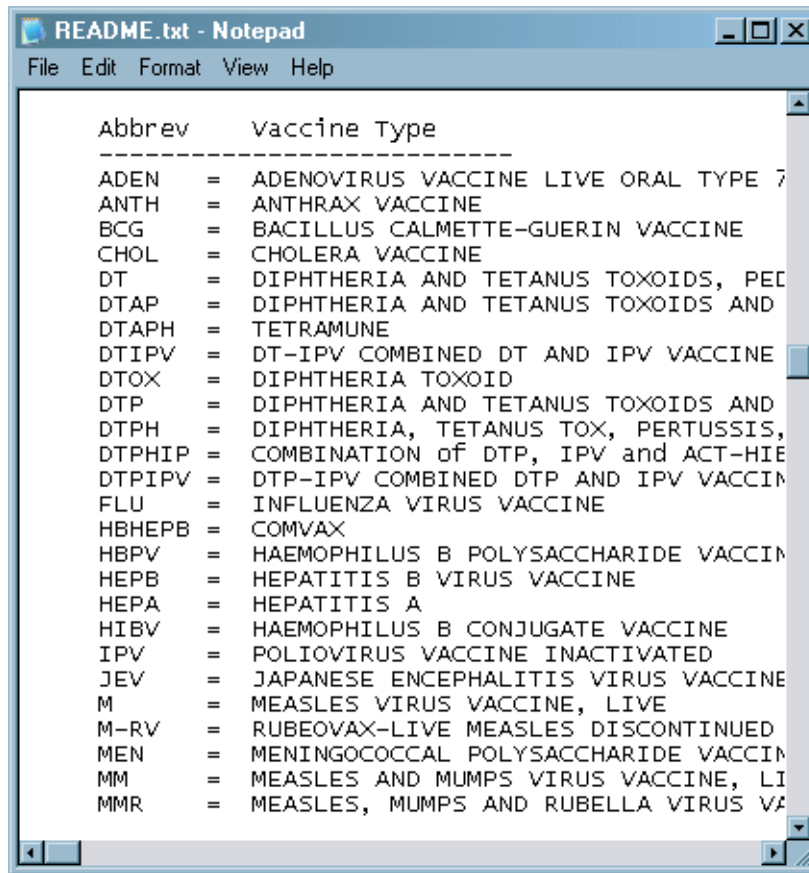
<i>About the Tasks that You Will Perform</i>	33
<i>Use a Synonym Data Set</i>	35
<i>Create a New Synonym Data Set</i>	37
<i>Examine Results Using Merged Synonym Data Sets</i>	40
<i>Create a Stop List</i>	43
<i>Explore Result Improvements</i>	46

About the Tasks that You Will Perform

Some clusters in the previous chapter's results probably seemed pretty logical and dealt with an interesting theme. But some clusters were probably duplicates of others, and some might have seemed a hodgepodge of different things. SAS Text Miner does a good job of finding themes that are clear in the data. When the data needs cleaning, SAS Text Miner is not as effective at uncovering useful themes. Analyzing messy data that needs cleaning is more common than analyzing clean data. In this example, you will deal with manually edited data with many misspellings and abbreviations. In this chapter, you will work on cleaning up the data to get better results.

The README.TXT file provided on the VAERS site contains a list of commonly used abbreviations in the adverse event reports. SAS Text Miner allows you to specify a synonym list for use in the tool. A VAER_ABBREV synonym list is provided for you in the Getting Started with SAS Text Miner 3.1 zip file. To create such a synonym list, the abbreviations list from README.TXT was copied into MS Excel. The list was manually edited there 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 Appendix 2, "Vaccine Adverse Event Reporting System Data Preprocessing," on page 71.

For more information about importing data into a SAS data set, see Importing and Exporting Data in the following URL: <http://support.sas.com/onlinedoc/913/docMainpage.jsp>.



The screenshot shows a spreadsheet window with the following data:

	TERM	PARENT	CATEGORY
1	a/	before	...
2	abd	abdomen	...
3	abn	abnormal	...
4	addl	additional	...
5	adm	admitted	...
6	admission	admitted	...
7	ADR	adverse drug reaction	...
8	ADL	activities of daily living	...
9	AE	Adverse Event	...
10	afeb	afebrile	...
11	AKA	also known as	...
12	alk phos	alkaline phosphatase	...
13	allerg	allergy	...
14	allergic	allergy	...
15	ALT	serum glutamic pyruvic transamina...	...
16	AMA	American Medical Association	...
17	amb	ambulate	...
18	ambulance	ambulate	...
19	amt	amount	...
20	ANA	antinuclear antibody	...

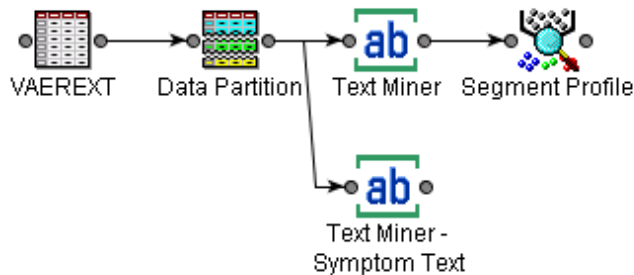
You will perform the following tasks to clean up the text and examine the results:

- 1 Use a synonym data set provided for you in the Getting Started with SAS Text Miner 3.1 zip file.
- 2 Create a new synonym data set using the SAS Code node and the %TEXTSYN macro. The %TEXTSYN macro will run through all the terms, automatically identify which ones are misspellings, and create synonyms mapping them 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 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.
- 5 Explore whether cleaning up the text improved the clustering results.

Use a Synonym Data Set

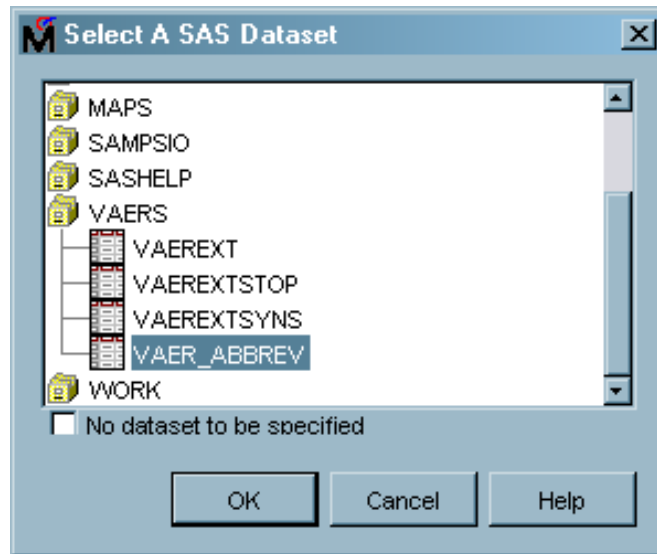
To use a synonym data set, complete the following steps:

- 1 Right-click the original Text Miner node and select **Copy**. 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 on it, select **Rename** and name it **Text Miner - Symptom Text**. Click **OK** in the Rename window.
- 3 Connect the Data Partition node to the new Text Miner - Symptom Text node.



- 4 Select the Text Miner - Symptom Text node in the diagram workspace. In the Properties panel, click the ellipsis to the right of **Synonyms**. The Select A SAS Dataset window opens.

- 5 Double-click the **VAERS** library to expand it, select **VAER_ABBREV**, and click **OK**.



- 6 Leave all other settings the same as in the original Text Miner node.

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

- 7 Select the Text Miner - Symptom Text node in the diagram workspace and run the node.
- 8 Click the ellipsis to the right of the Interactive property to open the Interactive Results 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 hand side of the VAERS.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. This shows 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. SAS Text Miner is now treating all these terms the same.

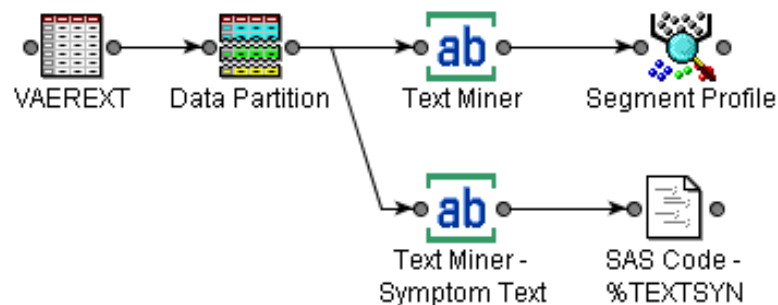
TERM ▲	Freq	# Docs	Keep	WEIGHT	Role
abscesses	1	1	<input checked="" type="checkbox"/>	1.238	
abd pain	1	1	<input checked="" type="checkbox"/>	0.342	NOUN_GROUP
abdomen	117	111	<input checked="" type="checkbox"/>	0.144	
abdomen	100	94			
abd	17	17			
abdomen area	1	1	<input checked="" type="checkbox"/>	0.342	NOUN_GROUP
abdomen then move	1	1	<input checked="" type="checkbox"/>	0.342	NOUN_GROUP
abdomen-laceration	1	1	<input checked="" type="checkbox"/>	1.238	
abdomen-on	1	1	<input checked="" type="checkbox"/>	0.342	
abdomen-on 8 da...	1	1	<input checked="" type="checkbox"/>	0.342	NOUN_GROUP
abdomen-red	1	1	<input checked="" type="checkbox"/>	0.342	
abdomen-red dot	1	1	<input checked="" type="checkbox"/>	0.342	NOUN_GROUP
abdomen/neck	1	1	<input checked="" type="checkbox"/>	0.342	
abdomen/neck flat	1	1	<input checked="" type="checkbox"/>	0.342	NOUN_GROUP

11 Close the Interactive Results window.

Create a New Synonym Data Set

You will use the SAS Text Miner `%TEXTSYN` macro to create a new synonym data set. The macro will run through all the terms, automatically identify which ones are misspellings, and create synonyms mapping them to the misspelled term. To create a new synonym data set, complete the following steps:

- 1 Select the **Utility** tab and drag a SAS Code node into the diagram workspace. Connect this node to the Text Miner - Symptom Text node. Rename the SAS Code node to **SAS Code - %TEXTSYN**.



- 2 Select the arrow that connects the Text Miner - Symptom Text node with the SAS Code - `%TEXTSYN` 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 libname `EMWS` in the Terms export **Table** property is dependent upon the diagram number within your Enterprise Miner project. If your diagram is the first one created, the libname will be `EMWS`, the second diagram will be `EMWS1`, the third will be `EMWS2`, and so on. △

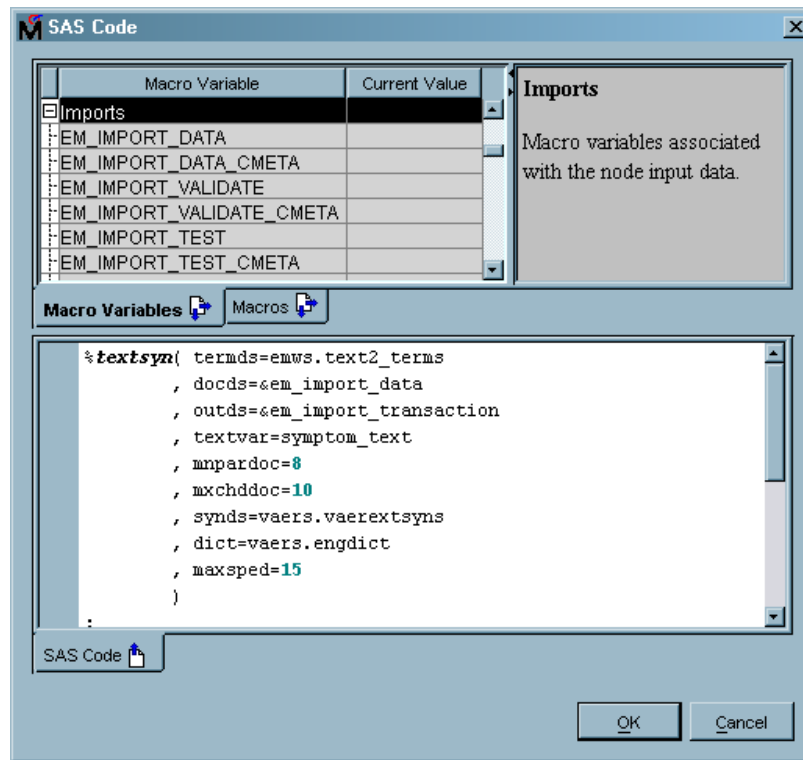
Property	Value
From	TEXT2
To	EMCODE
DOCUMENTS export	
Table	EMWS.TEXT2_DOCUMENTS ...
Variables	...
Role	Train
Validate export	
Table	EMWS.TEXT2_VALIDATE ...
Variables	...
Role	Validate
Test export	
Table	EMWS.TEXT2_TEST ...
Variables	...
Role	Test
Terms export	
Table	EMWS.TEXT2_TERMS ...
Variables	...
Role	Terms
CLUSTER export	
Table	EMWS.TEXT2_CLUSTER ...
Variables	...
Role	Cluster
OUT export	
Table	EMWS.TEXT2_OUT ...
Variables	...
Role	Transaction

- 3 Select the SAS Code - %TEXTSYN node and click the ellipsis to the right of the **SAS Code** property in the Properties panel.
- 4 Enter the following code in the **SAS Code** tab:

```

%textsyn( termds=emws.text2_terms
, docds=&em_import_data
, outds=&em_import_transaction
, textvar=symptom_text
, mnpardoc=8
, mxchddoc=10
, synds=vaers.vaerextsyns
, dict=vaers.engdict
, maxsped=15
) ;

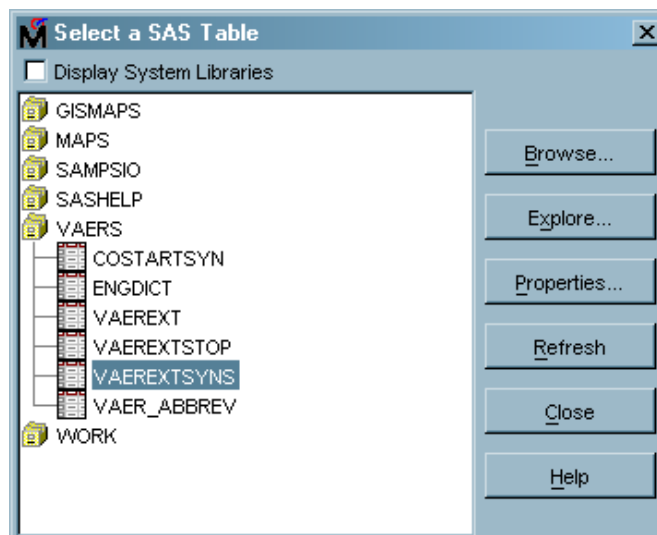
```



Note: For details on the %TEXTSYN macro, see SAS Text Miner online documentation. △

- 5 Click **OK** and run the SAS Code - %TEXTSYN node.
- 6 From the Enterprise Miner window, select **View ► Table**. The Select a SAS Table window opens.
- 7 Double-click **VAERS**, select **VAEREXTSYNS**, and click **Explore**. The Explore window opens.

Note: If the **VAERS** library is already expanded, you might need to click **Refresh** to see the **VAEREXTSYNS** data set. △



8 Examine the VAERS.VAEREXTSYNS table.

Obs #	EXAMPLE1	EXAMPLE2	Term	PARENT	CATEGORY	CHILDNDOCS	# Documents	MINSPEL	DICT
61	... with rest, alternating Tylenol !!650...	... thigh, redb...	650mg	... 60mg		2	9	12N	
62	... tabs and Depo Medrol !!80mg!! at ...		80mg	... 800mg		1	8	6N	
63	Immunizatio given at !!8:30a!l. m. o...		8:30a	... 8:30am		1	8	6N	
64	... cures done (no !!abcess!)	... for fever. ...	abcess	... abscess		8	43	8N	
65	... the rash (upper !!abdoment!!) w...		abdoment	... abdomen		1	94	4N	
66	... 20 mins later, upper !!abdomin!! an...	Does have...	abdomin	... abdominal		3	84	9N	
67	... 1ml of vaccine (!!Accidental ove...		accid	... acid		1	16	6N	
68	... received an injection of !!Acel-Im...		acel-immune	... acel-immune		1	18	2N	
69	... received series of Hib-Titer, !!Ace...		acel-immun	... acel-immune		1	18	3N	
70	... child received Hib-TITER, Tetramu...		acel0immune	... acel-immune		1	18	10N	
71	... , diphtheria -tetanus and !!acellua...		acelluar	... acellular		1	82	6N	
72	... cool wet compresses and !!aceta...		acetaminsph	... acetamino...		1	43	7N	
73	... continued to feel " !!achey!!" all ov...	... jawache, ...	achey	... ache		3	82	8N	
74	... the first dose of !!Act-H!!		act-h	... act-hib		1	10	14N	
75	... doses of Prevnar and !!Act-Hib!! (...	... right lateral...	act-hib	... acthib		10	46	8N	
76	The discharge summary states !!ac...		actue	... acute		1	101	10N	
77	... belly then she could !!acutally!! se...		acutally	... actually		1	11	6N	
78	... that the pt recovered. !!Addi!!		addi	... add		1	16	10N	
79	no 6385704700K or 640327.075		additional	... additional		1	249	5N	

Here is a list of what the VAEREXTSYNS columns provide:

- Term is the misspelled word.
- PARENT is an intelligent 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.
- MINSPEL 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.
- EXAMPLE1 and EXAMPLE2 are two examples of the term in a document.

For example, Observation 66 shows **abdomin** to be a misspelling of **abdominal**. Three documents contain **abdomin**, 84 documents contain the parent, **abdomin** is not a legitimate English word, and an example text that contains that misspelling is **20 mins later, upper !!abdomin!! an...** Note that double exclamation marks (!!) both precede and succeed the child term in the example text so you can see the term in context.

- 9 Examine this table to see if you disagree with some of the choices made. For this example, however, assume the %TEXTSYN macro has done a good enough job detecting misspellings.

Note: The table could be edited using any SAS table editor. You cannot edit this table in the SAS Enterprise Miner Java client. You could change a parent for any misspellings that appear incorrect or delete a row if the Term column contains a valid term. Make sure to save any changes you made. \triangle

- 10 Close the VAERS.VAEREXTSYNS table and the Select a SAS Table window.

Examine Results Using Merged Synonym Data Sets

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

- 1 Select the **Utility** tab and drag a SAS Code node into the diagram workspace. Connect this new SAS Code node to the existing SAS Code - %TEXTSYN node. Rename the new SAS Code node to SAS Code - Merge Synonym Lists.

- 2 Select the SAS Code - Merge Synonym Lists node and click the ellipsis to the right of the **SAS Code** property in the Properties panel.
- 3 Enter the following code in the **SAS Code** tab:

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

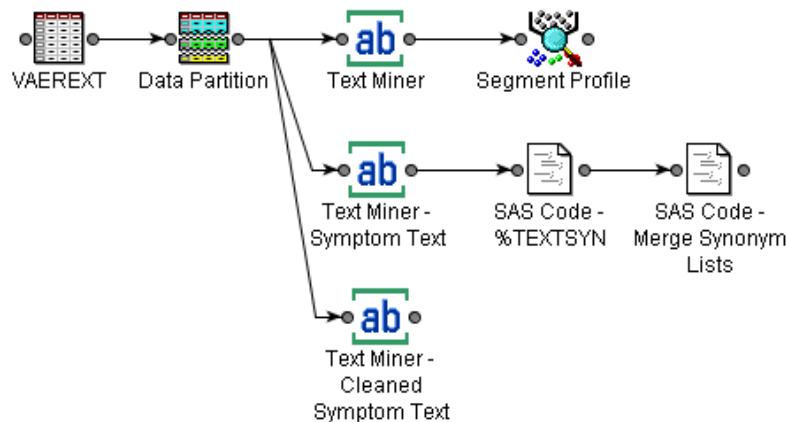
This code merges the resulting synonyms data set from the first SAS Code - %TEXTSYN node with the abbreviations data set.

- 4 Run the SAS Code — Merge Synonym Lists node. Click **Results** in the Run Status window.
- 5 From the Results window, select **View ► SAS Results ► Log**. The following display shows the SAS code where the new data set is created.

The screenshot shows the SAS Log window titled "Results - SAS Code - Merge synonym lists". The log contains the following text:

```
75 4419 * EMCODE2: User Code;
76 4420 *-----*;
77 4421 data vaers.vaerextsyms_new;
78 4422 set vaers.vaerextsyms vaers.vaer_abbrev;
79 4423 run;
80
81 NOTE: There were 1696 observations read from the data set VAERS.VAEREXTSYMS.
82 NOTE: There were 319 observations read from the data set VAERS.VAER_ABBREV.
83 NOTE: The data set VAERS.VAEREXTSYMS_NEW has 2015 observations and 9 variables.
84 NOTE: DATA statement used (Total process time):
85     real time          0.04 seconds
86     cpu time           0.03 seconds
```

- 6 Copy and paste the Text Miner - Symptom Text node and rename it Text Miner - Cleaned Symptom Text.
- 7 Connect it to the Data Partition node.



8 Select the Text Miner - Cleaned Symptom Text node. Set the following properties in the Properties panel:

- Click the ellipsis to the right of the **Synonyms** property. Select **VAERS.VAEREXTSYN_NEW** from the Select A SAS Dataset window.
- Set **Terms in a Single Document** back to **No**.

Parse	
Parse Variable	SYMPTOM_TEXT
Language	ENGLISH
Stop List	SASHELP.STOPLST ...
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	VAERS.VAEREXTSYNS_NEW ...
Find Entities	No
Types of Entities	...

9 Run the Text Miner - Cleaned Symptom Text node.

10 Click **OK** in the Run Status window.

11 Click the ellipsis to the right of the **Interactive** property in the Text Miner - Cleaned Symptom Text node Properties panel. The Interactive Results window opens.

12 Select **+patient** in the Terms table. Note that the misspellings **patien**, **patienn**, and **patie** are included as child terms.

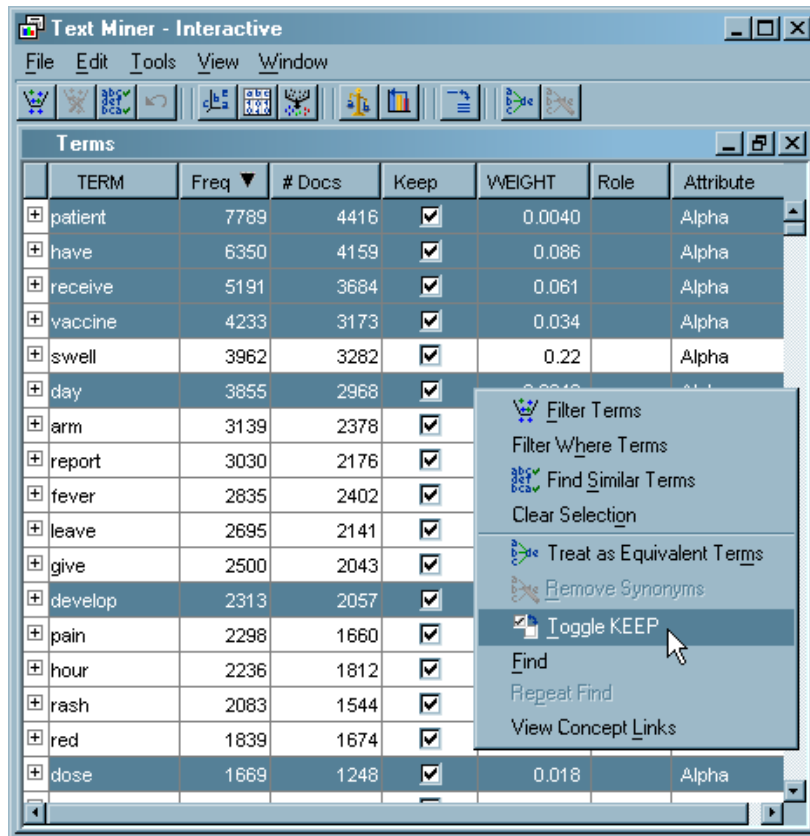
Terms			
	TERM \blacktriangle	Freq	# Docs
<input checked="" type="checkbox"/>	patient	7789	4416
<input type="checkbox"/>	pt	4239	2438
<input type="checkbox"/>	patient	3436	2043
<input type="checkbox"/>	patients	70	61
<input type="checkbox"/>	pts	27	27
<input type="checkbox"/>	patien	9	9
<input type="checkbox"/>	pts.	1	1
<input type="checkbox"/>	ppts	1	1
<input type="checkbox"/>	ppt	1	1
<input type="checkbox"/>	patinet	1	1
<input type="checkbox"/>	patienn	2	2
<input type="checkbox"/>	patie	2	2

Create a Stop List

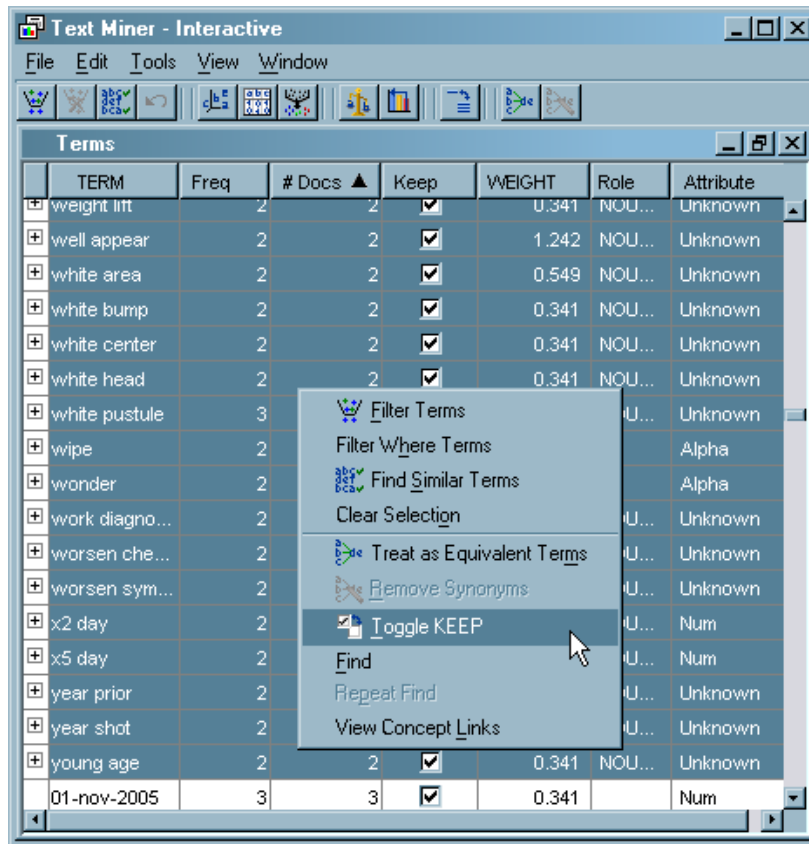
A stop list is a simple collection of low-information or extraneous words that you want to remove from the text, which has been saved as a SAS data set. To create a stop list, complete the following steps:

- 1 Click the Freq column heading to sort the Terms table by the frequency. Make sure that the Freq label has an arrow that points downward to indicate that the Freq column is sorted in descending order.
- 2 Drop some terms that really 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 invoke the pop-up menu. Select **Toggle KEEP** to uncheck the **Keep** attribute. This removes the checkmark from the Keep column for each term you have selected.

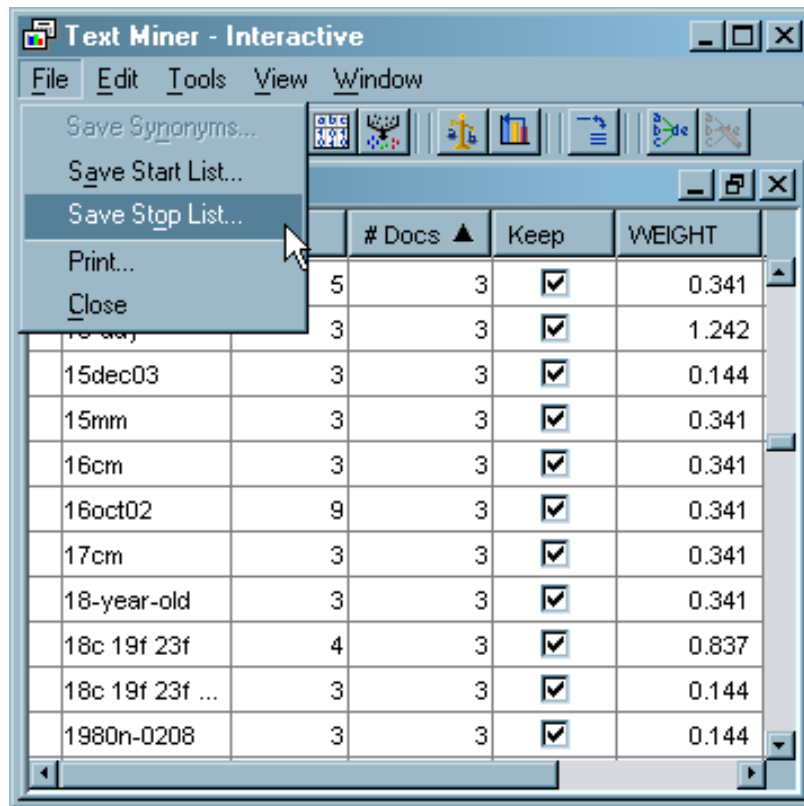
There are several more terms you could choose to exclude. Only a few are itemized here to demonstrate the concept and process. If additional terms are dropped from the analysis, note that different results will be obtained that will not match those later in this document.



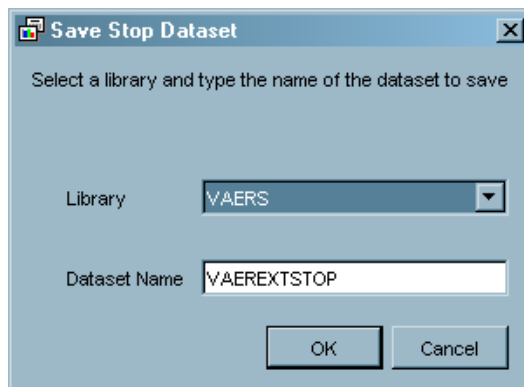
- 3 Double-click the # Docs column heading and ensure that the sort arrow is pointing upward. This sorts the terms by count.
- 4 Click and drag the mouse to select all terms with counts of 2. Right-click and select **Toggle KEEP** so that these terms are dropped from the analysis.



- 5 Select **File ► Save Stop List**.



- 6 Select the VAERS library and type VAEREXTSTOP in the **Dataset Name** box.



- 7 Click **OK**.
- 8 Close the Text Miner — Interactive Results window.

- 9 Note that the **Stop List** property of the Text Miner - Cleaned Symptom Text node is set to **VAERS.VAEREXTSTOP**.

Parse	
Parse Variable	SYMPTOM_TEXT
Language	ENGLISH
Stop List	VAERS.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	VAERS.VAEREXTSYNS_NEW
Find Entities	No
Types of Entities	

Explore Result Improvements

You will redo the 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 - Cleaned Symptom Text node are the same as in previous examples.

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

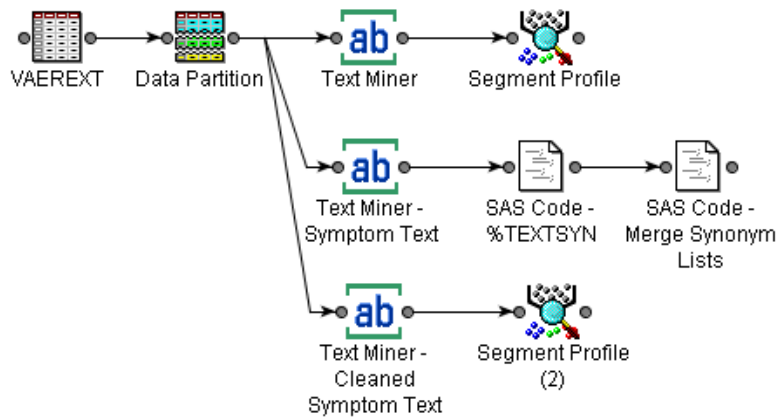
- 2 Run the Text Miner - Cleaned Symptom Text node to redo the clustering.

3 Open the Interactive Results window and look at the Clusters table.

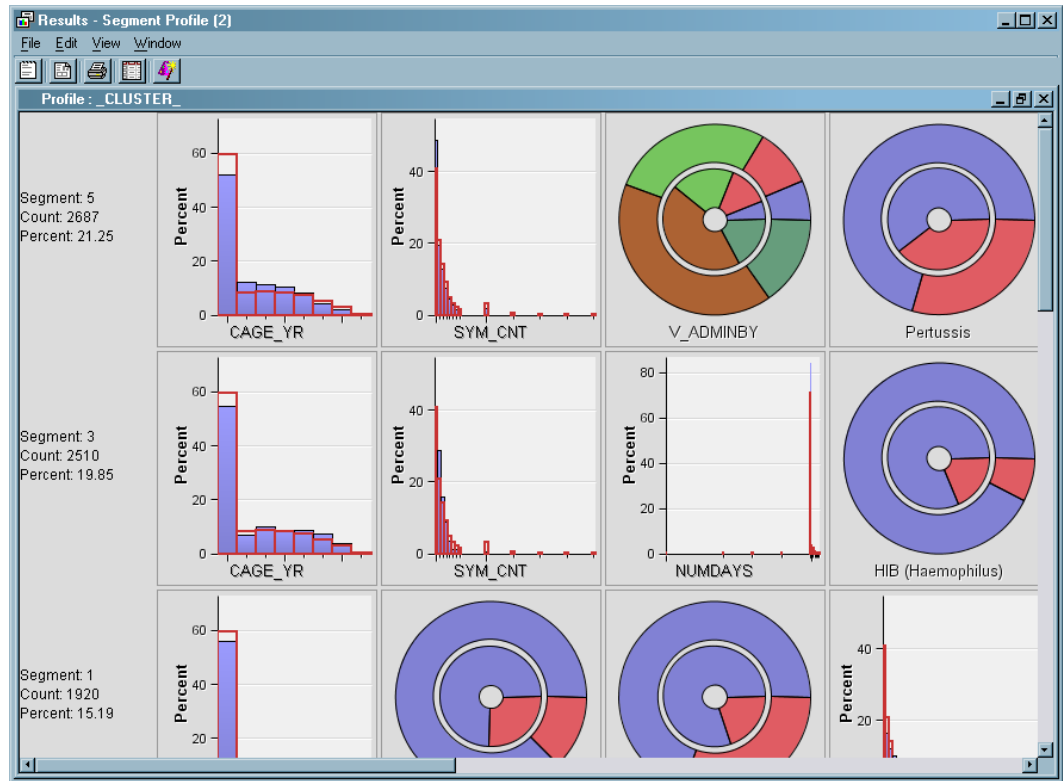
Clusters				
#	Descriptive Terms	Freq	Percentage	RMS Std.
1	+ episode, + week, + old, + year, + report, + minute, + experience, + female, + symptom, + feel, + emergency room, + time	1920	0.151862690...	0.1536012...
2	+ infant, + diagnosis, loss, + diagnose, + month, + late, + child, + vaccination, + symptom, + experience, + see, + information	976	0.077196867...	0.1503808...
3	+ elbow, injection site, + upper, + swell, + warmth, + itch, + pain, + injection, + delatoid, + arm, + leave arm, + leave	2510	0.198528830...	0.1048913...
4	+ emergency room, + seizure, + discharge, + admit, + hospital, + home, + minute, + temperature, + child, + call, + state, + give	787	0.062247884...	0.1252533...
5	+ emergency room, + benadryl, + feel, + start, + state, + give, + hour, + see, + hive, + fever, + rash, + face	2687	0.212528671...	0.1198837...
6	+ febrile seizure, + minute, + fever, + febrile, + activity, + seizure, + emergency room, + generalize, + second, + last, + immunization, + temperature	239	0.018903741...	0.0389836...
7	+ hospitalize, + information, + antibiotic, + vaccinate, + year, + recover, male, + old, + seizure, + report, + physician, + female	171	0.013525270...	0.0802450...
8	+ physician, + unspecified, medical history, + concomitant, + medical, + old, male, + allergy, + include, + report, + virus, + lot	912	0.072134778...	0.0913942...
9	+ leave, + red, hot, tender, + swell, + delatoid, + touch, + erythema, + size, + right, + upper, + site	1563	0.123625721...	0.1003592...

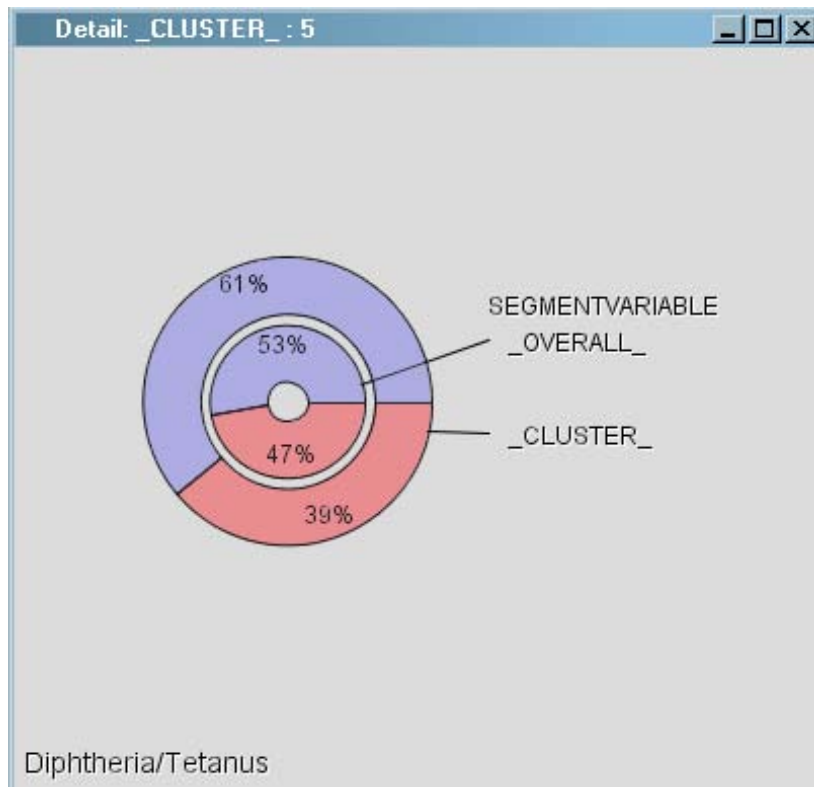
- 4 Compare these results to the first Text Miner node results from Chapter 4, “Analyzing the SYMPTOM_TEXT Variable,” on page 19. Does the clustering seem to have improved with the cleaned SYMPTOM_TEXT data?
- 5 Copy the old Segment Profile node, and paste it next to the Text Miner — Cleaned Symptom Text node.

- 6 Connect the Text Miner — Cleaned Symptom Text node to this Segment Profile (2) node.

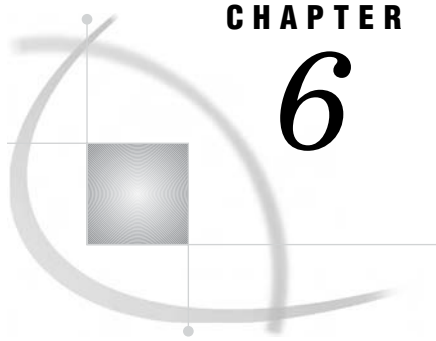


- 7 Click the **Variables** property for the Segment Profile (2) node. Make sure that the PROB variables and the SVD variables are set to **Use=No**.
- 8 Run the Segment Profile (2) node.
- 9 View the resulting profiles. Note the significant relationships in the table. Explore your results and note anything that seems of particular interest to you about the problem you are trying to solve, such as areas you'd like to explore further at a later time.





10 Do the relationships appear clearer with the cleaned text than they did with the uncleaned text?



CHAPTER

6

Predictive Modeling with Text Variables

<i>About the Tasks that You Will Perform</i>	51
<i>Use the COSTRING Variable to Model</i>	51
<i>Use the SYMPTOM_TEXT Variable to Model</i>	58
<i>Compare the Models</i>	61
<i>Additional Exercises</i>	64

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 pull out keywords or synonyms of keywords that could then 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 of your initial tasks was to find what factors influence whether a reaction becomes serious and how well these 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 how serious the adverse event was. This is called predictive modeling, and it is the goal of this chapter.

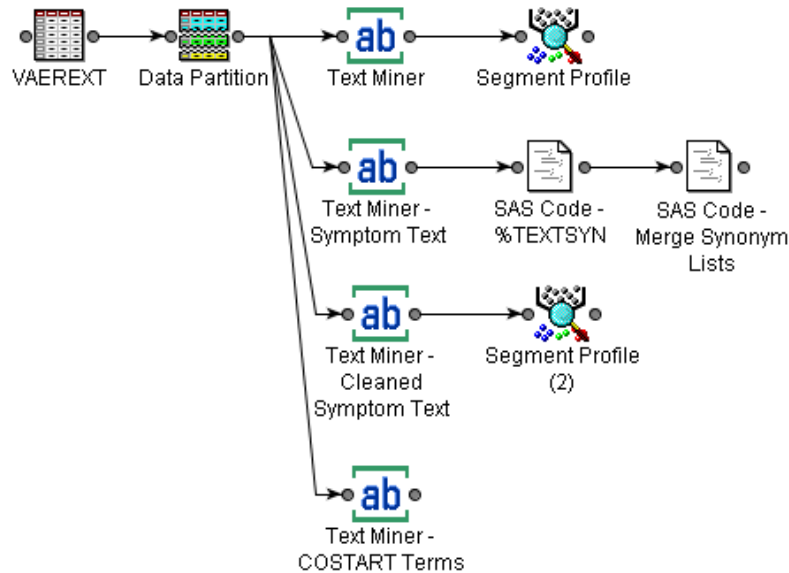
To analyze texts with predictive models, you will perform the following tasks:

- 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

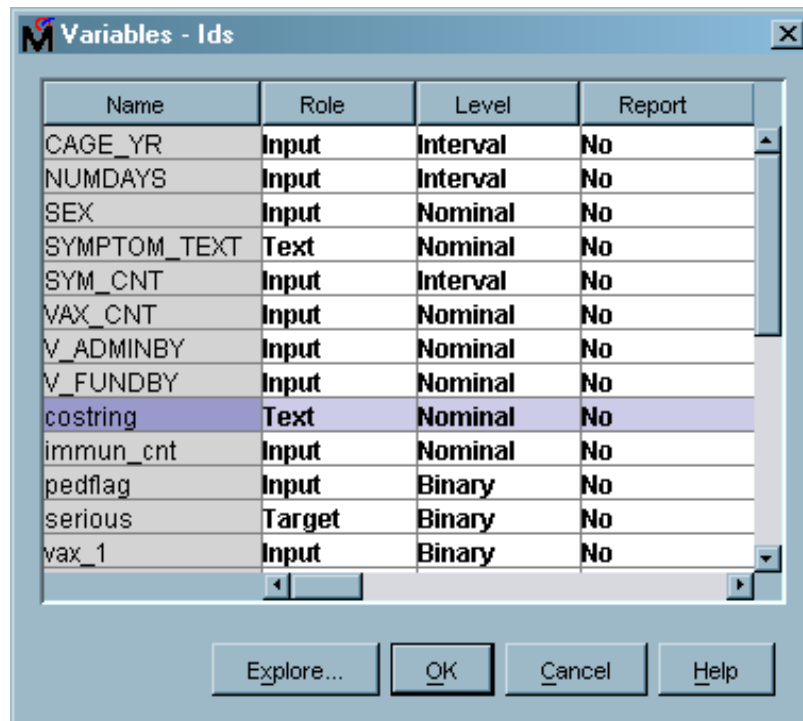
To use the COSTRING variable to create a model, complete the following steps:

- 1 Select the **Explore** tab on the toolbar and drag and drop a Text Miner node into the diagram workspace. Connect it to the Data Partition node.
- 2 To distinguish it from other Text Miner nodes, rename this new node Text Miner - COSTART Terms.



- 3 Select the node named VAEREXT in the diagram workspace. Click the ellipsis to the right of the **Variables** property in the Properties panel.

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.

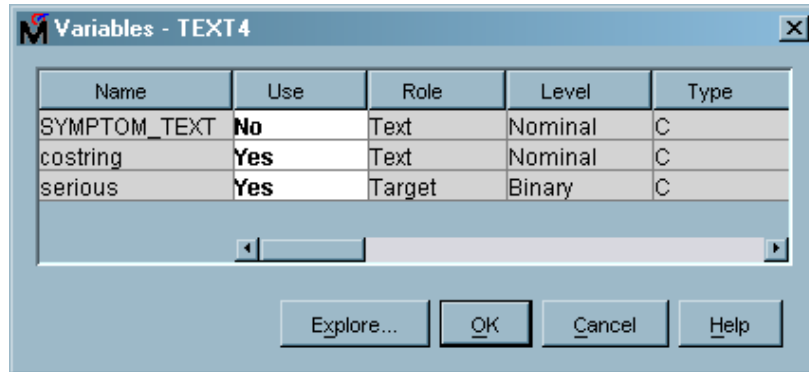


4 Select the Text Miner - COSTART Terms node. Set the following properties in the Properties panel:

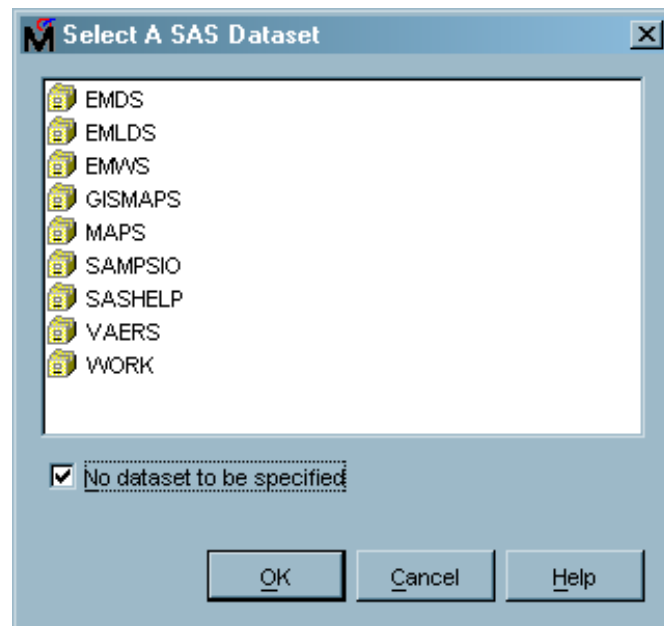
- Click the ellipsis to the right of the **Variables** property. In the Variables window, set **SYMPTOM_TEXT** to **Use=No**, **costring** to **Use=Yes**, and **serious** to **Use=Yes**.

Click **OK** to save your changes.

□



- Click the ellipsis to the right of the **Stop List** property. Select the **No dataset to be specified** check box in the Select A SAS Dataset window. 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 Run the Text Miner - COSTART Terms node.

6 In the Properties panel, make sure that the **Parse Variable** property of the Text Miner - COSTART Terms node is set to **costring**.

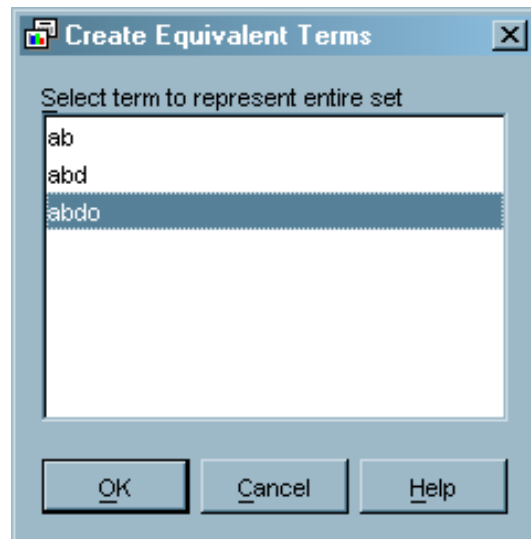
7 Click the ellipsis to the right of the **Interactive** property to open the Interactive Results window. One problem with COSTART is that it does not always use the same keyword to describe the same term or equivalent terms. For example,

abdomen is shown in COSTART as **ab** and as **abdo**. Sometimes there are modifiers that you do not need. You could run the %TEXTSYN macro, but because these are abbreviations, the macro will probably 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. Clean the COSTART data by fixing these differences. Select **ab**, **abd**, and **abdo** from the TERM column. Right-click and select **Treat as Equivalent Terms**.

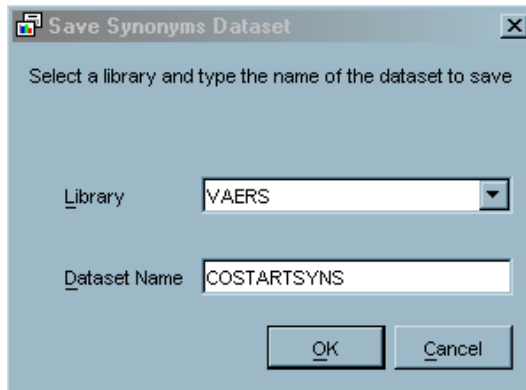
Terms						
TERM ▲	Freq	# Docs	Keep	WEIG...	Role	Attribute
a	8	8	<input checked="" type="checkbox"/>	0.78		Alpha
ab	21	8	<input checked="" type="checkbox"/>	0.673		Alpha
ab pain	3		<input type="checkbox"/>		ROUP	Unknown
abd	2		<input type="checkbox"/>			Alpha
abdo	163		<input type="checkbox"/>			Alpha
abdo pain	6		<input type="checkbox"/>		ROUP	Unknown
abdo pain chest	2		<input type="checkbox"/>		ROUP	Unknown
abdo pain neck	2		<input type="checkbox"/>		ROUP	Unknown
abdo pallor	4		<input type="checkbox"/>		ROUP	Unknown
abdo paresthesia	2		<input type="checkbox"/>		ROUP	Unknown
abdo pharyngitis	2		<input type="checkbox"/>		ROUP	Unknown
abdo pharyngitis r...	2		<input type="checkbox"/>		ROUP	Unknown

Select **abdo** from the Create Equivalent Terms window. 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**.

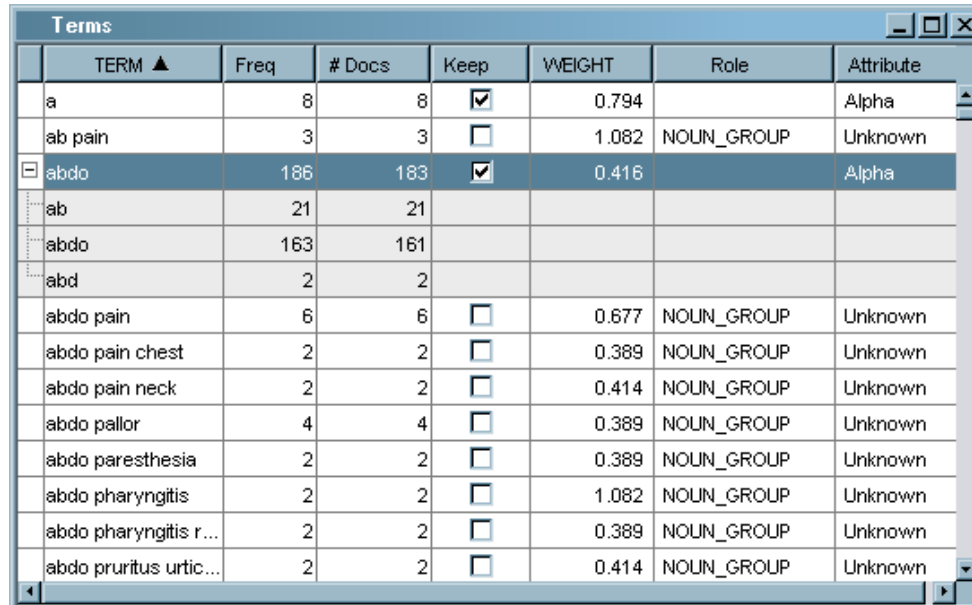
- 9 Repeat this process as many times as you need. It might be helpful to filter the terms so that you can view the full text of COSTART before combining terms.
- 10 Select **File ► Save Synonyms** from the Interactive Results window menu. Save your changes as VAERS.COSTARTSYNS.



- 11 Close the Interactive Results window.
- 12 Note that the **Synonyms** property in the Properties panel has been set to the new VAERS.COSTARTSYNS synonym data set.
- 13 COSTART terms should represent keywords, so you want to create variables for each keyword. Set the following **Transform** properties on 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**.

Property	Value
Parse	
Parse Variable	costring
Language	ENGLISH
Stop List	...
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	VAERS.COSTARTSYNS
Find Entities	No
Types of Entities	...
Transform	
Compute SVD	No
SVD Resolution	Low
Max SVD Dimensions	100
Scale SVD Dimensions	No
Frequency weighting	Log
Term Weight	Mutual Information
Roll up Terms	Yes
No. of Rolled-up Terms	400
Drop Other Terms	Yes
Cluster	
Automatically Cluster	No
Exact or Maximum Number	Maximum
Number of Clusters	40
Cluster Algorithm	EXPECTATION-MAXIMIZATION
Ignore Outliers	No
Hierarchy Levels	.
Descriptive Terms	5
What to Cluster	SVD Dimensions

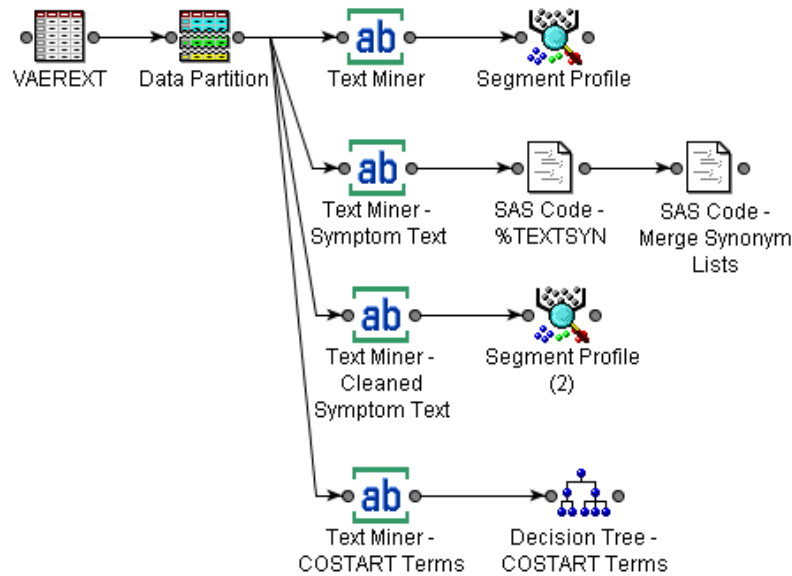
- 14 Rerun the Text Miner - COSTART Terms node using the newly created synonym list.
- 15 Open the Interactive Results window and view the resulting Terms window.
- 16 Sort the TERM column until the arrow on the column heading is pointing down. Note that terms with a plus (+) sign indicate the synonyms you have specified. Click the + sign to expand the child terms underneath the representative parent term.



TERM ▲	Freq	# Docs	Keep	WEIGHT	Role	Attribute
a	8	8	<input checked="" type="checkbox"/>	0.794		Alpha
ab pain	3	3	<input type="checkbox"/>	1.082	NOUN_GROUP	Unknown
abdo	186	183	<input checked="" type="checkbox"/>	0.416		Alpha
ab	21	21				
abdo	163	161				
abd	2	2				
abdo pain	6	6	<input type="checkbox"/>	0.677	NOUN_GROUP	Unknown
abdo pain chest	2	2	<input type="checkbox"/>	0.389	NOUN_GROUP	Unknown
abdo pain neck	2	2	<input type="checkbox"/>	0.414	NOUN_GROUP	Unknown
abdo pallor	4	4	<input type="checkbox"/>	0.389	NOUN_GROUP	Unknown
abdo paresthesia	2	2	<input type="checkbox"/>	0.389	NOUN_GROUP	Unknown
abdo pharyngitis	2	2	<input type="checkbox"/>	1.082	NOUN_GROUP	Unknown
abdo pharyngitis r...	2	2	<input type="checkbox"/>	0.389	NOUN_GROUP	Unknown
abdo pruritus urtic...	2	2	<input type="checkbox"/>	0.414	NOUN_GROUP	Unknown

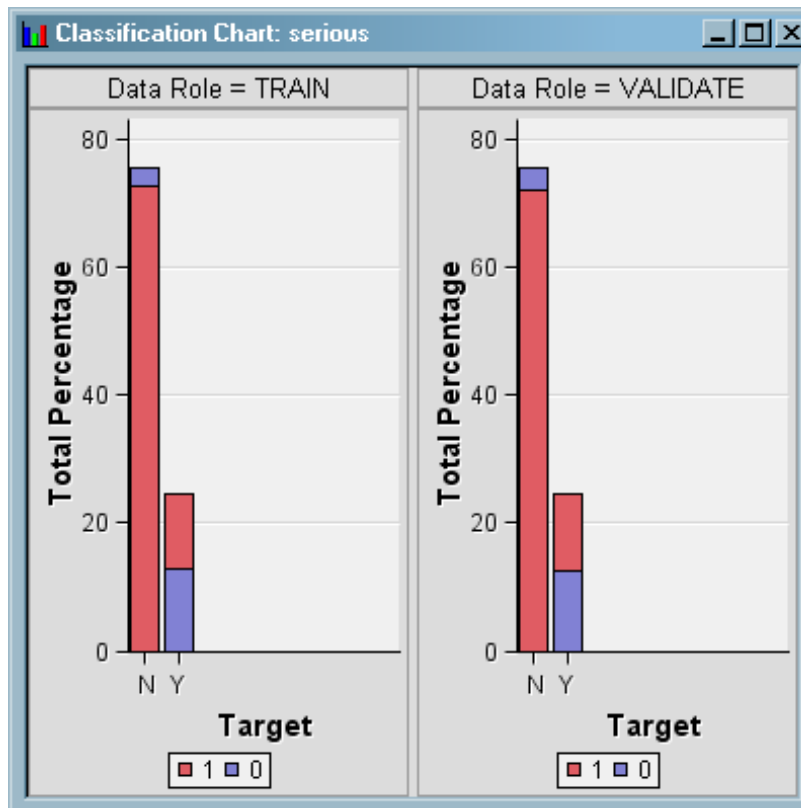
- 17 Scroll down until you see terms that do not have a checkmark beneath the Keep column. A separate variable will not be created for these terms. They were not considered significant enough (based on only rolling up 400 variables) to create a separate variable. Recall that you set the **Roll up Terms** property to **Yes** and **No. of Rolled-up Terms** 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 on the document collection.
- 18 Close the Interactive Results window.

- 19 From the **Model** tab, drag and drop a Decision Tree node and connect it to the Text Miner - COSTART Terms node. Later on you will use another Decision Tree node. To distinguish them, rename this node Decision Tree - COSTART Terms.



- 20 Run the Decision Tree - COSTART Terms node with the default settings. Click **Yes** in the Confirmation window. Recall that when you created the VAEREXT data set, you set **serious** as the target variable.
- 21 When the node has run, click **Results** in the Run Status window to explore the results.
- 22 Select **View** \blacktriangleright **Assessment** \blacktriangleright **Classification Chart: serious** from the menu at the top of the Results - Decision Tree - COSTART Terms window to view the Classification Chart.

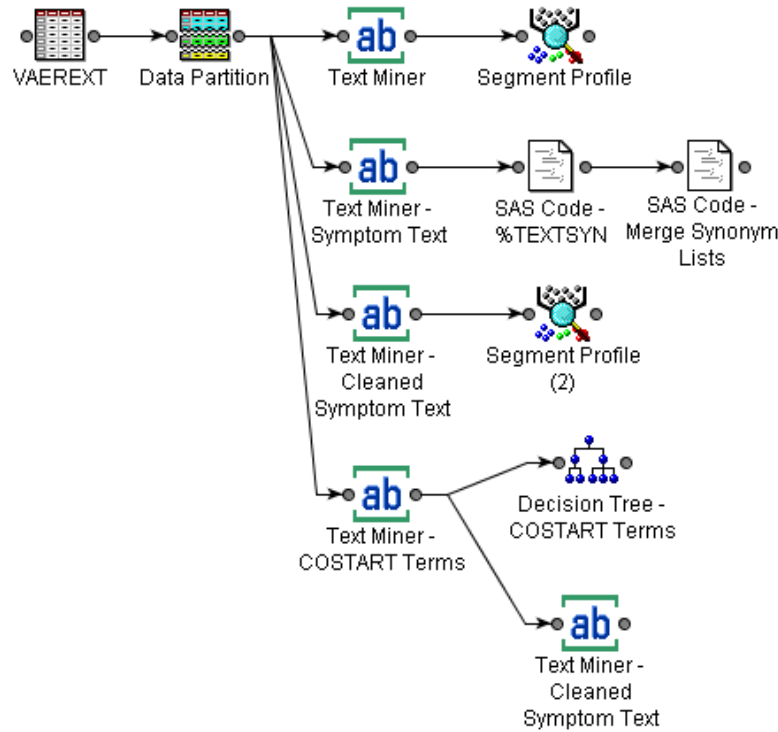
Note: Red indicates correct classification and blue indicates incorrect classification. The Decision Tree model does an excellent job of classifying the adverse events that are not considered serious. The model correctly predicts half of them to be serious and half of them to be minor. Δ



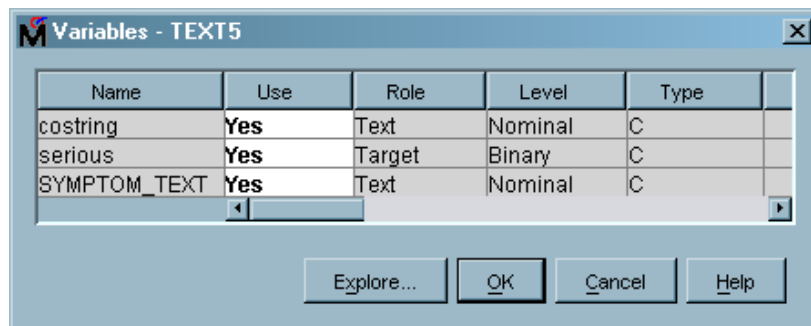
Use the SYMPTOM_TEXT Variable to Model

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

- 1 Copy the Text Miner - Cleaned Symptom Text node and connect it to the Text Miner - COSTART Terms node.

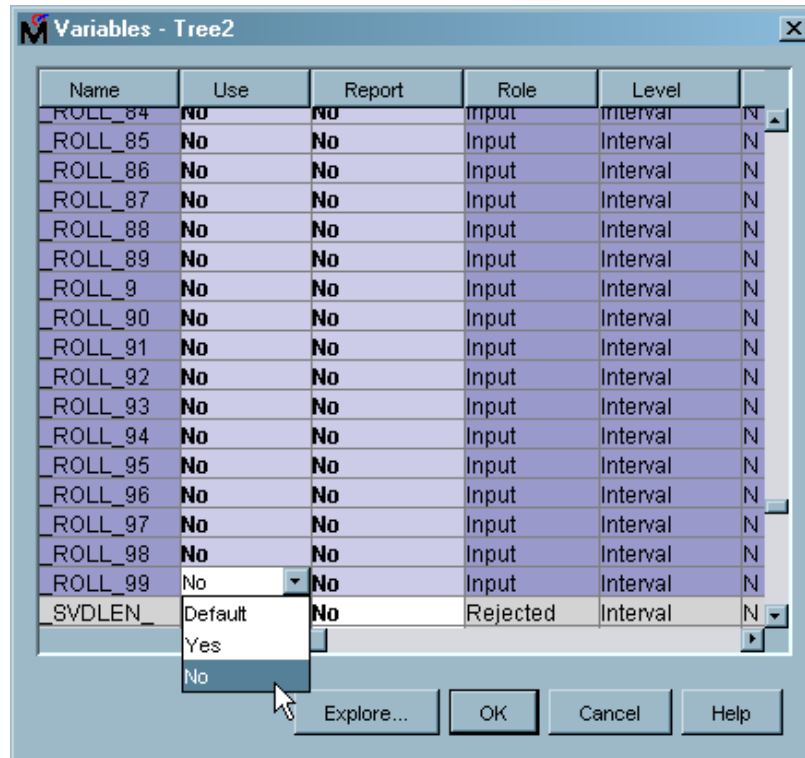


- 2 This second Text Miner - Cleaned Symptom Text node will be used to analyze the SYMPTOM_TEXT variable. SYMPTOM_TEXT will be the default parse variable because it is the longest text field in the data set. So you need to specify COSTRING as a parse variable as well. Select the second Text Miner - Cleaned Symptom Text node and click on the **Variables** property in the Properties panel.
- 3 In the Variables window, set the following:
 - Set **SYMPTOM_TEXT** to **Use=Yes**.
 - Set **costring** to **Use=Yes**.
 - Set **serious** to **Use=Yes**.

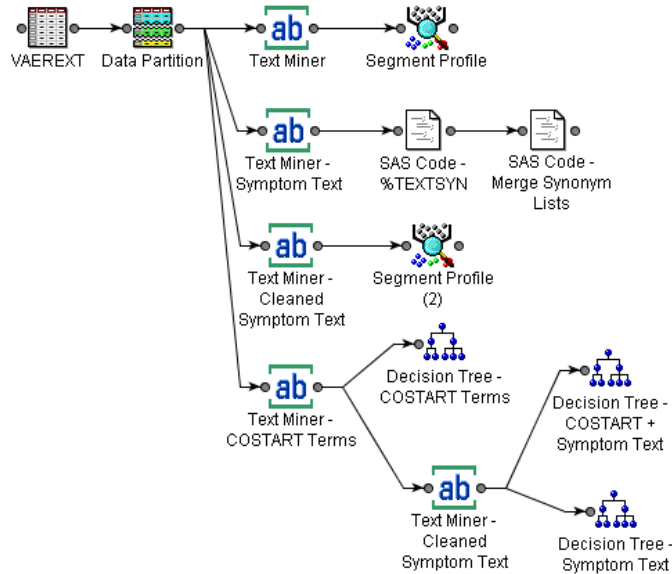


- 4 Set the following Text Miner - Cleaned Symptom Text node properties in the Properties panel:
 - Set **Compute SVD** to **Yes**.
 - Set **SVD Resolution** to **Low**.
 - Set **Term Weight** to **Mutual Information**.
- 5 Run the Text Miner - Cleaned Symptom Text node.

- 6 From the **Model** tab, drag and drop a Decision Tree node and connect it to the Text Miner - Cleaned Symptom Text 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 Rename this node Decision Tree - Symptom Text.
- 8 Click the **Variables** property in the Decision Tree - Symptom Text Properties panel. The Variable window opens.
- 9 Click and scroll to select all of the **_ROLL_** variables and set them to **Use=No**.



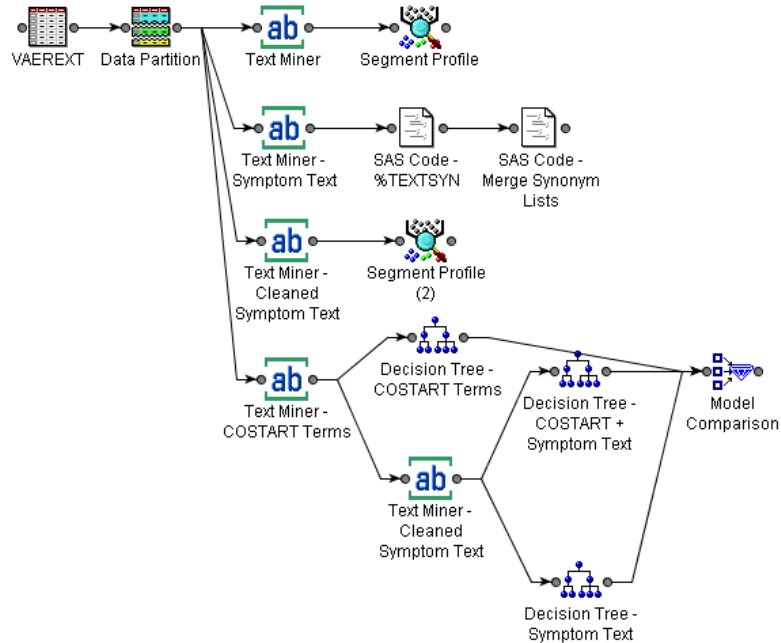
- 10 Click **OK** to save your changes.
- 11 Run the Decision Tree - Symptom Text node. Click **OK** in the Run Status window.
- 12 From the **Model** tab, drag and drop a Decision Tree node and connect it to the Text Miner - Cleaned Symptom Text node.
- 13 Rename this new Decision Tree node Decision Tree - COSTART + Symptom Text. 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

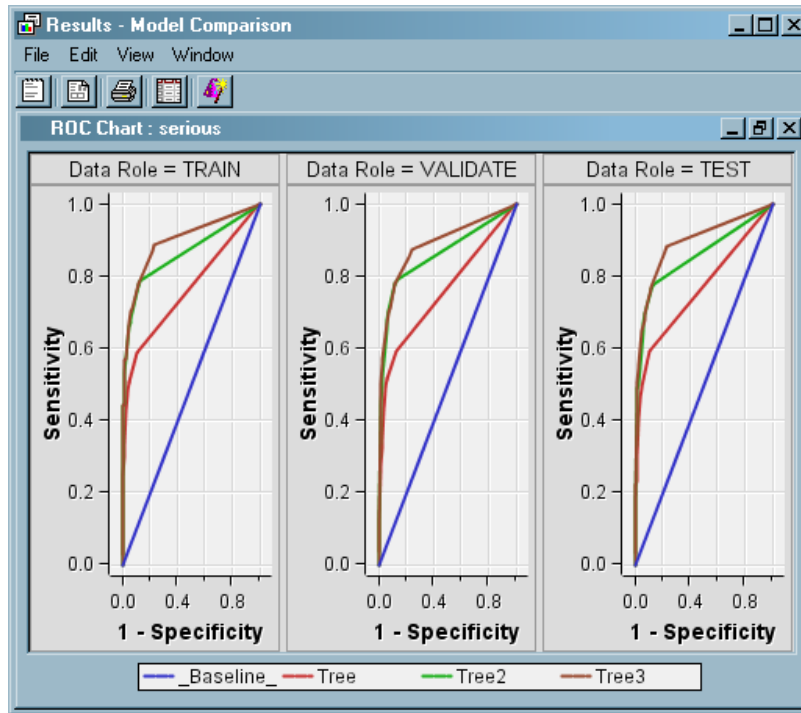
To compare the models, complete the following steps:

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



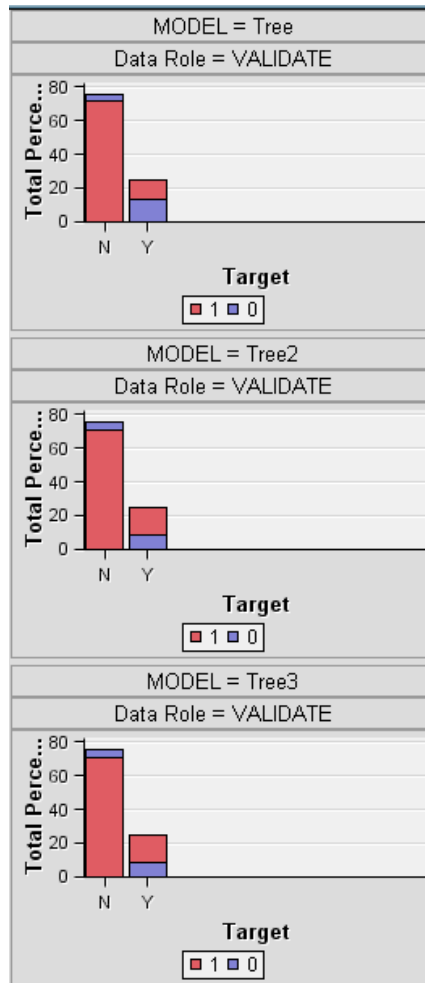
- 2 Run the Model Comparison node and click **Results** in the Run Status window to view the results.
- 3 Click the ROC Chart to maximize the window.

The greater the area under the curve, the better the model. 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 the 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 pull-down menu at the top of the Results window to view the Classification Chart.

Note: Red indicates correct classification and blue indicates incorrect classification. In the combined model, the majority of serious events are now classified as serious. \triangle



- 5 Close the Model Comparison Results window. It would be useful to see which variables are most important in the combined model for predicting serious events.
- 6 Right-click on the Decision Tree - COSTART + Symptom Text 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 themselves in predicting a serious adverse event. △

Interestingly, none of the demographic information, such as the patient’s age or sex, improves a model that consists entirely of textual data. None of these variables are listed in the **Variable Importance** results.

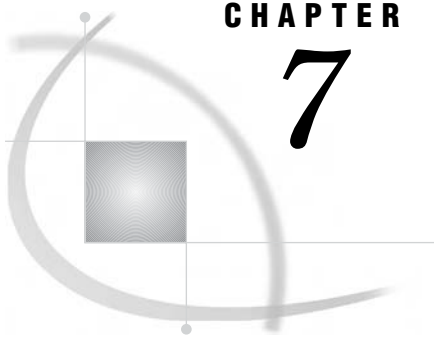
Obs	NAME	LABEL	NRULES	IMPORTANCE	VIMPORTANCE	RATIO
48	_SVD_3		2	1.00000	1.00000	1.00000
49	_SVD_5		4	0.62601	0.59947	0.95761
50	SYM_CNT		4	0.54074	0.48425	0.89552
51	_SVD_1		1	0.28926	0.22279	0.77019
52	_SVD_8		1	0.25067	0.23989	0.95700
53	_SVD_4		1	0.24757	0.16123	0.65126
54	_ROLL_1	abnorm	2	0.18414	0.16856	0.91538
55	_SVD_19		1	0.14889	0.13113	0.88068
56	_ROLL_43	anaphyl	1	0.13033	0.13310	1.02125
57	_ROLL_17	v	1	0.08072	0.08921	1.10514
58	_ROLL_2	convuls	1	0.06804	0.04305	0.63276
59	_SVD_20		1	0.06555	0.08338	1.27202

- Minimize the Output window, and then maximize the window containing 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:

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



CHAPTER

7

Next Steps: A Quick Look at Additional Features

<i>The %TMFILTER Macro</i>	65
<i>The %TMPUNC Macro</i>	65
<i>Tips for Text Mining</i>	66
<i>Processing a Large Collection of Documents</i>	66
<i>Dealing with Long Documents</i>	66
<i>Processing Documents from an Unsupported Language or Encoding</i>	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 for SAS Text Miner 3.1 Java Help for more information.

The %TMPUNC Macro

The %TMPUNC macro strips unwanted punctuation from terms in your document collection. If your documents contain terms with run-on punctuation, such as ****people** or **+bags**, these punctuation characters become part of the terms when SAS Text Miner parses the documents. The %TMPUNC macro enables you to convert terms with run-on punctuation by putting spaces before and after the punctuation characters to prevent them from appearing as part of the term. Without the %TMPUNC macro, these two

examples would parse out as two terms, ****people** and **+bags**. After running the %TMPUNC macro, they would parse as five terms:

- *
- *
- **people**
- **+**
- **bags**

See Other Macros: %TEXTSYN and %TMPUNC in the Text Miner node documentation for SAS Text Miner 3.1 Java Help for more information.

Tips for Text Mining

Processing a Large Collection of Documents

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

- Use a sample of the document collection.
- Set some of the parse properties to **No**, such as **Find Entities**, **Noun Groups**, and **Terms in Single Document**.
- Reduce the number of SVD dimensions or roll-up terms. If you are running into memory problems with the SVD approach, you can roll up a certain number of terms, and then the remaining terms are automatically dropped.
- Limit parsing to high information words by turning off all parts of speech other than nouns, proper nouns, noun groups, and verbs.
- Structure sentences properly for best results, including correct grammar, punctuation, and capitalization. Entity extraction does not always generate reasonable results.

Dealing with Long Documents

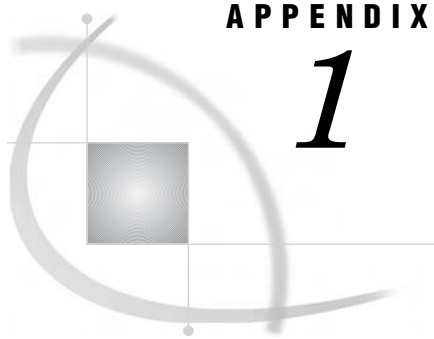
SAS Text Miner uses the "bag-of-words" approach to represent documents. That means that documents are represented with a vector that contains the frequency with which each term occurs in each document. In addition, word order is ignored. This approach is very effective for short, paragraph-sized documents, but it can cause a harmful loss of information with longer documents. You may want to consider preprocessing your long documents in order to isolate the content that is really of use in your model. For instance, if you are analyzing journal papers, you may 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 may 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 may 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.



APPENDIX

1

Recommended Reading

Recommended Reading 69

Recommended Reading

Here is the recommended reading list for this title:

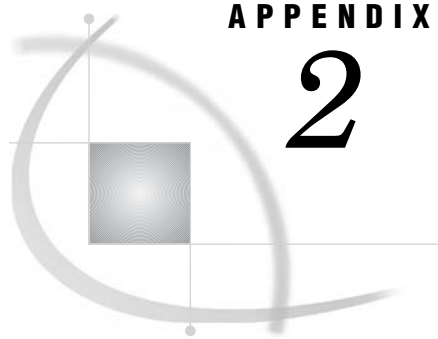
- *Getting Started with SAS Enterprise Miner 5.2*
- *Getting Started with SAS 9.1 Text Miner*

For a complete list of SAS publications, see the current *SAS Publishing Catalog*. To order the most current publications or to receive a free copy of the catalog, contact a SAS representative at

SAS Publishing Sales
SAS Campus Drive
Cary, NC 27513
Telephone: (800) 727-3228*
Fax: (919) 677-8166
E-mail: sasbook@sas.com
Web address: support.sas.com/pubs

* For other SAS Institute business, call (919) 677-8000.

Customers outside the United States should contact their local SAS office.



APPENDIX

2

Vaccine Adverse Event Reporting System Data Preprocessing

VAERS Data Preprocessing 71

VAERS 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 3.1 zip file. For more information about the `IMPORT` procedure, see SAS OnlineDoc 9.1.3 in the following URL: <http://support.sas.com/onlinedoc/913/docMainpage.jsp>.

```
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, for example, listed earlier, this would flag both the Pertussis and Diphtheria/Tetanus variables.

The SAS code Vaerssetup.Sas used to generate the resulting table, VAEREXT, is in the Getting Started with Text Miner 3.1 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 coterm $ 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','DTOX','TD','TTOX') 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;
  end;
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;
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;

```

The following figure shows some columns of the VAEREXT SAS data set that was created by the Vaerssetup.Sas code.

VIEWTABLE: TMP1.vaerext				
	costrng	SYMPTOM_TEXT	serious	pedflag
1	ANXIETY CELLULITIS EDEMA INJECT SITE FEVER HYPERTENS HYSN INJECT SITE LEUKOCYTOSIS MYALGIA PAIN INJECT SITE VASODILAT	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	Y	N
2	ABORTION LAB TEST ABNORM	Information has been received from an NP concerning a 29 year old female pt who on an unspecified date was vaccinated with varicella virus vaccine live. The NP indicated that the pt was 2 weeks pregnant when she received the vaccination. No adverse experiences were reported. The pt sought unspecified medical attention. Follow-up information was received from a physician assistant who indicated that the pt (gravida 4, para 3) was vaccinated on 9/18/01 with a 1st dose of varicella virus vaccine li	Y	N
3	AMNESIA DELUSIONS	Memory loss, family loss. Mother and father are dead. Permanent coma. Delusional thoughts if my brain died. Adopted. Birth certificate. Orange birth certificate. White birth certificate is necessary.	Y	Y
4	REACT UNEVAL	Sabin tri vaccines were not good ones. They make you taller and handicapped looking.	Y	Y

Glossary

clustering

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

concept linking

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

data source

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

diagram

See process flow diagram.

entity

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

libref (library reference)

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

file NEWBDAY. You assign a libref with a LIBNAME statement or with an operating system command.

model

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

node

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

parse

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

partition

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

process flow diagram

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

roll-up terms

the highest-weighted terms in the document collection.

SAS data set

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

score

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

segmentation

the process of dividing a population into subpopulations of similar individuals. Segmentation can be done in a supervisory mode (using a target variable and various techniques, including decision trees) or without supervision (using clustering or a Kohonen network).

singular value decomposition

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

stemming

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

stop list

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

test data

currently available data that contains input values and target values that are not used during training, but which instead are used for generalization and to compare models. See also training data and test data.

training data

currently available data that contains input values and target values that are used for model training. See also test data and validation data.

validation data

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

variable

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

Index

A

accessibility features 3
Automatically Cluster property 22

C

cleaning data 33
 creating a stop list 43
 creating a synonym data set 37
 examining results using merged synonym data sets 40
 exploring result improvements 46
 using a synonym data set 35
Cluster properties 22
clusters 25
concept linking 26
converting files into data sets 65
COSTART coding system 51
COSTART Terms node 51
COSTRING variable 51

D

data cleaning
 See cleaning data
Data Partition node 20
data segments 28
data sets
 converting files into 65
 importing 33
 merged synonym data sets 40
 synonym data sets 35, 37
data source
 creating for projects 14
descriptive mining 1
Descriptive Terms property 22
diagrams
 creating 18
Different Parts of Speech property 21
document requirements 1
documents
 from unsupported language or encoding 67
 large collection of 66
 long 66

E

encoding, unsupported 67

F

files
 converting into data sets 65

H

Help 10

I

Ignore Outliers property 23
Ignore Parts of Speech property 21
importing data sets 33
input data
 identifying 19
 partitioning 20
interactive results
 viewing 23

L

languages, unsupported 67
large collection of documents 66
long documents 66

M

MedDRA program 64
merged synonym data sets 40
misspelled terms 37
Model Comparison node 61
modeling
 See predictive modeling

N

node properties 20

P

Parse properties 21
Parse Variable 23
partitioning input data 20
path for projects 13
predictive mining 1
predictive modeling 51
 comparing models 61
 COSTRING variable for 51

- exercises 64
- SYMPTOM_TEXT variable for 58
- projects
 - creating 11
 - creating data source 14
 - creating diagrams 18
 - path for 13
 - setting up 11
- properties, node 20
- punctuation
 - stripping from terms 65

R

- results
 - examining with merged synonym data sets 40
 - exploring result improvements 46
 - viewing 23

S

- SAS Enterprise Miner 2
- SAS Text Miner 2
 - accessibility features 3
 - Help 10
- Section 508 standards 3
- Segment Profile node 28
- segments 28
- Set Terms in Single Document property 21
- stems 36
- stop lists
 - creating 43
 - defined 35
- SYMPTOM_TEXT variable analysis 19
 - examining data segments 28
 - identifying input data 19
 - partitioning input data 20

- setting node properties 20
- viewing interactive results 23
- SYMPTOM_TEXT variable for modeling 58
- synonym data sets 35
 - creating 37
 - merged 40
- Synonyms property 21

T

- Term Weight property 22
- text cleaning
 - See* cleaning data
- Text Miner node 2
- text mining 1
 - descriptive mining 1
 - document requirements for 1
 - general order for 3
 - large collection of documents 66
 - long documents 66
 - predictive mining 1
 - process 3
 - tips for 66
 - unsupported language or encoding 67
- tips for text mining 66
- %TMFILTER macro 2, 65
- %TMPUNC macro 65
- Transform properties 22

U

- unsupported languages or encoding 67

V

- VAERS data preprocessing 71

Your Turn

We want 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**.

SAS® Publishing gives you the tools to flourish in any environment with SAS!

Whether you are new to the workforce or an experienced professional, you need to distinguish yourself in this rapidly changing and competitive job market. SAS® Publishing provides you with a wide range of resources to help you set yourself apart.

SAS® Press Series

Need to learn the basics? Struggling with a programming problem? You'll find the expert answers that you need in example-rich books from the SAS Press Series. Written by experienced SAS professionals from around the world, these books deliver real-world insights on a broad range of topics for all skill levels.

support.sas.com/saspress

SAS® Documentation

To successfully implement applications using SAS software, companies in every industry and on every continent all turn to the one source for accurate, timely, and reliable information—SAS documentation. We currently produce the following types of reference documentation: online help that is built into the software, tutorials that are integrated into the product, reference documentation delivered in HTML and PDF—free on the Web, and hard-copy books.

support.sas.com/publishing

SAS® Learning Edition 4.1

Get a workplace advantage, perform analytics in less time, and prepare for the SAS Base Programming exam and SAS Advanced Programming exam with SAS® Learning Edition 4.1. This inexpensive, intuitive personal learning version of SAS includes Base SAS® 9.1.3, SAS/STAT®, SAS/GRAPH®, SAS/QC®, SAS/ETS®, and SAS® Enterprise Guide® 4.1. Whether you are a professor, student, or business professional, this is a great way to learn SAS.

support.sas.com/LE



**THE
POWER
TO KNOW®**

