Text Analytics

Handout

Text Analytics Handout was developed by Terry Woodfield. Additional contributions were made by Peter Christie and Rich Perline. Editing and production support was provided by the Curriculum Development and Support Department.

Text Analytics Handout

Book code E2847, course code SGF15TA, prepared date 3/27/2015. SGF15TA_001

Table of Contents

To learn more…

For information about other courses in the curriculum, contact the SAS Education Division at 1-800-333-7660, or send e-mail to training@sas.com. You can also find this information on the web at http://support.sas.com/training/ as well as in the Training Course Catalog.

For a list of other SAS books that relate to the topics covered in this course notes, USA customers can contact the SAS Publishing Department at 1-800-727-3228 or send e-mail to sasbook@sas.com. Customers outside the USA, please contact your local SAS office.

Also, see the SAS Bookstore on the web at http://support.sas.com/publishing/ for a complete list of books and a convenient order form.

Chapter 1 Text Analytics

1.1 Quick Introduction to Text Analytics

Objectives

- Explain some of the jargon of Text Analytics.
- Describe the basic interface of SAS Enterprise Miner.
- Illustrate key elements of Text Mining using SAS Text Miner.

3

Text Analytics

- You use the terms text analytics, text data mining, and text mining synonymously in this course.
- Text analytics uses *algorithms* for turning free-form text into data that can then be analyzed by applying *statistical and machine learning methods*, as well as *natural language processing techniques*.
- Text analytics encompasses many sub areas.

4

Text Mining

Text mining as presented here has the following characteristics:

- operates with respect to a *corpus* of documents
- creates a *dictionary* or *vocabulary* to identify relevant terms
- accommodates a variety of *metrics* to quantify the contents of a document within the corpus

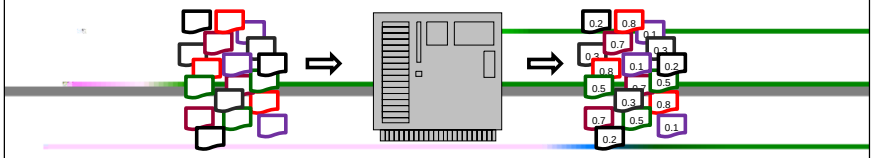
5

Text Mining

- Derives a *structured vector* of measurements for each document relative to the corpus
- Uses *analytical methods* that are applied to the structured vector of measurements based on the goals of the analysis (for example, groups documents into segments)

6

Text Mining – Training



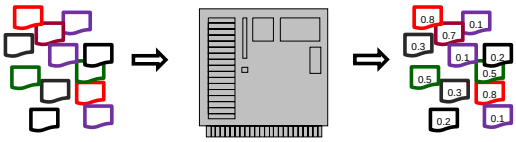
Corpus = Original Documents

Text Mining Training

Scores

7

Text Mining – Scoring New Documents



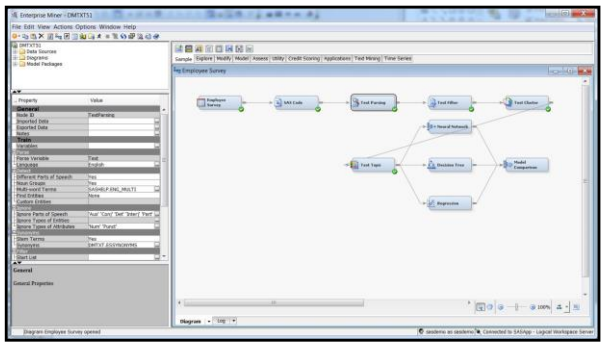
New Documents

Text Mining Trained Model

Scores

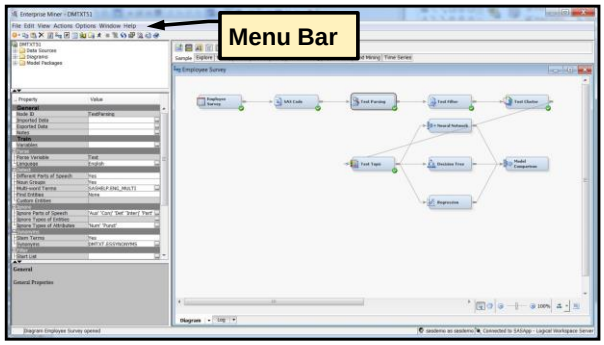
8

SAS Enterprise Miner



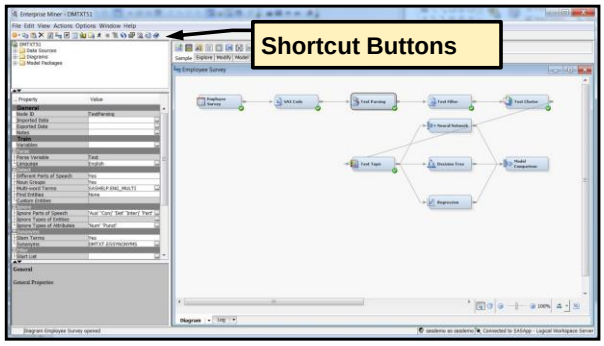
9

SAS Enterprise Miner



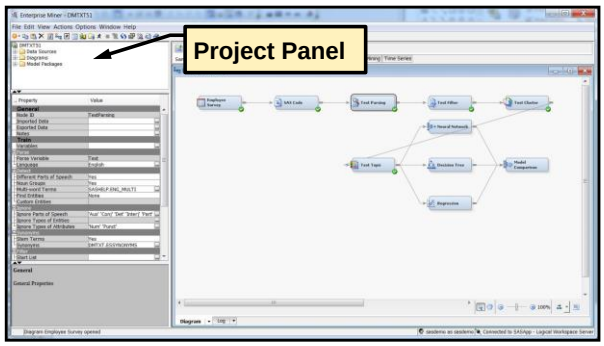
10

SAS Enterprise Miner



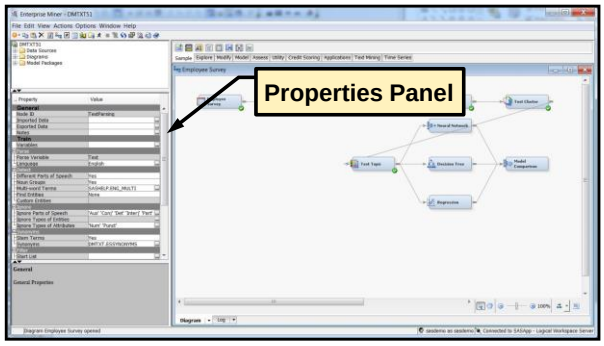
11

SAS Enterprise Miner



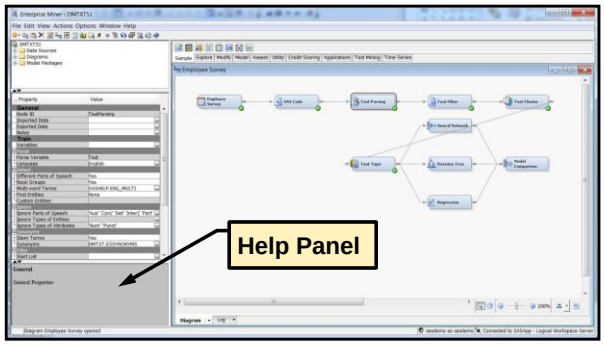
12

SAS Enterprise Miner



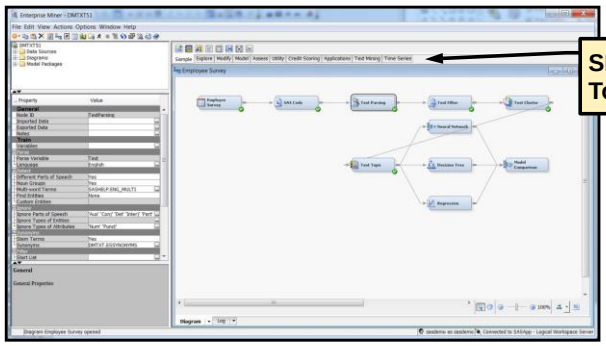
13

SAS Enterprise Miner



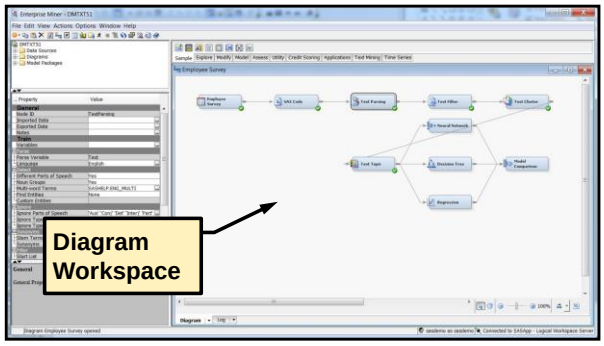
14

SAS Enterprise Miner



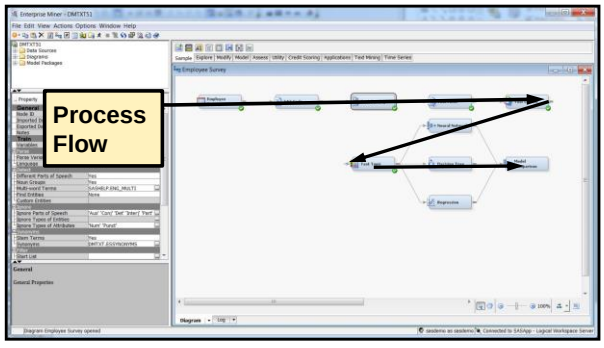
15

SAS Enterprise Miner



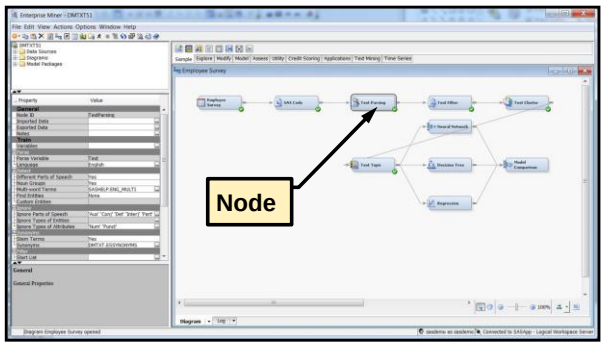
16

SAS Enterprise Miner



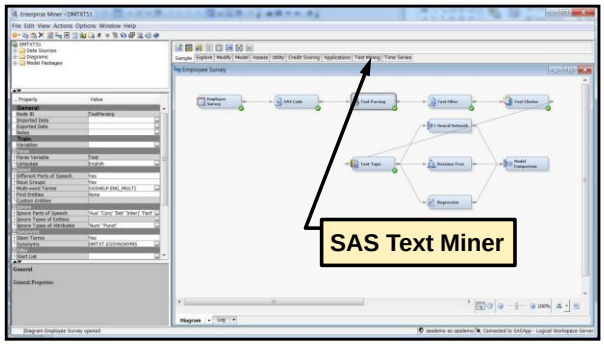
17

SAS Enterprise Miner



18

SAS Enterprise Miner



19

Text Mining Tools

The following SAS Enterprise Miner text mining nodes are discussed:

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

20

Using SAS Enterprise Miner

- Table types
  - Raw
  - Train
  - Validate
  - Test
  - Score
  - Transaction
- Metadata
  - Variable roles
  - Variable levels

21

1.2 From Documents to Data

Importing Text Mining Data Sources

- When documents are stored in separate files in the same directory, or subdirectories under the same directory, then the *Text Import* node can be used to create an appropriate SAS data set for text mining.
- When documents are stored together (for example, one document per row in a Microsoft Excel spreadsheet), then the *Import Data Wizard* or *File Import* node can be used to create a text mining data set.
- Sometimes special SAS programming might be required if you are combining text data with other data.

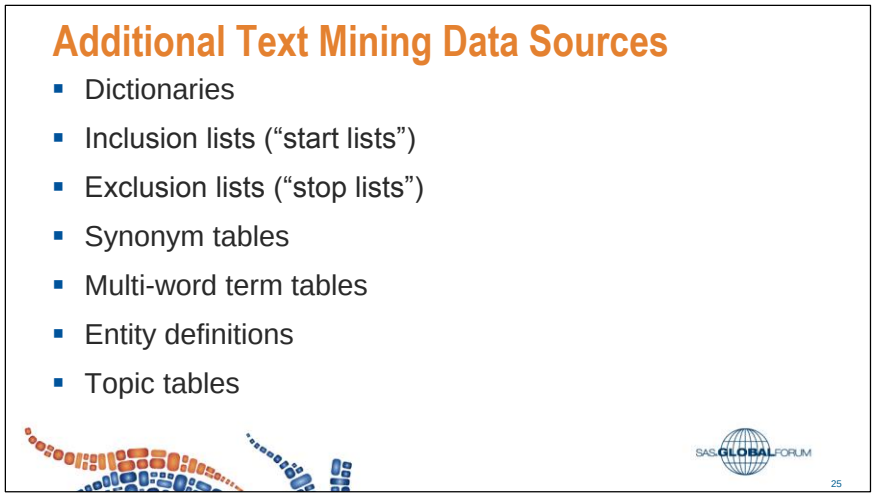
23

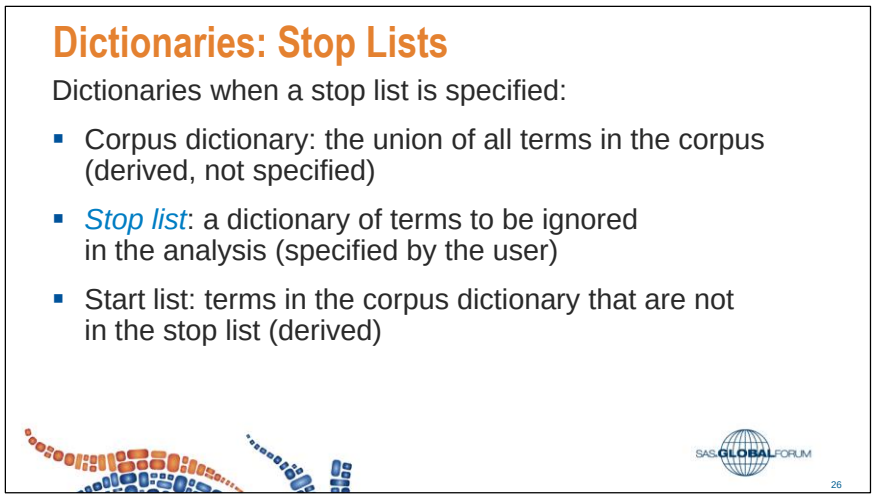
Working with Text Mining Data Sources

Two supported types of textual data:

- *Text* with each document stored as a SAS character variable in the analysis data set. Appropriate when the largest document is smaller than 32,767 characters, which is about 10 pages.
- *Text Location* with the full pathname of the document with respect to the Text Miner server. Appropriate when the document cannot be completely stored as a SAS character variable.

24





A stop list is typically used to remove low information terms that add only noise to an analysis. Noisy data has no descriptive or predictive value. Stop lists are commonly used when documents are unedited and contain numerous misspellings and typographical errors.

Dictionaries: Start Lists

Dictionaries when a start list is specified:

- Corpus dictionary: the union of all terms in the corpus (derived, not specified)
- *Start list*: a dictionary of terms to be used in the analysis (specified by the user)
- Stop list: terms in the corpus dictionary that are not in the start list (derived)

27

A start list can be a business or technical dictionary that is developed by the analyst or acquired from other sources. Start lists are commonly used when documents are well edited and contain few typographical errors.

SASHELP.ENGSTOP

a	across
aboard	actually
about	after
above	afterwards
according	against
accordingly	…

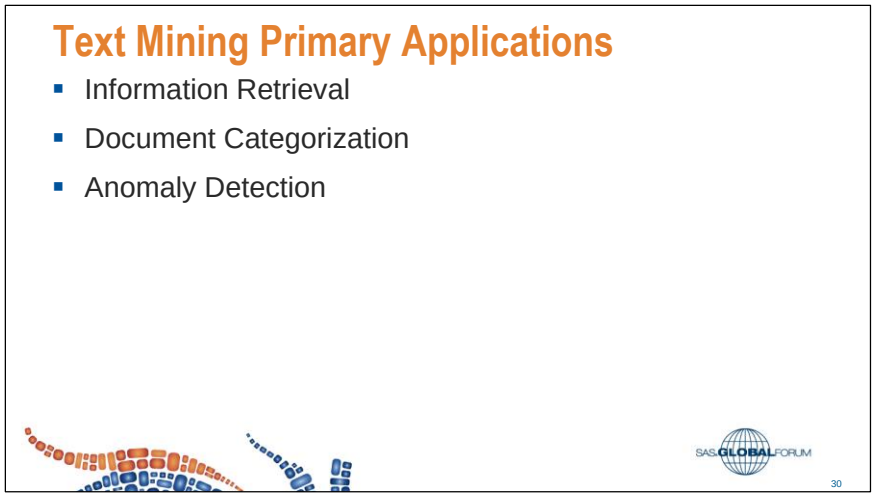
509 low information English words

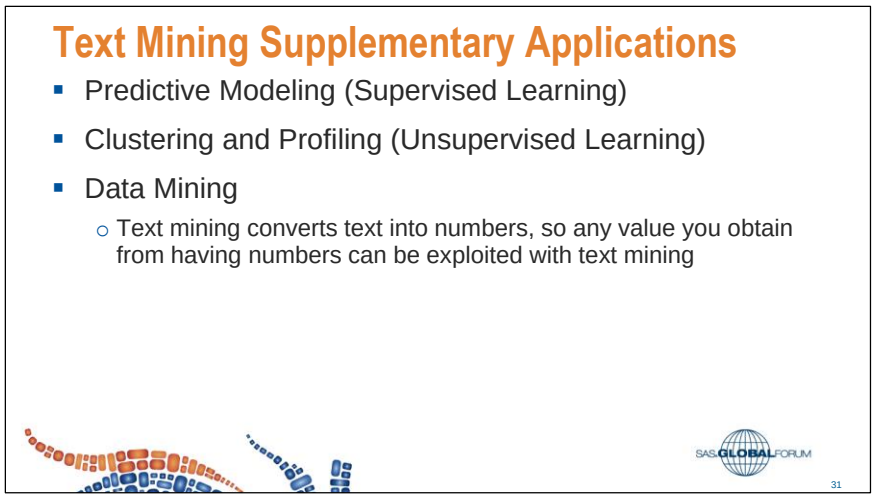
28

Other Text Mining Topics Not Covered Here

- Zipf's Law and frequency filtering
- Stemming and part-of-speech tagging
- Spell checking
- Term weighting schemes
- Entity processing
- Latent Semantic Analysis (LSA) and the Singular Value Decomposition (SVD)

29

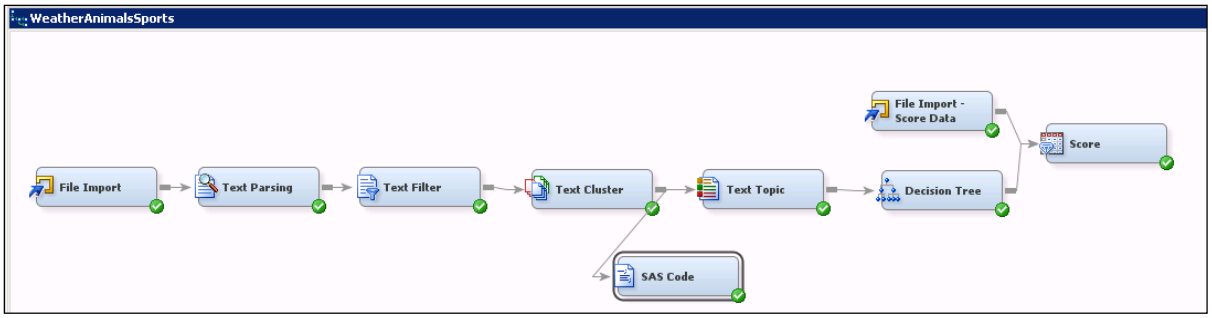




Text Analytics Illustrated with a Simple Data Set

This demonstration illustrates some text analytic results using a simple data set that is designed to be easy to interpret. You can learn a lot about many features of the major Text Mining nodes by working through this example. You also use a SAS Code node to show you how to "get under the hood" and examine some results.

In this class, the project that you use and the diagrams are already set up, at least partially. However, for each demonstration, you should rebuild your own version of each diagram. In some cases, you can make additions to an existing diagram. You start by opening the project called **SGF15TA**. Then, select the diagram **WeatherAnimalsSports**. Set up the flow for this diagram as it is shown below.



1. Insert a **File Import** node in the diagram. This is the first node on the left that you see in the diagram above. In this demonstration, you use a data set that is completely stored in a single Excel spreadsheet. This is one way of getting relatively small text mining data sets into SAS Enterprise Miner. On the Property Sheet for the File Import node, specify the import file as the data set **D:\workshop\winsas\SGF15TA\WeatherAnimalsSports.xls**. Then run this node. Run the File Import node.

   In some classroom configurations, the import file is located at **D:\workshop\SGF15TA\WeatherAnimalsSports.xls.**

2. To see the data set after the **File Import** node is run, go to the **Exported Data** line of the Property Sheet. Click the ellipsis button (...). Then select the **Train** data and click **Browse** near the bottom of the window. You see the rows of the data set. The first seven rows are shown in the display below.

	Target_Subject	TextField
1	A	Bob has two dogs and one cat. The cat is bigger than either of the dogs.
2	S	Carmelo Anthony scored 42 points to lead the NY Knicks basketball team to a win over the Florida Pelicans.
3	S	Come play baseball with us.
4	S	Derek Jeter, the captain of the New York Yankees baseball team, said 2014 will be his last season playing.
5	S	Do you have a baseball or a football that we could play with? You can be on my team.
6	A	Do you like big dogs or little dogs? Dogs are such wonderful animals.
7	W	During the winter, the sun is lower in the sky than it is during the summer. That's why winter days are colder than summer days.

   The data set has two fields: **Target_Subject** (with values **`A`**, **`S`**, **`W`**) and **TextField**, which consists of short sentences. The sentences are about with one of three subjects: **Animals (A)**, **Sports (S)**, a **Weather (W)**.

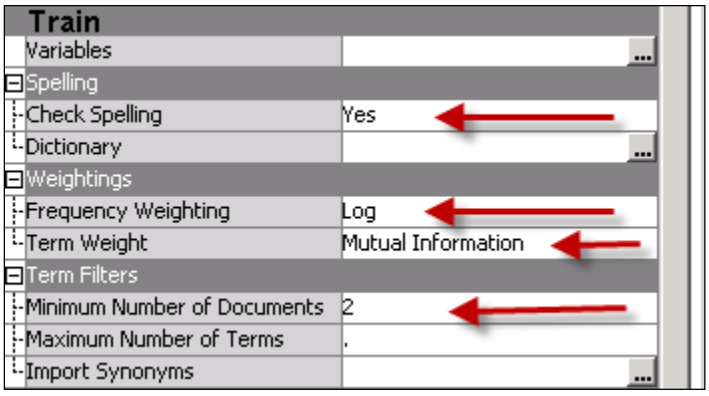
   It is important to understand that the **Target_Subject** field was created by a person interpreting the content of each **TextField**. It was not created automatically by the Text Miner nodes.

   Read through a few of the rows and make sure that you understand the nature of the data set and how it is structured. The variable **TextField** is what is referred to as a *document*. All the rows of **TextField** together (47 rows of data) are referred to as the *corpus collection*.

3. Attach a **Text Parsing** node to the **Text Import** node. This node has the language processing algorithms and has many different options that can be set by the user. For this demonstration, use the default settings. Run the **Text Parsing** node.

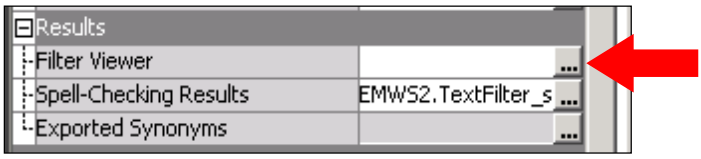
4. Attach a **Text Filter** node to the **Text Parsing** node. Change **Frequency Weighting** from **Default** to **Log**. Change **Term Weight** from **Default** to **Mutual Information**. Notice that **Mutual Information** is recommended for data where a target variable is present and predictive modeling is the goal. Also change the **Minimum Number of Documents** value in the Property Sheet to **2**. This option filters out terms that are not used in at least two documents in the corpus collection. Because you use a very small data set, this number is reduced from the default 4 to 2. Also, change the **Check Spelling** property from **No** to **Yes**. (It is easy to forget that **Check Spelling** is in the **Text Filter** node and not on the **Text Parsing** node. In general, changing this to **Yes** can add a lot of time to processing, so be cautious about its use.)

   The settings for the **Text Filter** node now resemble the following:

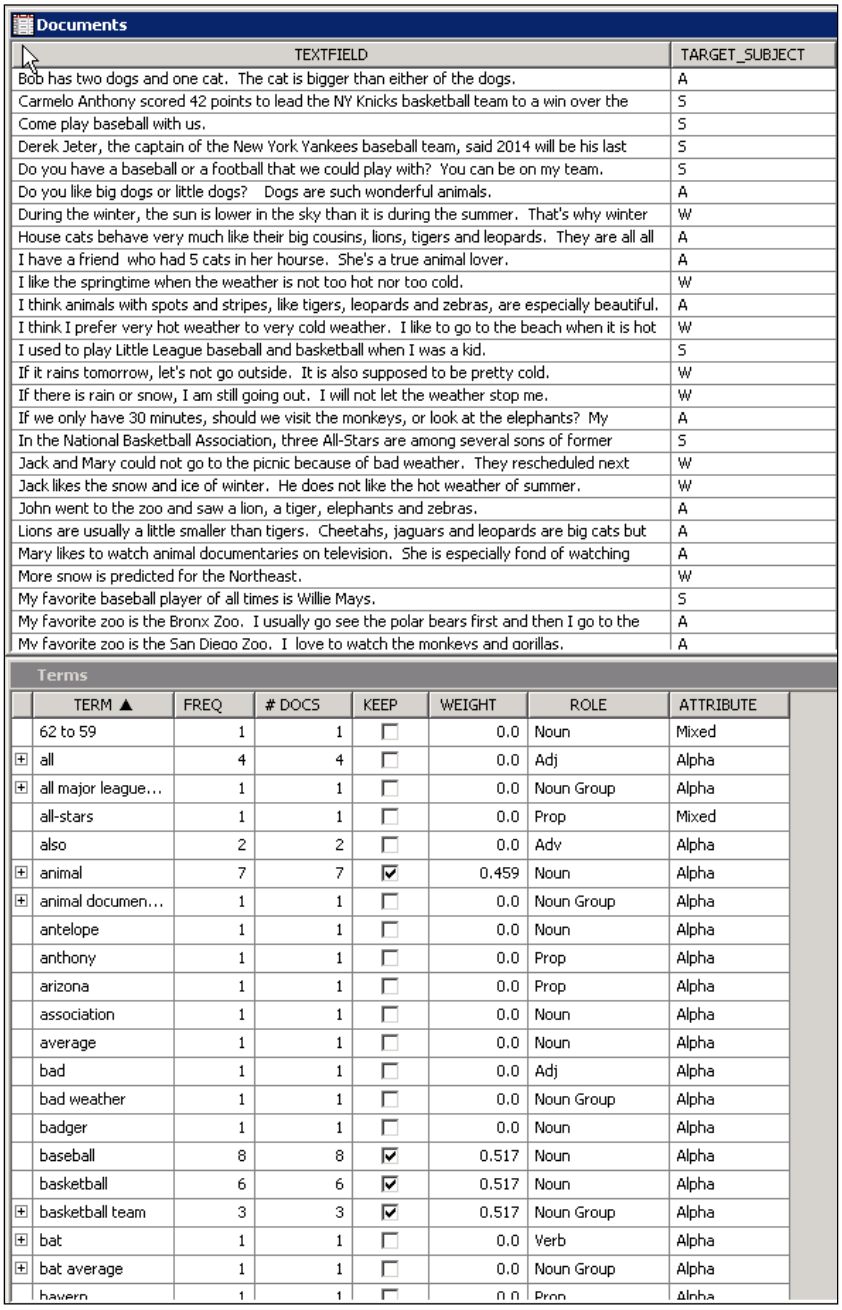


   Run the **Text Filter** node.

5. Open the **Filter Viewer** in the Property Sheet. This is also called the *Interactive Filter Viewer*.



Look at the two main windows that open in the Filter Viewer. You see what is shown in the display below. The first window, labeled **Documents**, simply lists each document and any other variables on the data set; in this case, only the variable **Target_Subject**. The second window, labeled **Terms**, gives information about each of the terms that came out of the **Text Parsing** node. Notice that a *term* does not need to be a single word.



Documents

TEXTFIELD	TARGET_SUBJECT
Bob has two dogs and one cat. The cat is bigger than either of the dogs.	A
Carmelo Anthony scored 42 points to lead the NY Knicks basketball team to a win over the	S
Come play baseball with us.	S
Derek Jeter, the captain of the New York Yankees baseball team, said 2014 will be his last	S
Do you have a baseball or a football that we could play with? You can be on my team.	S
Do you like big dogs or little dogs? Dogs are such wonderful animals.	A
During the winter, the sun is lower in the sky than it is during the summer. That's why winter	W
House cats behave very much like their big cousins, lions, tigers and leopards. They are all all	A
I have a friend who had 5 cats in her hourse. She's a true animal lover.	A
I like the springtime when the weather is not too hot nor too cold.	W
I think animals with spots and stripes, like tigers, leopards and zebras, are especially beautiful.	A
I think I prefer very hot weather to very cold weather. I like to go to the beach when it is hot	W
I used to play Little League baseball and basketball when I was a kid.	S
If it rains tomorrow, let's not go outside. It is also supposed to be pretty cold.	W
If there is rain or snow, I am still going out. I will not let the weather stop me.	W
If we only have 30 minutes, should we visit the monkeys, or look at the elephants? My	A
In the National Basketball Association, three All-Stars are among several sons of former	S
Jack and Mary could not go to the picnic because of bad weather. They rescheduled next	W
Jack likes the snow and ice of winter. He does not like the hot weather of summer.	W
John went to the zoo and saw a lion, a tiger, elephants and zebras.	A
Lions are usually a little smaller than tigers. Cheetahs, jaguars and leopards are big cats but	A
Mary likes to watch animal documentaries on television. She is especially fond of watching	A
More snow is predicted for the Northeast.	W
My favorite baseball player of all times is Willie Mays.	S
My favorite zoo is the Bronx Zoo. I usually go see the polar bears first and then I go to the	A
My favorite zoo is the San Diego Zoo. I love to watch the monkeys and gorillas.	A

Terms

	TERM ▲	FREQ	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
	62 to 59	1	1	☐	0.0	Noun	Mixed
⊞	all	4	4	☐	0.0	Adj	Alpha
⊞	all major league...	1	1	☐	0.0	Noun Group	Alpha
	all-stars	1	1	☐	0.0	Prop	Mixed
	also	2	2	☐	0.0	Adv	Alpha
⊞	animal	7	7	☑	0.459	Noun	Alpha
⊞	animal documen...	1	1	☐	0.0	Noun Group	Alpha
	antelope	1	1	☐	0.0	Noun	Alpha
	anthony	1	1	☐	0.0	Prop	Alpha
	arizona	1	1	☐	0.0	Prop	Alpha
	association	1	1	☐	0.0	Noun	Alpha
	average	1	1	☐	0.0	Noun	Alpha
	bad	1	1	☐	0.0	Adj	Alpha
	bad weather	1	1	☐	0.0	Noun Group	Alpha
	badger	1	1	☐	0.0	Noun	Alpha
	baseball	8	8	☑	0.517	Noun	Alpha
	basketball	6	6	☑	0.517	Noun	Alpha
⊞	basketball team	3	3	☑	0.517	Noun Group	Alpha
⊞	bat	1	1	☐	0.0	Verb	Alpha
⊞	bat average	1	1	☐	0.0	Noun Group	Alpha
	bavern	1	1	☐	0.0	Prop	Alpha

The information shown in the Terms window is the following:

FREQ = number of times the term appeared in the entire corpus

#DOCS = total number of documents in which the term appeared

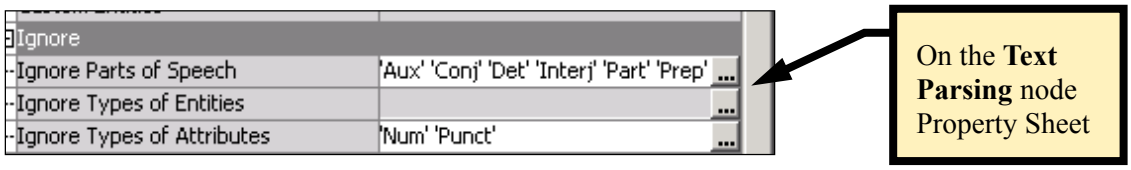
KEEP = whether the term is kept for calculations

WEIGHT = a term weight (in this case, Mutual Information) because you specified **entropy** as the term weight to use in the Property Sheet

ROLE = part of speech of the term

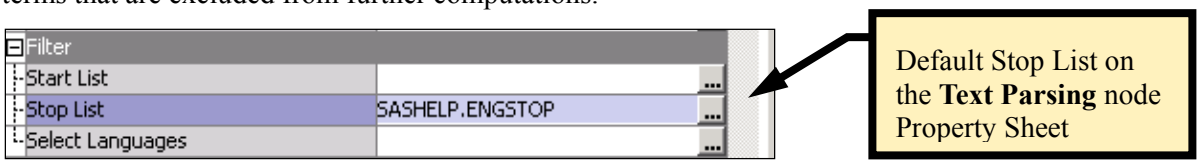
ATTRIBUTE = the different categories are listed later in the chapter

Go to the **Terms** window and confirm that "the" is not listed. (If the column of **Terms** is not already in alphabetical order, you can sort a column by clicking on the heading.) Why does the most common word in the English language not appear on the list? To understand why, click the **Text Parsing** node so that the Property Sheet for that node is visible. Look at the properties near the bottom. You can see that there is an **Ignore Parts of Speech** property. By default, this excludes certain terms that are very common. In particular, *Det* represents *Determiner*, which is a class of common words and phrases such as **`the`**, **`that`**, **`an`**, and so on. These are eliminated unless you modify this property.

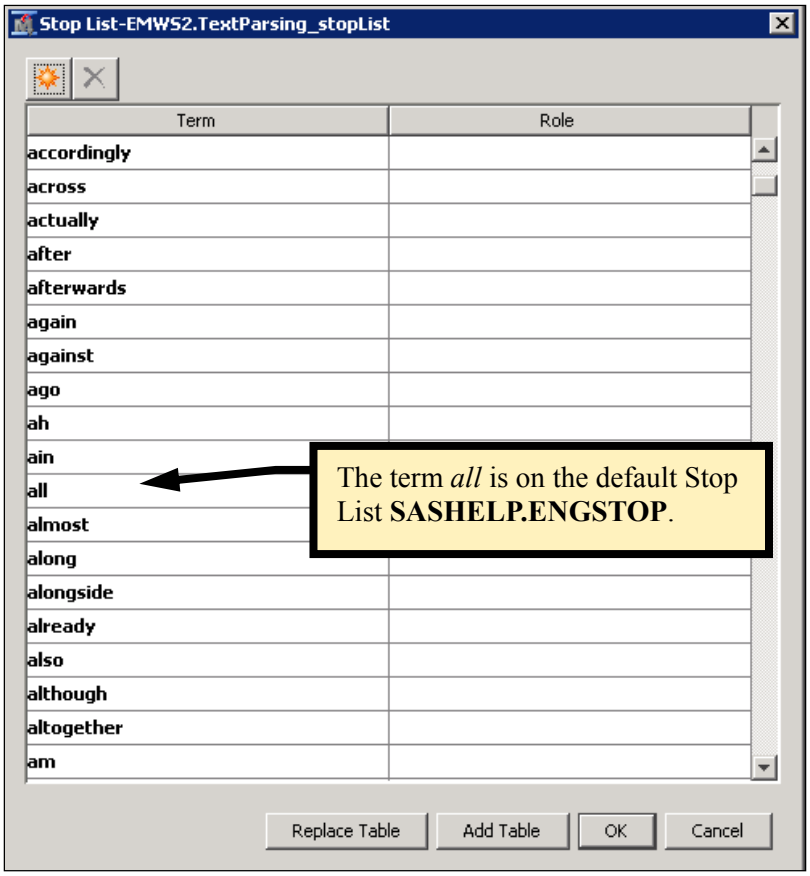


Go back to the **Text Filter** node. Why are some of the terms kept (KEEP is checked), but others are not kept (KEEP is unchecked)? There are several reasons why a word is not kept, and these can depend on settings in both the Text Parsing node and the Text Filter node. One reason, such as for the word **`antelope`**, is that it does not appear in enough documents. You previously set the Minimum Number of Documents property to **2** for the Text Filte**r** node. Because **`antelope`** occurs only in one document, it is not kept.

Another reason a term is not kept is if it appears on a stop list used in the Text Parsing node. The default stop list is **SASHELP.ENGSTOP**. If you open and look at it, you see a list of many terms that are excluded from further computations.



If you open **SASHELP.ENGSTOP** from the Text Parsing node, you see that **`all`** is listed as a term not to be used, as in the display below. Therefore, **`all`** is not selected as **KEEP** in the Text Filter node.



6. You now use the two main analytic text mining tools, the **Text Clustering** node and the **Text Topic** node. Attach a **Text Clustering** node to the **Text Filter** node as in the first diagram of this section. The Text Clustering node takes the 47 documents in the example data set and separates them into *mutually exclusive* and *exhaustive* groups (that is, clusters). The number of clusters to be used is under user control. You modify four of the default settings.

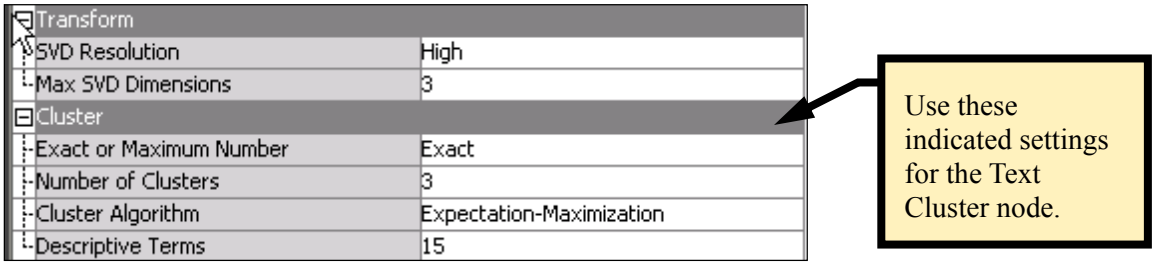
   a. Change **SVD Resolution** from **Low** to **High**.

   b. Change **Max SVD Dimensions** from **100** to **3**.

   c. Change **Exact or Maximum Number** (of clusters) to **Exact**.

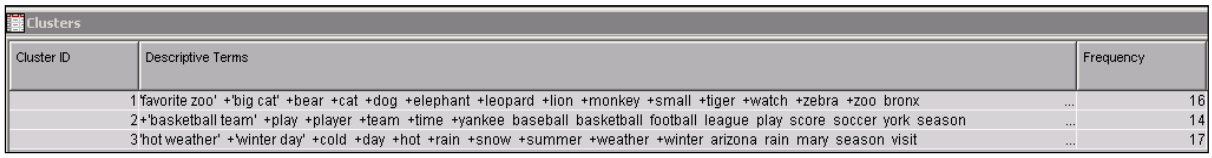
   d. Change **Number of Clusters** from **40** to **3**.

   The settings resemble the ones below.



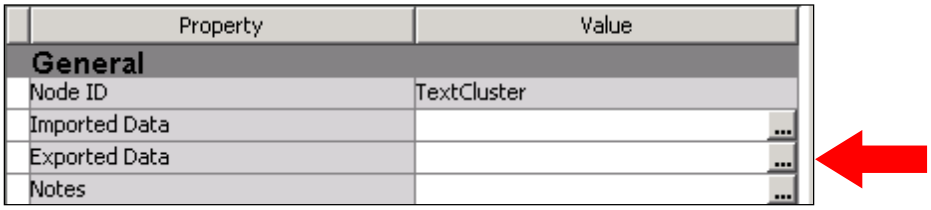
Regarding the **Text Cluster** properties, remember that you are using a very small and simple data set. You know that there are basically three types of documents (animals, sports, weather). It is most reasonable to think in terms of creating a small number of clusters (for example, three to five). Use ***three***. In practice, with real and complex text data, you want to experiment with these parameters. You might want to start with the default property settings. Run the node.

7. Open the **Text Cluster** node results and examine the left side of the Clusters window as shown.

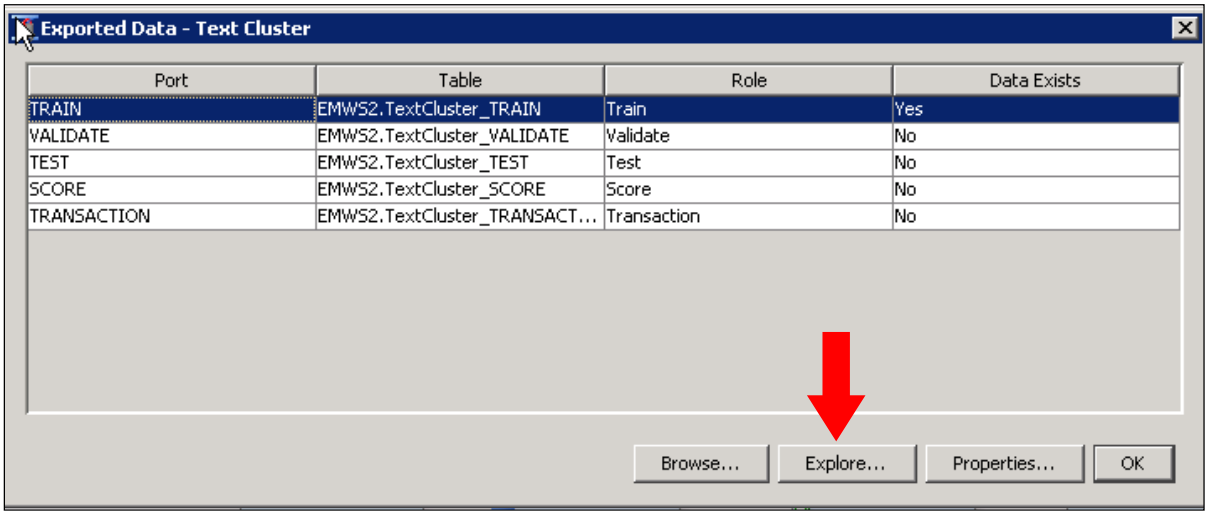


Exactly three clusters were created, as requested in the Property Sheet. The **Descriptive Terms** column shows up to 15 terms that are given to help the user interpret the types of documents that are put into each cluster. (The number can be changed.) These terms are selected by the underlying algorithm as being the most important for characterizing the documents placed into a given cluster. Reading these, you can see that Cluster 1, which has 16 documents, has terms such as **`favorite zoo`**, **`big cat`**, and so on. These documents are likely about animals. The + indicates a stemmed term. Cluster 2 has 14 documents that are likely related to sports. Cluster 3 has 17 documents that likely deal with weather.

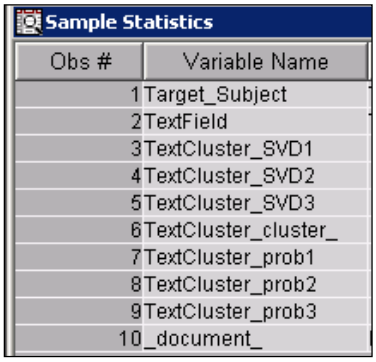
8. To see the new variables that were generated by the Text Cluster node, close out of the results. Select **Exported Data** from the Property Sheet.



Then select the **Train** data set and click **Explore**.



The upper right window (Sample Statistics) shows a list of variables that were exported from the Text Cluster node.


Sample Statistics

Obs #	Variable Name
1	Target_Subject
2	TextField
3	TextCluster_SVD1
4	TextCluster_SVD2
5	TextCluster_SVD3
6	TextCluster_cluster_
7	TextCluster_prob1
8	TextCluster_prob2
9	TextCluster_prob3
10	_document_

Several new variables have been added to the original variables **Target_Subject** and **TextField**:

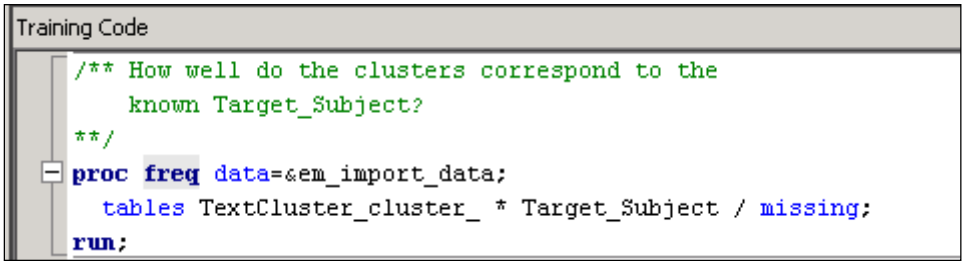
TextCluster_SVD1-TextCluster_SVD3 – These are numeric variables calculated from a singular value decomposition of the (usually weighted) document-term frequency matrix. Each document is represented by its values on these three new variables. The values are also normalized so that for each document all the squared SVD values sum to 1.0. These are the variables that are used to cluster the documents

TextCluster_cluster_ – This is the Cluster ID, a categorical variable. In this example, it is simply a number from 1 to 3 because three clusters were created. The clusters were generated by performing a cluster analysis on the three **TextCluster_SVD** variables. The interpretation of the clusters begins with looking at the descriptive terms given for each cluster, as you did earlier.

TextCluster_prob1 - TextCluster_prob3 – These variables are the probabilities of membership in each cluster for a given document. The sum of these probabilities is 1. A document is assigned to the cluster where it has the highest membership probability.

document – This is a document ID.

9. Many ways to do further explorations with these results can be helpful for learning about what the text mining nodes are doing and for looking more deeply at certain aspects of the analysis. SAS Enterprise Miner provides a SAS Code utility node that is especially good for this. Attach a **SAS Code** node to the **Text Cluster** node and go into the Code Editor on the Property Sheet. Enter the following code:


Training Code

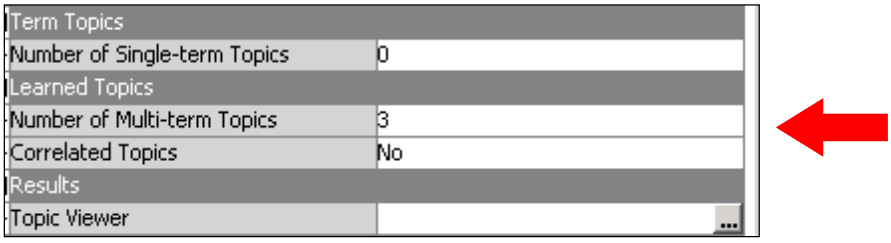
```
/** How well do the clusters correspond to the
    known Target_Subject?
**/
proc freq data=&em_import_data;
  tables TextCluster_cluster_ * Target_Subject / missing;
run;
```

The macro variable **&em_import_data** refers to the Training data set ***after*** it is processed by the Text Cluster node and imported into the SAS Code node. Because there is a target variable (**Target_Subject**) created by a person who read the documents, it is interesting to see how the clusters automatically created by the Text Cluster node align with how the documents were labeled by the person. The PROC FREQ step does this by cross-tabulating the cluster variable (**TextCluster_cluster_**) with the target. Run this code and look at the results.

```
The FREQ Procedure

Table of TextCluster_cluster_ by Target_Subject

TextCluster_cluster_
          Target_Subject(Target_Subject)

Frequency|
Percent  |
Row Pct  |
Col Pct  |A       |S       |W       |  Total
         |        |        |        |
         |        |        |        |
---------+--------+--------+--------+
       1 |     16 |      0 |      0 |     16
         |  34.04 |   0.00 |   0.00 |  34.04
         | 100.00 |   0.00 |   0.00 |
         |  94.12 |   0.00 |   0.00 |
---------+--------+--------+--------+
       2 |      0 |     14 |      0 |     14
         |   0.00 |  29.79 |   0.00 |  29.79
         |   0.00 | 100.00 |   0.00 |
         |   0.00 | 100.00 |   0.00 |
---------+--------+--------+--------+
       3 |      1 |      0 |     16 |     17
         |   2.13 |   0.00 |  34.04 |  36.17
         |   5.88 |   0.00 |  94.12 |
         |   5.88 |   0.00 | 100.00 |
---------+--------+--------+--------+
Total          17       14       16       47
             36.17    29.79    34.04   100.00
```

This crosstabulation shows that Cluster 1 (which was seen previously to have descriptive terms such as **`favorite zoo`**, **`big cat`**, and so on) consists of 16 documents defined that have to do with animals (A) as labeled by the human reader. Cluster 2 (**`basketball team`**, **`play`**, and so on) consists of 14 documents with a target value always equal to S. Cluster 3 (**`hot weather`**, **`winter day`**, and so on) consists of 17 documents, and 16 of them were defined as weather-related. The three clusters line up almost perfectly with the labels given to the documents. ***It would be wonderful if real data worked out this well, but do not expect that!***

10. Set up and run the Text Topic node. Look at some results to see how they compare with the Text Cluster node results. Although a cluster is a mutually exclusive category (that is, each document can belong to one and only one cluster), a document can have more than one topic or it can have none of the topics. Attach a **Text Topic** node directly to the **Text Cluster** node. Make one change to the default properties by specifying **3** as the number of multi-term topics to create. Just as the number of clusters created is a parameter with which you want to experiment when you use the Text Cluster node, this parameter for the number of topics to create is typically something that you might try with different values. In this example, the artificial data set was purposely created with three different topics, so a reasonable value to start with would be 3 to 5 and not the default value of 25. You use 3.



Run the node. Then click on the ellipsis for the **Topic Viewer** on the Property Sheet. The Topic Viewer is an interactive group of windows. The Topics window shows the topics created by the node.

Topics

Topic	Category	Term Cutoff	Document Cutoff	Number of Terms	# Docs
+snow,+cold,+winter,+weather,+hot	Multiple	0.189	0.469	7	9
baseball,+team,basketball,+play,+basketball team	Multiple	0.192	0.351	5	8
+lion,+tiger,+zoo,+animal,+zebra	Multiple	0.196	0.363	7	8

The three topics created by the algorithm also have key descriptive terms to guide interpretation. The five most descriptive terms for each topic are shown. By default, the first topic is selected when you open the viewer. In this example, its descriptive terms start with **`snow`**, **`cold`**, ... This is evidently a topic related to weather. The second topic has descriptive terms starting with **`baseball`**, **`team`**, ... and relating to sports. The descriptive terms for the third topic (**`lion`**, **`tiger`**, …) are interpretable as having to do with animals. With this simple data set, the algorithm did very well in identifying what are known to be the three underlying topics in the documents.

In the Topics window, there is a column labeled **Term Cutoff**. For each created topic, the algorithm computes a topic weight for every term in the corpus. This measures how strongly the term represents or is associated with the given topic. Terms that are above a certain value, called the *Term Cutoff*, appear in yellow in the Terms window shown below.

Look at the Terms window. You can see all the terms above the cutoff value.

Terms

Topic Weight	+	Term	Role	# Docs	Freq
0.496	+	snow	Noun	8	10
0.487	+	cold	Adj	9	9
0.385	+	winter	Noun	6	7
0.338	+	weather	Noun	7	8
0.288	+	hot	Adj	6	7
0.247	+	summer	Noun	4	5
0.243	+	day	Noun	5	6
0.135		rain	Noun	3	3
0.102	+	rain	Verb	2	2
0.095	+	winter day	Noun Group	2	2
0.089		hot weather	Noun Group	3	3
0.04	+	animal	Noun	8	8

You should know, however, that ***all*** terms have a topic weight for each topic, although it might be a very small value.

Part of the Documents window at the bottom of the Interactive Viewer is show below. Every document receives a topic weight for each topic.

Documents

Topic Weight	TextField	Target_Subject
0.959	The weather in NYC is hot in the summer and cold in the winter, but we do not get as much snow as in	W
0.774	Jack likes the snow and ice of winter. He does not like the hot weather of summer.	W
0.704	Winter is my favorite season. I love the cold and the snow.	W
0.682	This has been a very difficult winter, much colder than usual with lots of snow, ice and rain.	W
0.677	During the winter, the sun is lower in the sky than it is during the summer. That's why winter days are	W
0.612	I think I prefer very hot weather to very cold weather. I like to go to the beach when it is hot and sunny.	W
0.599	I like the springtime when the weather is not too hot nor too cold.	W
0.569	We have had snow, snow and more snow for 10 days in a row.	W
0.556	Winter days are often so pretty, even if it is cold.	W
0.446	If there is rain or snow, I am still going out. I will not let the weather stop me.	W
0.41	Phoenix, Arizona, had it's fourth hottest day on record in June, 2013, when the temperature reached 119	W
0.38	More snow is predicted for the Northeast.	W

Notice that in the Documents window, the documents with topic weight values above the document cutoff for this topic (.469) are shown in yellow. However, it is important to observe that there are several documents ***below*** this cutoff value that are nevertheless related to a weather topic. For example, the first document below the cutoff ("If there is rain or snow....") has a topic weight equal to .446 and is not highlighted in yellow. However, this document certainly involves weather. Although a cutoff value for a document can be useful in helping to understand what the topic represents, for some purposes, it is the topic weight itself that is used, such as in predictive modeling. It is also possible to change the cutoff.

To see what variables were generated by the Text Topic node, as was done previously with the Text Cluster node, go to **Exported Data** in the Property Sheet. Select the **Train** data set and click **Explore**. The list of all the variables shows what was previously created by the Text Cluster node and the new **TextTopic** and **TextTopic_raw** variables created by the Text Topic node.

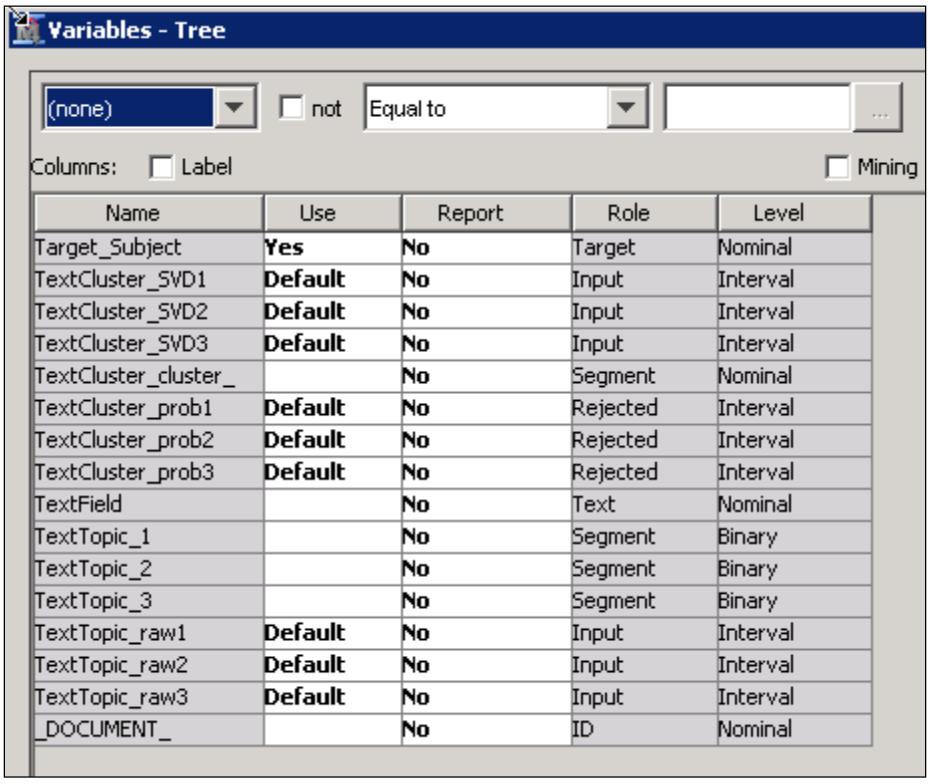
Sample Statistics

Obs #	Variable Name	Label
1	Target_Subject	Target_Subject
2	TextField	TextField
3	TextCluster_SVD1	
4	TextCluster_SVD2	
5	TextCluster_SVD3	
6	TextCluster_cluster_	
7	TextCluster_prob1	
8	TextCluster_prob2	
9	TextCluster_prob3	
10	TextTopic_1	_1_0_+snow,+cold,+winter,+weather,+hot
11	TextTopic_2	_1_0_baseball,+team,basketball,+play,+basketball team
12	TextTopic_3	_1_0_+lion,+tiger,+zoo,+animal,+zebra
13	TextTopic_raw1	+snow,+cold,+winter,+weather,+hot
14	TextTopic_raw2	baseball,+team,basketball,+play,+basketball team
15	TextTopic_raw3	+lion,+tiger,+zoo,+animal,+zebra
16	_DOCUMENT_	Document

TextTopic_raw1 - TextTopic_raw3 – These are numeric variables that indicate the strength a particular topic has within a given document. Three topics were generated because this was specified on the Property Sheet. These variables are the same as the topic weight values for the documents that were previously looked at in the Documents window of the interactive Topic Viewer. Each of these variables (topics) has a label (the five most descriptive terms) to identify it and help the user interpret the topic.

TextTopic_1 - TextTopic_3 – These are *binary* variables defined for each document and constructed from the **TextTopic_raw1 - TextTopic_raw3** values based on the document cutoff values described earlier. For example, **TextTopic_1** is set to **1** if a document has a **TextTopic_raw1** value greater than the cutoff value for this particular topic. Otherwise, it is set to 0. The labels for the **TextTopic** variables are the same as for the **TextTopic_raw** raw variables, except that they have _1_0_ as prefixes. This indicates that they are binary variables. Each label shows the five most descriptive terms that are identified with that topic.

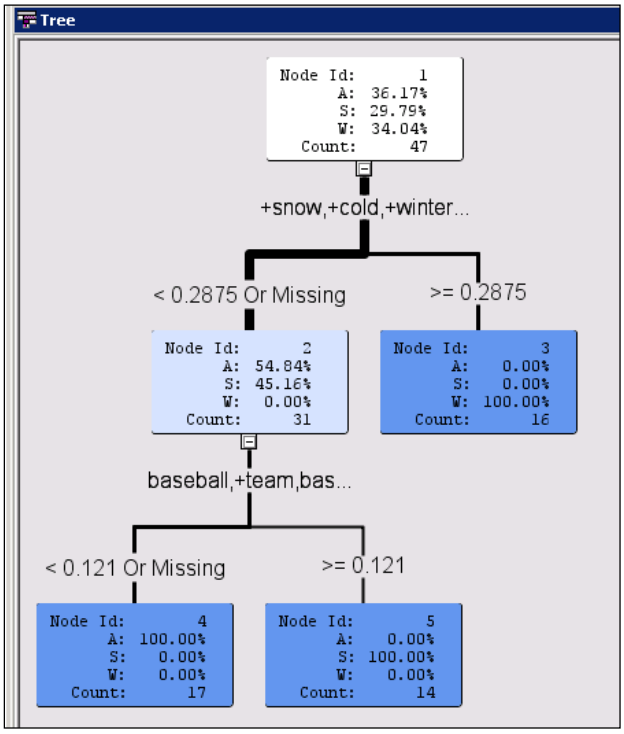
11. The emphasis in this class is on text mining for prediction (including supervised classification). To that end, continue this demonstration by attaching a **Decision Tree** node to the output of the **Text Topic** node. This node is found on the Model tab at the top. Among all model types, decision tree models are especially good for interpretation. After you attach the Decision Tree node but before running it with the default settings, go to the **Variables** ellipsis button in the Property Sheet to view the following window:



Name	Use	Report	Role	Level
Target_Subject	**Yes**	**No**	Target	Nominal
TextCluster_SVD1	**Default**	**No**	Input	Interval
TextCluster_SVD2	**Default**	**No**	Input	Interval
TextCluster_SVD3	**Default**	**No**	Input	Interval
TextCluster_cluster_		**No**	Segment	Nominal
TextCluster_prob1	**Default**	**No**	Rejected	Interval
TextCluster_prob2	**Default**	**No**	Rejected	Interval
TextCluster_prob3	**Default**	**No**	Rejected	Interval
TextField		**No**	Text	Nominal
TextTopic_1		**No**	Segment	Binary
TextTopic_2		**No**	Segment	Binary
TextTopic_3		**No**	Segment	Binary
TextTopic_raw1	**Default**	**No**	Input	Interval
TextTopic_raw2	**Default**	**No**	Input	Interval
TextTopic_raw3	**Default**	**No**	Input	Interval
DOCUMENT		**No**	ID	Nominal

By default, the only text mining variables that are considered as candidate prediction variables are those that have a role of **Input**. These are **TextCluster_SVD** and **TextTopic_raw** variables. Others, such as the **TextCluster_cluster_** or the **TextTopic** variables, which some analysts would consider using for prediction or classification, must be redefined as Input variables using the Metadata node.

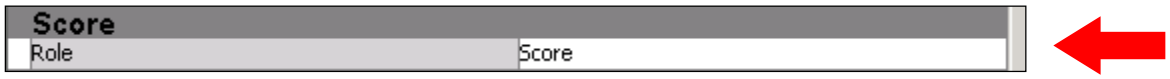
12. Run the default Decision Tree node. Open the results and maximize the Tree window.



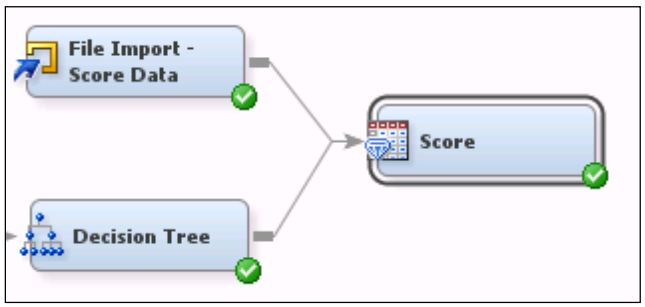
The decision tree resulted in three leaves. They are 100% accurate in classifying the documents as either A, S, or W. The variables used for this prediction or classification are the **TextTopic_1** and **TextTopic_2** text mining variables because the labels for these variables are displayed in the tree. The rules for the tree are obvious. The leaf with Node Id=3 comprises all documents where **TextTopic_1** (+snow, +cold, + winter …) is quite high, that is, >=.2875. Then Node 4 consists of documents where **TextTopic_1** is less than .2875 and **TextTopic_2** (baseball,+team, …) is less than .121. That is, Node 4, which has 100% animal documents, consists of documents that do not contain much information about either weather or sports. Finally, Node 5 is defined as consisting of documents that have **TextTopic_1** less than .2875 and **TextTopic_2** greater than .121. In other words, these are documents relatively high on the sports topic and low on the weather topic.

There are many approaches that an analyst can use to interpret the results of text mining. In this demonstration, the situation is easy to understand. In most realistic applications, you might need to do some creative analytic work to dig more deeply.

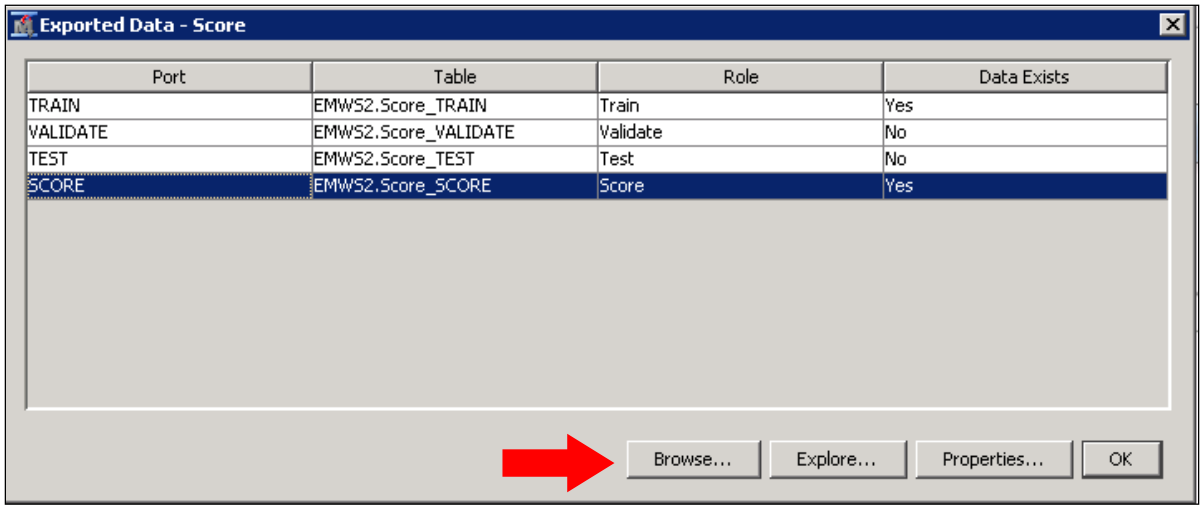
13. The final part of this demonstration is to use the Score node to score new data set. Following the top part of the diagram shown at the beginning of this demonstration, bring in a new **File Import** node. Rename it **File Import Score Data**. The import file for scoring is **D:\workshop\winsas\SGF15TA\Score_WeatherAnimalsSports.xls**. In the Property Sheet, change the role of the data set to **Score**.



Run the node and look at the Exported Data window. This **Score** data set has 16 documents. They are related to the three subjects (animals, sports, or weather). (As is usually the case with a data set to be scored, there is no target field on this data set.) The object now is to classify these documents using the Decision Tree model that was previously obtained on the training data, **WeatherAnimalsSports.xls**. To do that, bring in a **Score** node and connect it to the output of the **File Import Score Data** node and also to the output of the **Decision Tree** node.



Run the **Score** node. Then go to the Exported Data window through the Property Sheet. Select the **SCORE** data to view and click **Browse**.



14. When the Browse window appears, move. the column headings so that **TextField** is the first column heading and **Into: Target_Subject** is the second heading (See the display below.) Into: Target_Subject is the label for the variable **I_Target_Subject**. This variable is the predicted classification (A,S, or W) of **Target_Subject** based on the **TextTopic_1** and **TextTopic_2** variables used in the decision tree.

    Read through the 16 rows and check to see whether any of the classifications looks incorrect to you. Generally, all of them should look right except observation #14, which is about seasons and therefore is really a document about weather. However, it was incorrectly classified as A, an animal document. (Also, observation #16 is probably a better fit in A, but does fit in W, as it was used here.)

EMWS2.Score_SCORE

	TextField	Into: Target_Subject
1	We have a dog in our house. His name is Princey and he is a part of our family.	A
2	I spend a lot of time on the weekend watching sports shows: football, baseball, basketball and soccer are all fun for me.	S
3	The World Cup in soccer is held every four years. Brazil has had the winning team more than any other country.	S
4	The 2013 World Series was won by the Boston Red Sox team. They beat the St. Lous Cardinals in 6 games.	S
5	The winter weather has been very harsh in many parts of the U.S. during the months of January and February.	W
6	Yesterday, I watched a documentary about the big cats of Africa.	A
7	In our neighborhood, one man has 5 small dogs that he walks every day.	A
8	We have a problem with feral cats.	A
9	Professional basketball players are paid enormous salaries.	S
10	We have friends who live in a rural area and they sometimes see bears near their house.	A
11	We have had 5 inches of snow since this morning.	W
12	Rainy weather in the summer can be very pleasant to cool things off.	W
13	I could watch elephants all day. They are such beautiful, graceful animals.	A
14	I would like to spend summers in Maine and winters in Florida.	A
15	Leopards are nocturnal hunters.	A
16	I watched a nature movie about bears hibernating in winter and waking up in the springtime.	W

There is an endless number of reasons why the underlying text mining and modeling algorithms make mistakes. One possibility in this case is the very small number of training examples that were used.

1.3 Information Retrieval

Retrieving Medical Information

You often want to explore a document collection by searching on various terms of interest. This does not require a target variable and is efficiently done with the Text Filter node. As always, you first run the Text Parsing node. This demonstration illustrates how to do this using medical information from Medline data.

The **MEDLINES** data source contains a sample of 4,000 abstracts from medical research papers that stored in the MEDLINE data repository.

1. Create a new diagram and name it **Medline Information Retrieval**. Drag the **MEDLINES** data source into the diagram. Look at the variables.

Variables - Ids

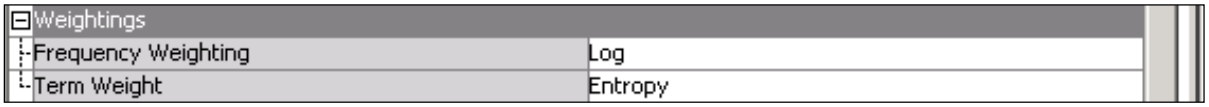
(none) not Equal to

Columns: Label

Name	Role	Level
ABSTRACT	**Text**	**Nominal**
AUTHOR	**Text**	**Nominal**
INDEX	**Input**	**Interval**
MEDLINEID	**Input**	**Interval**
MESHTERMS	**Text**	**Nominal**
PUBTYPE	**Input**	**Nominal**
SOURCE	**Rejected**	**Nominal**
TITLE	**Text**	**Nominal**

There is more than one variable with the role of **Text**. In cases like this, the Text variable with the longest length is the one that is selected for analysis by the Text Parsing node. If two or more Text variables have the same length, the one appearing first in alphabetical order is selected. In this example, **ABSTRACT** (2730 bytes in length) is the longest of the Text variables and is the one that is analyzed.

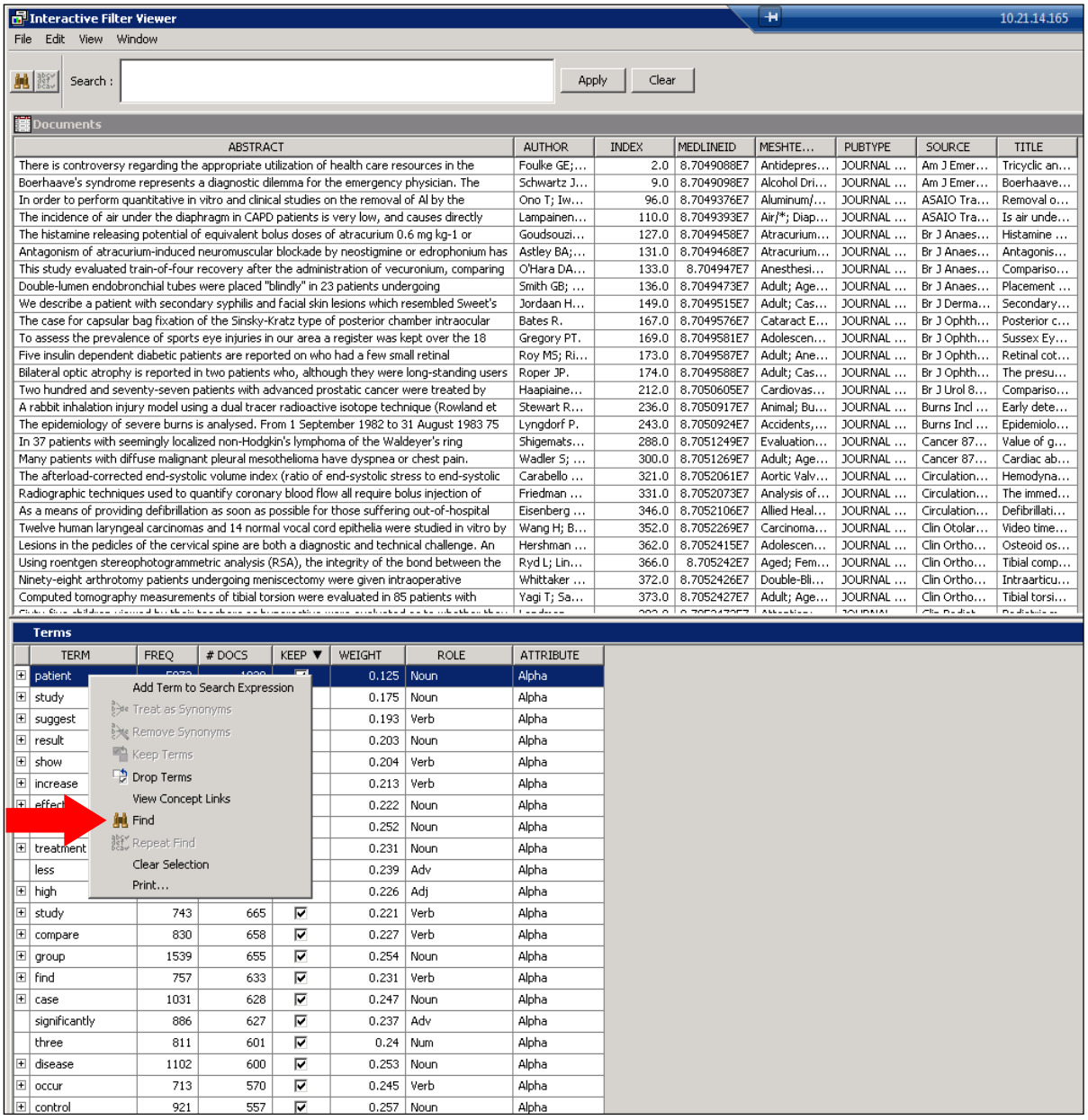
2. Attach a default **Text Parsing** node to the **Input Data Source** node. Notice that the default Text Parsing node populates the Properties Panel with certain tables. For example, there is a default Synonyms table called **SASHELP.ENGSYNMS**. (This actually contains only one row (one synonym) and is present only as a template). There is also a default stop list called **SASHELP.ENGSTOP**.
3. Attach a **Text Filter** node to the **Text Parsing** node. The default frequency weighting is **Log**. When there is no target variable, the default term weight is **Entropy**. It is a good idea to make this explicit, as shown.



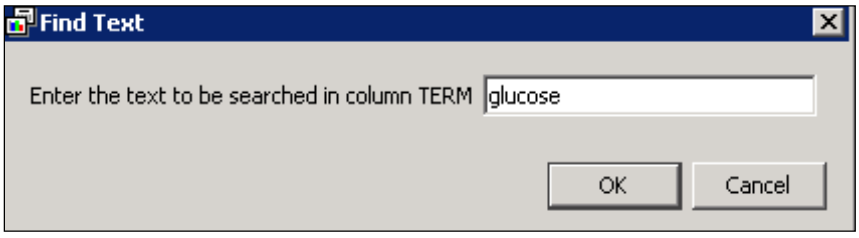
4. Run the default Text Parsing and Text Filter nodes.

5. Select **Filter Viewer** from the Properties Panel. This accesses the Interactive Filter Viewer where searching on terms in the documents is performed.

6. In the Terms window, right-click any term in the table. Select **Find**.



7. Enter **glucose** as the term to find.

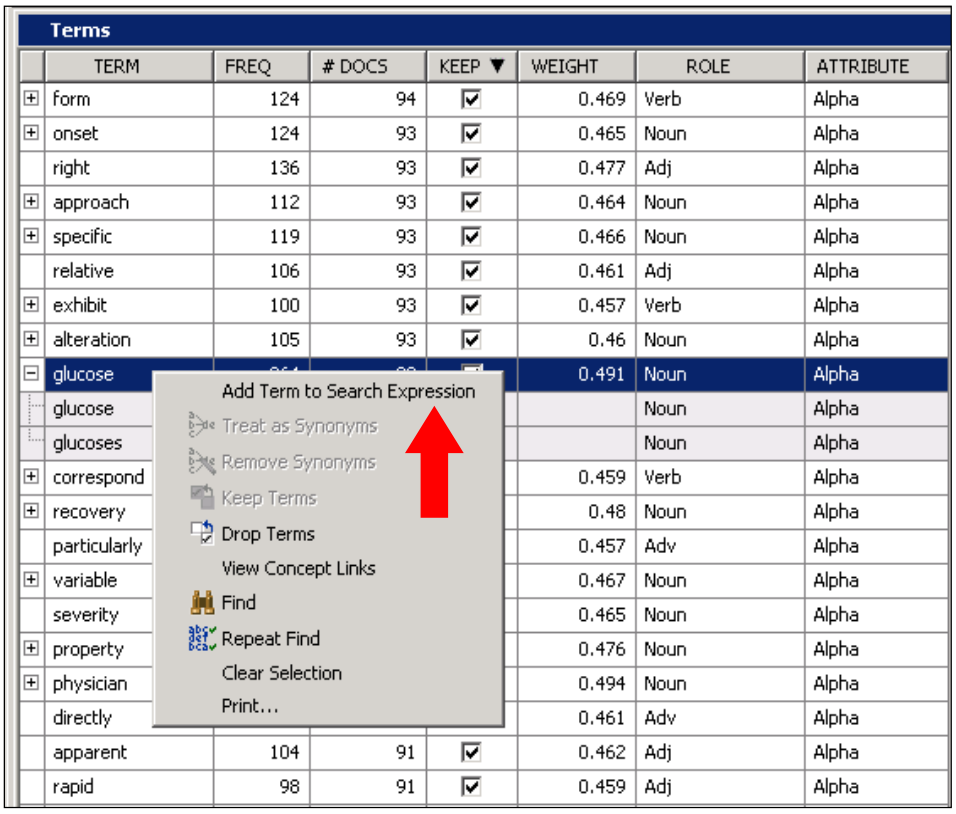


The table jumps to the portion of the table that contains the term **glucose**.

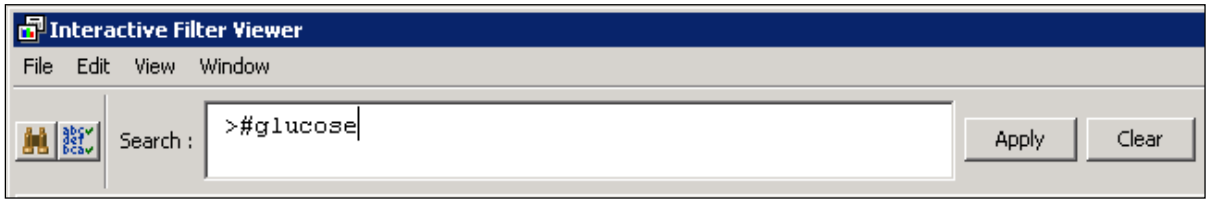
⊟	glucose	264	93	☑	0.491	Noun	Alpha
	glucose	263	93			Noun	Alpha
	glucoses	1	1			Noun	Alpha

Expand ⊞ to see the stemmed versions of **glucose**. It occurred 263 times in its singular form and one time as a plural.

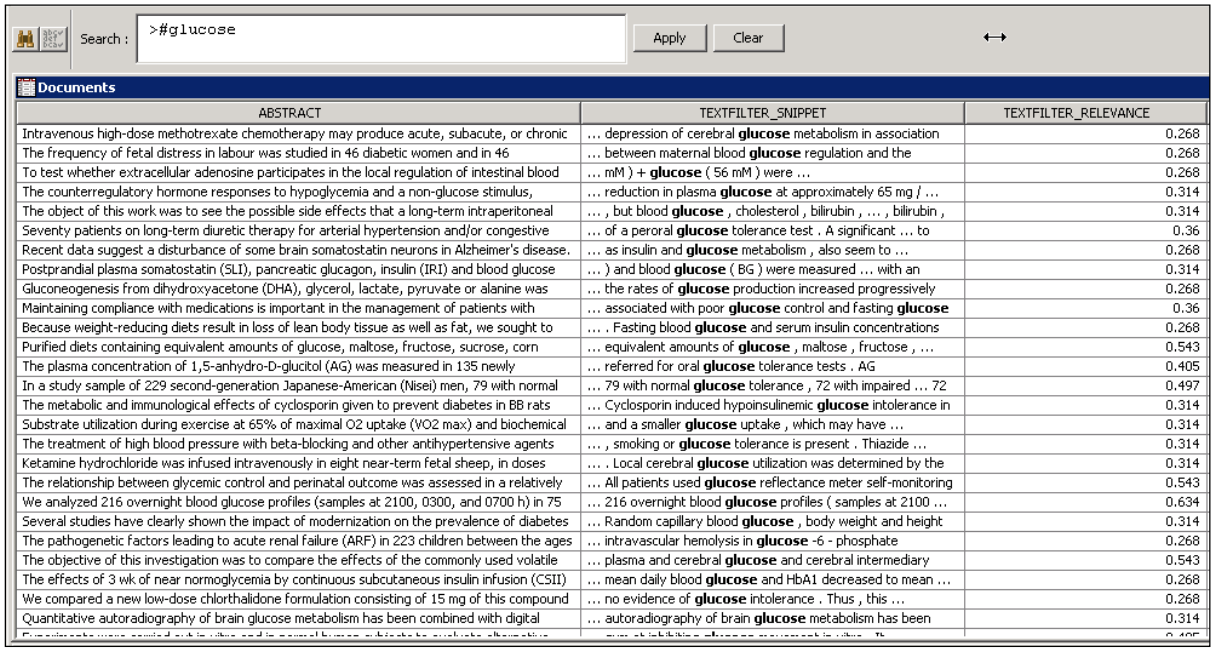
8. Right-click on the first row of **glucose** and select **Add Term to Search Expression**.



The term **glucose** is added to the Search window. The preceding symbols (>#) indicate that all stemmed versions (or synonyms if any were defined) of the term are searched for.



9. Click **Apply**. The following results appear in the **Documents** window:

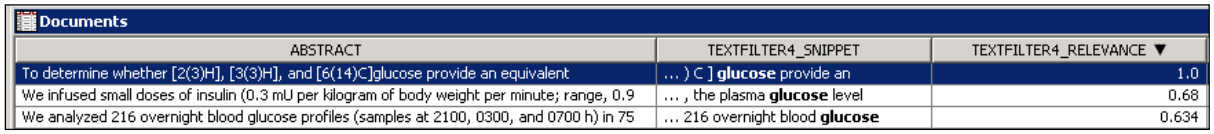


ABSTRACT	TEXTFILTER_SNIPPET	TEXTFILTER_RELEVANCE
Intravenous high-dose methotrexate chemotherapy may produce acute, subacute, or chronic	... depression of cerebral **glucose** metabolism in association	0.268
The frequency of fetal distress in labour was studied in 46 diabetic women and in 46	... between maternal blood **glucose** regulation and the	0.268
To test whether extracellular adenosine participates in the local regulation of intestinal blood	... mM) + **glucose** (56 mM) were ...	0.268
The counterregulatory hormone responses to hypoglycemia and a non-glucose stimulus,	... reduction in plasma **glucose** at approximately 65 mg / ...	0.314
The object of this work was to see the possible side effects that a long-term intraperitoneal	... , but blood **glucose** , cholesterol , bilirubin , ... , bilirubin ,	0.314
Seventy patients on long-term diuretic therapy for arterial hypertension and/or congestive	... of a peroral **glucose** tolerance test . A significant ... to	0.36
Recent data suggest a disturbance of some brain somatostatin neurons in Alzheimer's disease.	... as insulin and **glucose** metabolism , also seem to ...	0.268
Postprandial plasma somatostatin (SLI), pancreatic glucagon, insulin (IRI) and blood glucose	...) and blood **glucose** (BG) were measured ... with an	0.314
Gluconeogenesis from dihydroxyacetone (DHA), glycerol, lactate, pyruvate or alanine was	... the rates of **glucose** production increased progressively	0.268
Maintaining compliance with medications is important in the management of patients with	... associated with poor **glucose** control and fasting **glucose**	0.36
Because weight-reducing diets result in loss of lean body tissue as well as fat, we sought to	... . Fasting blood **glucose** and serum insulin concentrations	0.268
Purified diets containing equivalent amounts of glucose, maltose, fructose, sucrose, corn	... equivalent amounts of **glucose** , maltose , fructose , ...	0.543
The plasma concentration of 1,5-anhydro-D-glucitol (AG) was measured in 135 newly	... referred for oral **glucose** tolerance tests . AG	0.405
In a study sample of 229 second-generation Japanese-American (Nisei) men, 79 with normal	... 79 with normal **glucose** tolerance , 72 with impaired ... 72	0.497
The metabolic and immunological effects of cyclosporin given to prevent diabetes in BB rats	... Cyclosporin induced hypoinsulinemic **glucose** intolerance in	0.314
Substrate utilization during exercise at 65% of maximal O2 uptake (VO2 max) and biochemical	... and a smaller **glucose** uptake , which may have ...	0.314
The treatment of high blood pressure with beta-blocking and other antihypertensive agents	... , smoking or **glucose** tolerance is present . Thiazide ...	0.314
Ketamine hydrochloride was infused intravenously in eight near-term fetal sheep, in doses	... . Local cerebral **glucose** utilization was determined by the	0.314
The relationship between glycemic control and perinatal outcome was assessed in a relatively	... All patients used **glucose** reflectance meter self-monitoring	0.543
We analyzed 216 overnight blood glucose profiles (samples at 2100, 0300, and 0700 h) in 75	... 216 overnight blood **glucose** profiles (samples at 2100 ...	0.634
Several studies have clearly shown the impact of modernization on the prevalence of diabetes	... Random capillary blood **glucose** , body weight and height	0.314
The pathogenetic factors leading to acute renal failure (ARF) in 223 children between the ages	... intravascular hemolysis in **glucose** -6 - phosphate	0.268
The objective of this investigation was to compare the effects of the commonly used volatile	... plasma and cerebral **glucose** and cerebral intermediary	0.543
The effects of 3 wk of near normoglycemia by continuous subcutaneous insulin infusion (CSII)	... mean daily blood **glucose** and HbA1 decreased to mean ...	0.268
We compared a new low-dose chlorthalidone formulation consisting of 15 mg of this compound	... no evidence of **glucose** intolerance . Thus , this ...	0.268
Quantitative autoradiography of brain glucose metabolism has been combined with digital	... autoradiography of brain **glucose** metabolism has been	0.314

The abstract is shown on the left. Stretch the column labeled **TEXTFILTER_SNIPPET** so that you can see the term **glucose** in every row. This indicates the part of the abstract where **glucose** appears. (This is the first occurrence if there are multiple occurrences.)

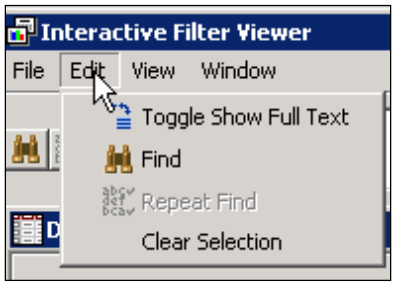
Place your mouse pointer above the **TEXTFILTER_SNIPPET** label. You see the following message: “Left-click on column header to sort 93 rows of the table.” This indicates that 93 documents were selected because either **glucose** or **glucoses** (or both) are found at least once in each document.

The **TEXTFILTER_RELEVANCE** column returns a measure of how strongly each document is associated with the search term. This is a relative measure. The most relevant document is given the highest value of 1. The calculation of this metric considers factors such as the number of times a term (or its stemmed versions and synonyms) appears in a document. To get an idea of this, click twice on the column heading for **TEXTFILTER_RELEVANCE** until you see the most relevant document in the first row (the one with **TEXTFILTER_RELEVANCE**=1.0). Then select that row.



ABSTRACT	TEXTFILTER4_SNIPPET	TEXTFILTER4_RELEVANCE ▼
To determine whether [2(3)H], [3(3)H], and [6(14)C]glucose provide an equivalent	...) C] **glucose** provide an	1.0
We infused small doses of insulin (0.3 mU per kilogram of body weight per minute; range, 0.9	... , the plasma **glucose** level	0.68
We analyzed 216 overnight blood glucose profiles (samples at 2100, 0300, and 0700 h) in 75	... 216 overnight blood **glucose**	0.634

Select **Edit** ⇨ **Toggle Show Full Text** to see the complete document with the highest relevance score.

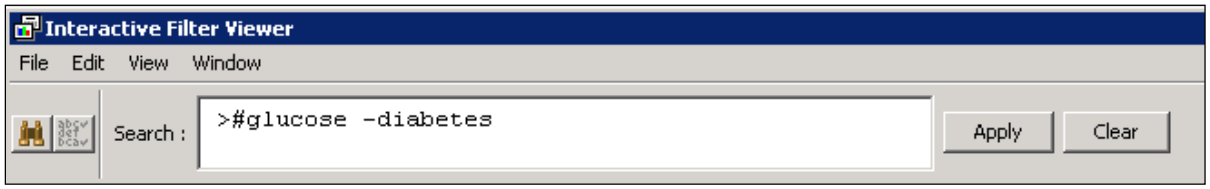


The full text for this abstract can be read.

ABSTRACT	TEXTFILTER4_SNIPPET
To determine whether [2(3)H], [3(3)H], and [6(14)C]glucose provide an equivalent assessment of glucose turnover in insulin-dependent diabetes mellitus (IDDM) and nondiabetic man, glucose utilization rates were measured using a simultaneous infusion of these isotopes before and during hyperinsulinemic euglycemic clamps. In the nondiabetic subjects, glucose turnover rates determined with [6(14)C]glucose during insulin infusion were lower (P less than 0.02) than those determined with [2(3)H]glucose and higher (P less than 0.01) than those determined with [3(3)H]glucose. In IDDM, glucose turnover rates measured with [6(14)C]glucose during insulin infusion were lower (P less than 0.05) than those determined with [2(3)H]glucose, but were not different from those determined with [3(3)H]glucose. All three isotopes indicated the presence of insulin resistance. However, using [3(3)H]glucose led to the erroneous conclusion that glucose utilization was not significantly decreased at high insulin concentrations in the diabetic patients. [6(14)C] and [3(3)H]glucose but not [2(3)H]glucose indicated impairment in insulin-induced suppression of glucose production. These results indicate that tritiated isotopes do not necessarily equally reflect the pattern of glucose metabolism in diabetic and nondiabetic man.	...) C] **glucose** provide an equivalent assessment of ... equivalent assessment of **glucose** turnover in insulin-dependent diabetes mellitus ... nondiabetic man , **glucose** utilization rates were measured using ...

Reading through the full document, it is obvious that **glucose** is used many times. This explains why this document has the highest relevance measure for a query based on this term. Select **Edit** ⇨ **Toggle Show Full Text** to go back to the original way of viewing the documents.

10. Ninety-three documents were retrieved by the query. It is also useful to be able to retrieve documents that contain one term, but do ***not*** contain another term. For example, in order to take these 93 documents and eliminate any that contain the term **diabetes**, in the Search window, enter **–diabetes**. (That is, precede the term with a minus sign as shown below.) Select **Apply**.



Be sure to separate the two terms with a space.

Verify that 72 documents of the original 93 remain after you eliminate any documents with the term **diabetes**.

1.4 Text Categorization

Categorizing ASRS Documents

This demonstration illustrates some document categorization using SAS Text Miner.

The ASRS data set can be accessed from the following link: **http://asrs.arc.nasa.gov/**

From the website:

“ASRS captures confidential reports, analyzes the resulting aviation safety data, and disseminates vital information to the aviation community.”

"More than 850,000 reports have been submitted (through October 2009) and no reporter's identity has ever been breached by the ASRS. ASRS de-identifies reports before entering them into the incident database. All personal and organizational names are removed. Dates, times, and related information, which could be used to infer an identity, are either generalized or eliminated."

As with other data sets used in this course, data sets derived from ASRS have been modified. The original data for this demonstration was extracted from the ASRS, pre-processed, and provided to competitors in a text mining competition sponsored by SIAM and the NASA Ames Research Center. The competition results were presented at the Seventh SIAM International Conference on Data Mining held in 2007 in Minneapolis, Minnesota. Participants were prohibited from using the R language, SAS software, and most commercial software.

A link that provides access to the original data follows. **https://c3.nasa.gov/dashlink/resources/138/**

A single report in the ASRS database can be a composite derivation of two or more reports filed for the same incident. For example, one runway incursion incident can result in three reports: one from the pilot, one from the copilot, and one from an air traffic controller. An incident involving two or more aircraft can have reports filed from pilots of all aircraft involved, as well as from air traffic controllers. In both examples, there will be only one ASRS report, but that report will be prepared by NASA professionals based on all reports submitted.

Reports can be submitted by aviation professionals, such as pilots, flight attendants, and mechanics. Reports can also be submitted by non-professionals, such as private pilots.

A report in the ASRS database has many fields, with one field representing a primary narrative describing the incident. This primary narrative is stored in the **Text** variable. All of the other fields have been omitted to simplify the text mining component of the analysis. In practice, an automated labeling system would attempt to use all fields.

NASA manually assigns to each report 1 or more of 54 anomalies, 1 or more of 32 results, 1 or more of 16 contributing factors, and 1 or more of 17 primary problems. For example, the report might describe an event that was a "runway ground incursion" anomaly, with a "took evasive action" result, that was a "human factor" contributing factor, and a "human factor" primary problem. These fields are not available in the contest data. Instead, the contest data has 22 labels, with a value of 1 "if document i has label j;" Otherwise, the label has a value of -1. Labels correspond to the topics identified by NASA to aid in the analysis of the reports. The labels are not defined in the competition. For the course data, the 22 labels are named **Target01**, **Target02** up through **Target22**, and an original coding of (-1,1) has been changed to (0,1), with a code of 1, which indicates the presence of the label in the document. A document can be associated with one or more labels.

The ASRS training data we will use contains columns that indicate which of the 22 manually assigned labels relates to a given report. The goal is to develop a system to automatically detect incidents to avoid the time, cost, and error associated with manually labeling the reports. In other words, we will be building a model on a data set where experts have already read the reports and made evaluations. (This is the sort of process that many people will use for sentiment analysis: create a data set of labeled cases and then build an automatic classification/prediction system based on these known cases.) In an actual operational system for this example, we would build 22 models to evaluate whether each of these 22 types of events occurred. This collection of models would provide 22 predicted values, one for each target.

The 22 target events vary considerably with respect to the difficulty of modeling them. Descriptions of several of these target events (as given in E.G. Allan et al 2008), along with published ROC index values for models that Allan et al obtained using an analytic method known as Nonnegative Matrix Factorization, are shown in the table below.

Some of the 22 ASRS target events with their ROC index values from published model results:

Event Label in Course Data	Description of Event	Reported ROC Index From Allan et al Model Results
Target02	Operation Noncompliance	.6009
Target05	Incursion (collision hazard)	.8977
Targer13	Weather Issue	.6287
Targer21	Illness or Injury event	.8201
Target22	Security concern / threat	.9040

1. Create a new diagram and name it **ASRS Text Categorization**.
2. Create a data source for the ASRS data using **SGF15TA.ASRS**.
   Use the following metadata:

Name	Role	Level
ID	**ID**	**Nominal**
Size	**Rejected**	**Interval**
Target01	**Rejected**	**Binary**
Target02	**Rejected**	**Binary**
Target03	**Rejected**	**Binary**
Target04	**Rejected**	**Binary**
Target05	**Target**	**Binary**
Target06	**Rejected**	**Binary**
Target07	**Rejected**	**Binary**
Target08	**Rejected**	**Binary**
Target09	**Rejected**	**Binary**
Target10	**Rejected**	**Binary**
Target11	**Rejected**	**Binary**
Target12	**Rejected**	**Binary**
Target13	**Rejected**	**Binary**
Target14	**Rejected**	**Binary**
Target15	**Rejected**	**Binary**
Target16	**Rejected**	**Binary**
Target17	**Rejected**	**Binary**
Target18	**Rejected**	**Binary**
Target19	**Rejected**	**Binary**
Target20	**Rejected**	**Binary**
Target21	**Rejected**	**Binary**
Target22	**Rejected**	**Binary**
Text	**Text**	**Nominal**

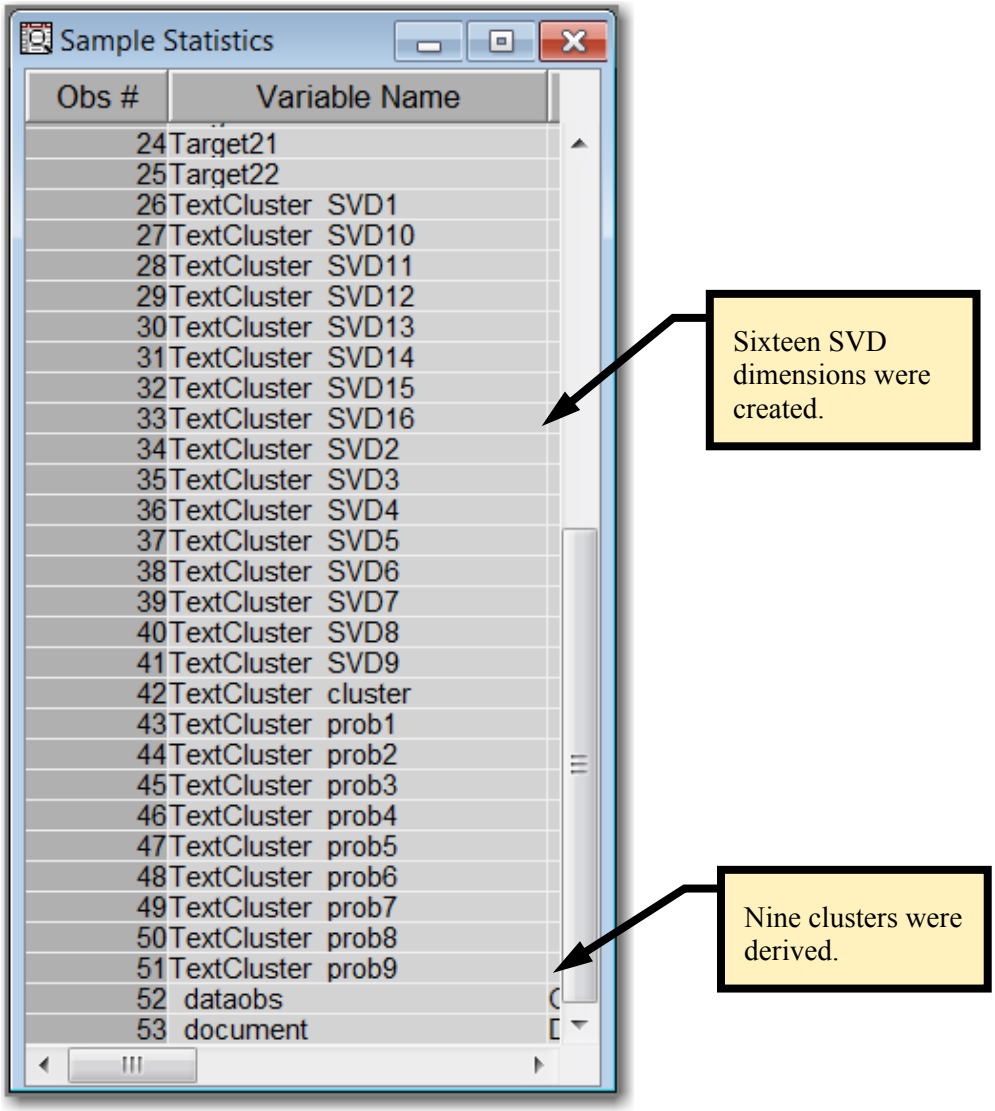
The data set contains 22 variables with the names **Target01** through**Target22**. We will use **Target05** for this demonstration, so all the others should be rejected. From a table in the E.G. Allan et al article "Anomaly Detection Using Nonnegative Matrix Factorization" (2008, p. 215), the incident for **Target05** has to do with the occurrence of a collision hazard event. The variable **Size** is just the length of the report in bytes and will not be used. Only the report itself (**Text**) is needed, but **ID** can be left as an ID variable.)

3. Drag the **ASRS** data source onto the diagram.
4. To investigate the robustness of the automated assignment, add a **Data Partition** node to the partition data set **SGF15TA.ASRS**. Use a 75/25/0 partition.
5. Attach a **Text Parsing** node to the **Data Partition** node. Leave all the defaults as is and run the node.
6. Attach a **Text Filter** node to the **Text Parsing** node. Change the default weightings to **Log** and **Mutual Information**. (Although in this case, these are the defaults.) Leave all else in default mode and run.

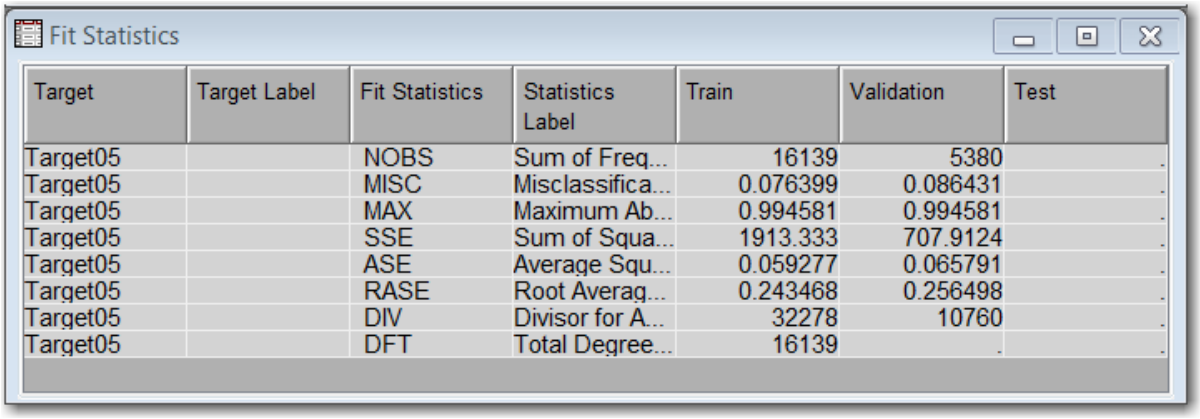
Weightings	
Frequency Weighting	Log
Term Weight	Mutual Information

7. Attach a **Text Cluster** node to **Text Parsing.** Use the default settings. Run the Text Cluster node.
8. This generates a 16-dimensional SVD solution. Go to **Exported Data** in the **Property Sheet**. Select the **TRAIN** data set and then click the **Explore** button. Verify that you have a 5-dimensional solution by looking at the number of **TextCluster_SVD** variables.

SVD=Singular Value Decomposition. In general, SVD provides the algorithmic component of **Latent Semantic Analysis** (LSA). A 16-dimensional SVD solution implies that each document will be assigned a 16-dimensional vector, which translates to the SAS data set observation associated with a document will have 16 additional numeric variables associated withthe document.



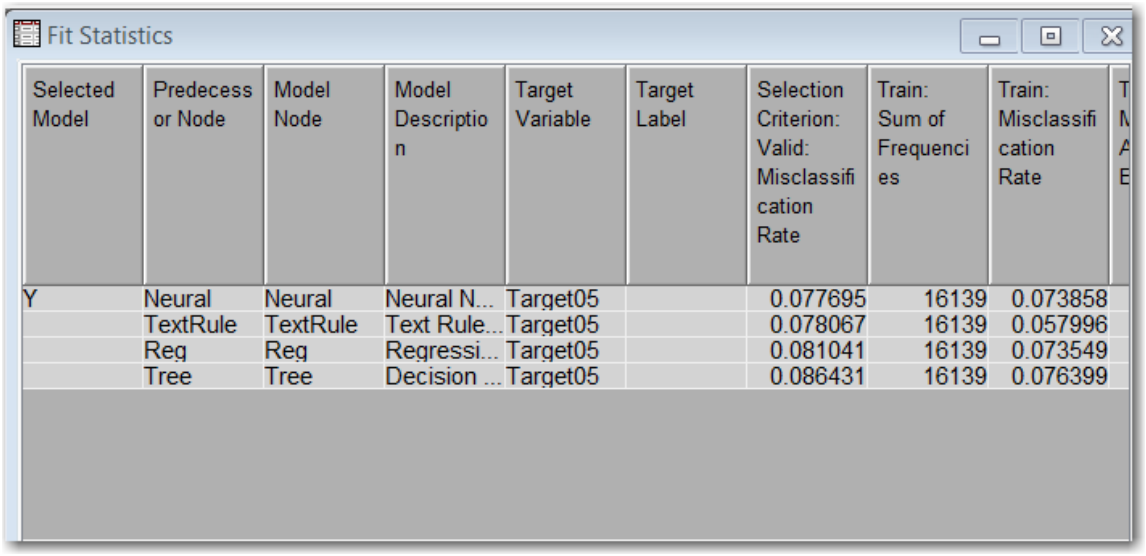
9. Attach a **Decision Tree** node to the **Text Cluster** node. Change the Assessment Measure property to **Average Square Error** and Leaf Size to **25**. (These are fairly routine changes that are often found to produce better results with trees. Obviously 25 is not some magic number, but the default Leaf Size of 5 is considered by many analysts to be too small.) Run the node, and open the results window. The Fit Statistics table follows.



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Target05		NOBS	Sum of Freq...	16139	5380	.
Target05		MISC	Misclassifica...	0.076399	0.086431	.
Target05		MAX	Maximum Ab...	0.994581	0.994581	.
Target05		SSE	Sum of Squa...	1913.333	707.9124	.
Target05		ASE	Average Squ...	0.059277	0.065791	.
Target05		RASE	Root Averag...	0.243468	0.256498	.
Target05		DIV	Divisor for A...	32278	10760	.
Target05		DFT	Total Degree...	16139	.	.

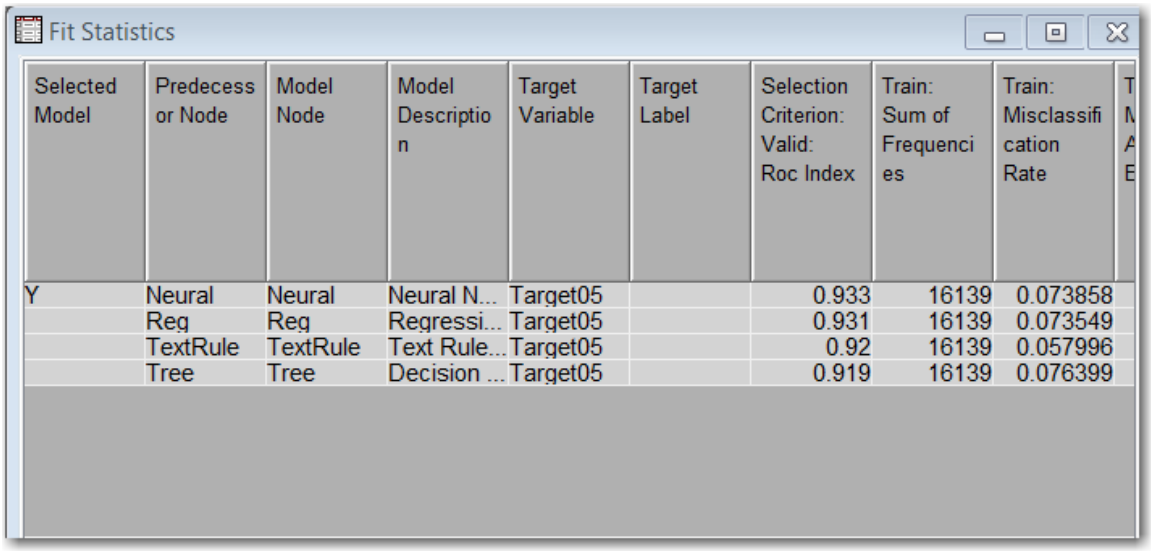
The decision tree exhibits an 8.6% misclassification rate for the validation data. If you apply Text Mining and use the 16 SVD variables as inputs for a predictive model, a decision tree model can be expected to correctly categorize 91.4% of the documents into the Target05 category (occurrence of a collision hazard event).

You can use the Text Rule Builder node as an alternative to decision tree modeling. This node provides a text mining predictive modeling solution within SAS Text Miner. To compare models in Enterprise Miner, use the Model Comparison node in the Assess tab. A summary of a comparison of a decision tree, a Text Rule Builder solution, a logistic regression model, and a neural network model are provided below.


Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	T M A E
Y	Neural	Neural	Neural N...	Target05		0.077695	16139	0.073858	
	TextRule	TextRule	Text Rule...	Target05		0.078067	16139	0.057996	
	Reg	Reg	Regressi...	Target05		0.081041	16139	0.073549	
	Tree	Tree	Decision ...	Target05		0.086431	16139	0.076399	

The Text Rule Builder is competitive with the other prediction tools, but the node can only use text input. To combine text inputs with other input variables, you must venture outside of the Text Mining tab. Allan, et al, used the C statistic for comparison. The table below shows that results depend on the accuracy statistic chosen. The C statistic for the neural network is 0.9330, which beats the value achieved by Allan, et al, which was 0.8977.


Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Roc Index	Train: Sum of Frequencies	Train: Misclassification Rate	T M A E
Y	Neural	Neural	Neural N...	Target05		0.933	16139	0.073858	
	Reg	Reg	Regressi...	Target05		0.931	16139	0.073549	
	TextRule	TextRule	Text Rule...	Target05		0.92	16139	0.057996	
	Tree	Tree	Decision ...	Target05		0.919	16139	0.076399	

This demonstration shows that document categorization with a well-defined target variable is just a special case of predictive modeling.

1.5 Enhancing a Bigger Project

Enhancing Text Mining with Custom Topics and Profiles

This demonstration illustrates how custom topics can be used to provide an alternative to predictive modeling for document categorization.

Use the **Movies** data set.

1. Create a diagram called **Movies**.
2. Create a data source using the Movies2015 data set in the SGF15TA library, **SGF15TA.Movies2015**.

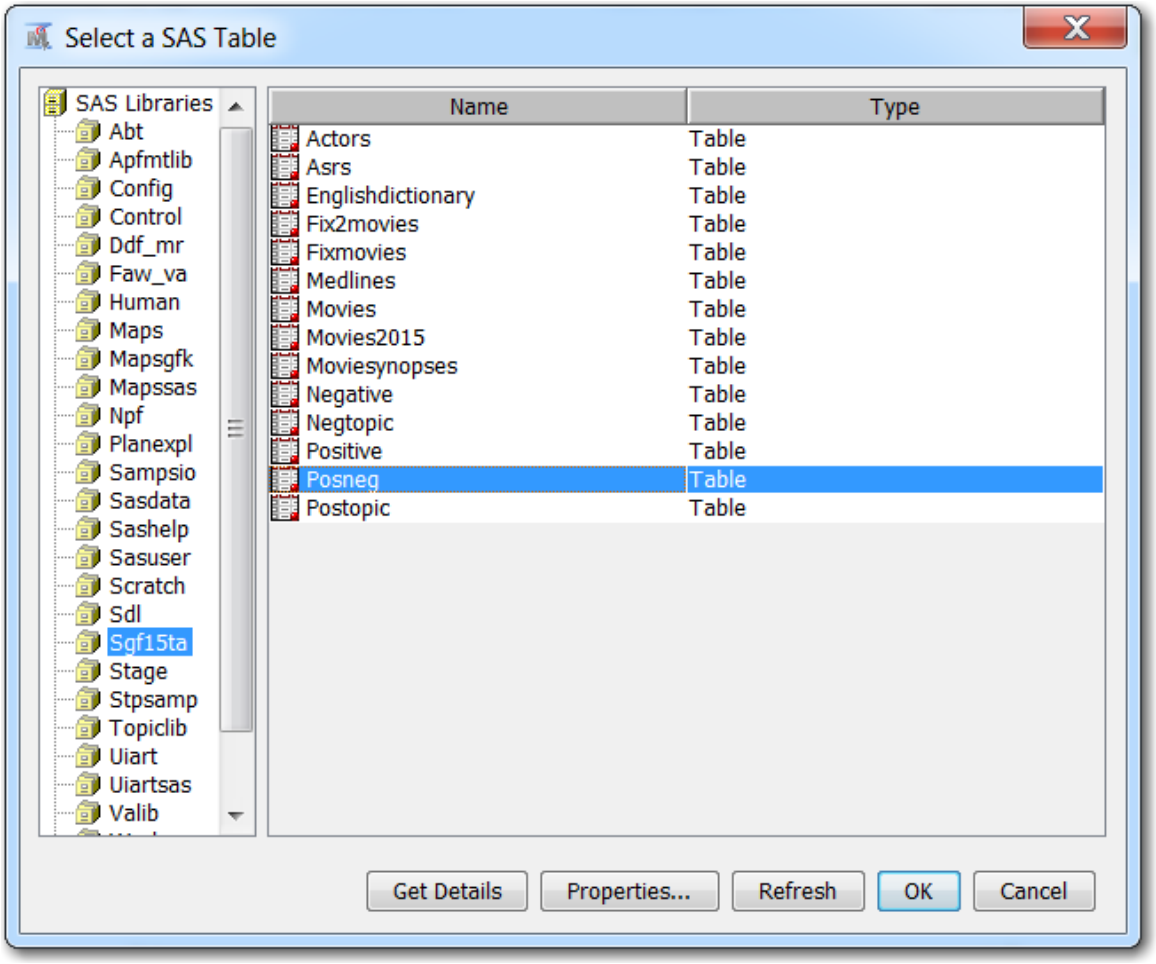
   The Metadata is given in the following table.

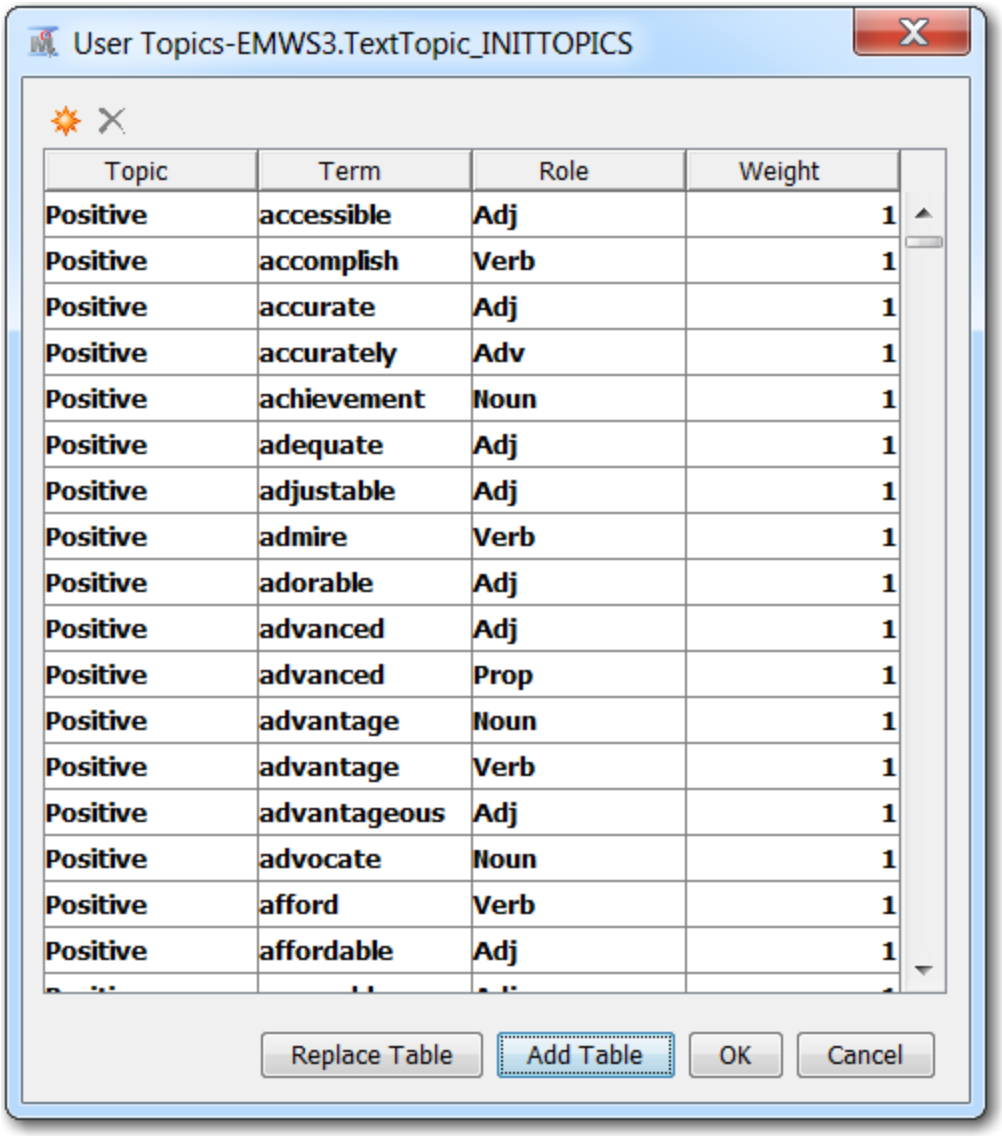
Name	Role	Level
Comedy	**Target**	**Binary**
Genre	**Input**	**Nominal**
MPAARating	**Input**	**Nominal**
Name	**Text**	**Nominal**
Size	**Input**	**Interval**
SortName	**ID**	**Nominal**
Synopsis	**Text**	**Nominal**
Title	**Text**	**Nominal**
ViewerRating	**Input**	**Nominal**
Year	**Input**	**Interval**

   When multiple text variables are present, the Text Parsing node will pick the variable with the largest length.

3. Drag the Movies2015 data source into the Movies diagram.
4. Attach a Data Partition node to the data source node. Change the Train/Validate/Test proportions to 75/25/0. Run the Data Partition node.
5. Attach a Text Parsing node. Use the default settings. Run the text Parsing node.
6. Attach a Text Filter node. Use the default settings. Run the Text Filter node.
7. Attach a Text Cluster node. Change Exact or Maxim Number to Exact. Change the Number of Clusters to 10. Run the node.

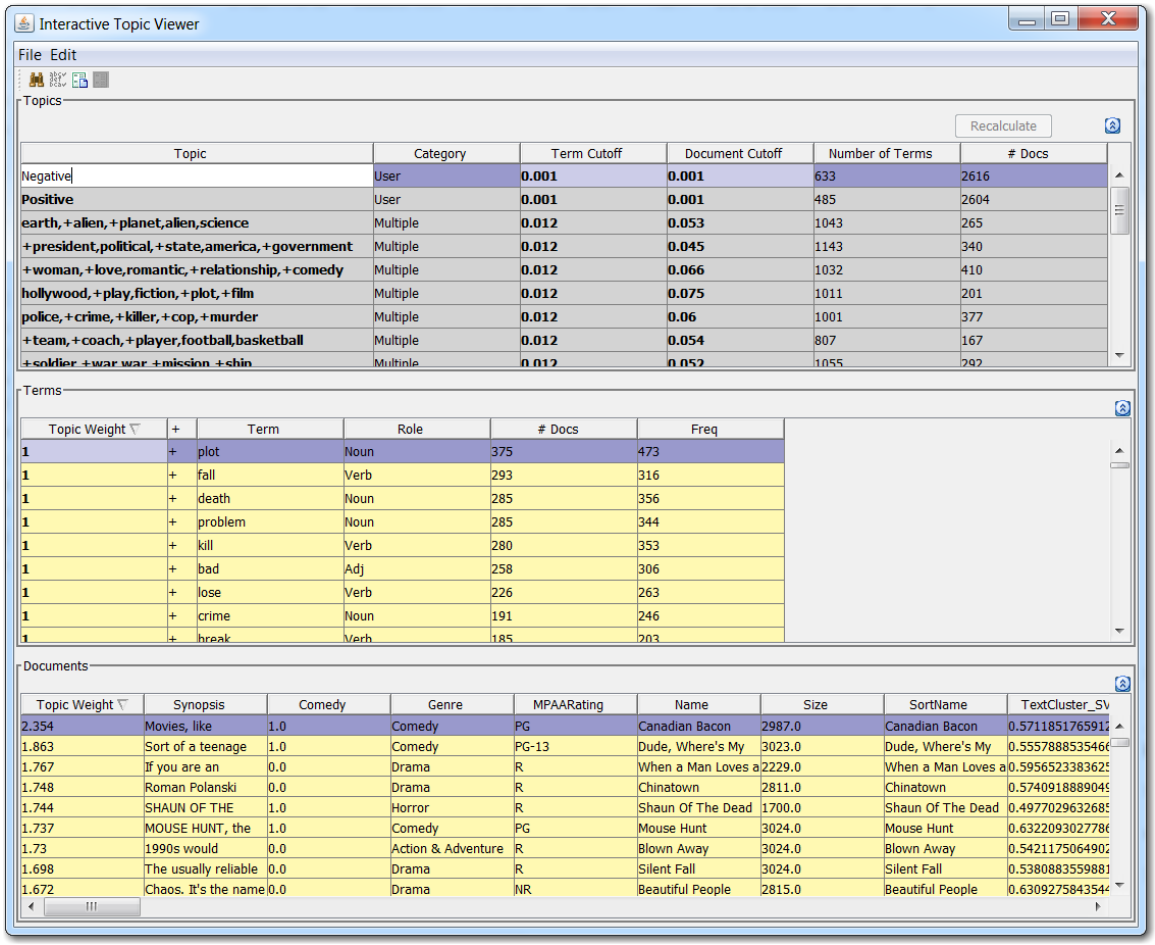
8. Attach a Text Topic node. Select **User Topics**. Select **Add Table**. Select the **SGF15TA.POSNEG** table.



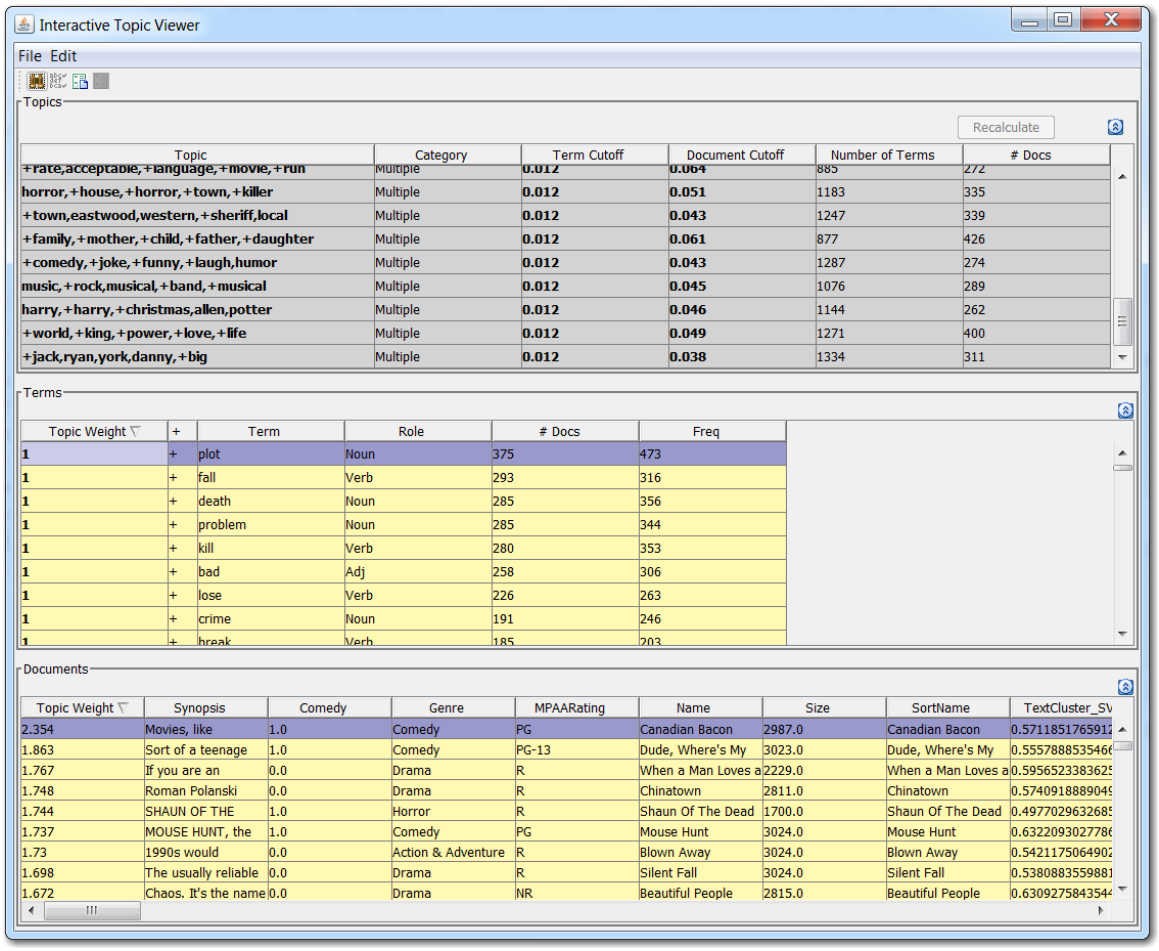


Click **OK**. Run the Text Topic node.

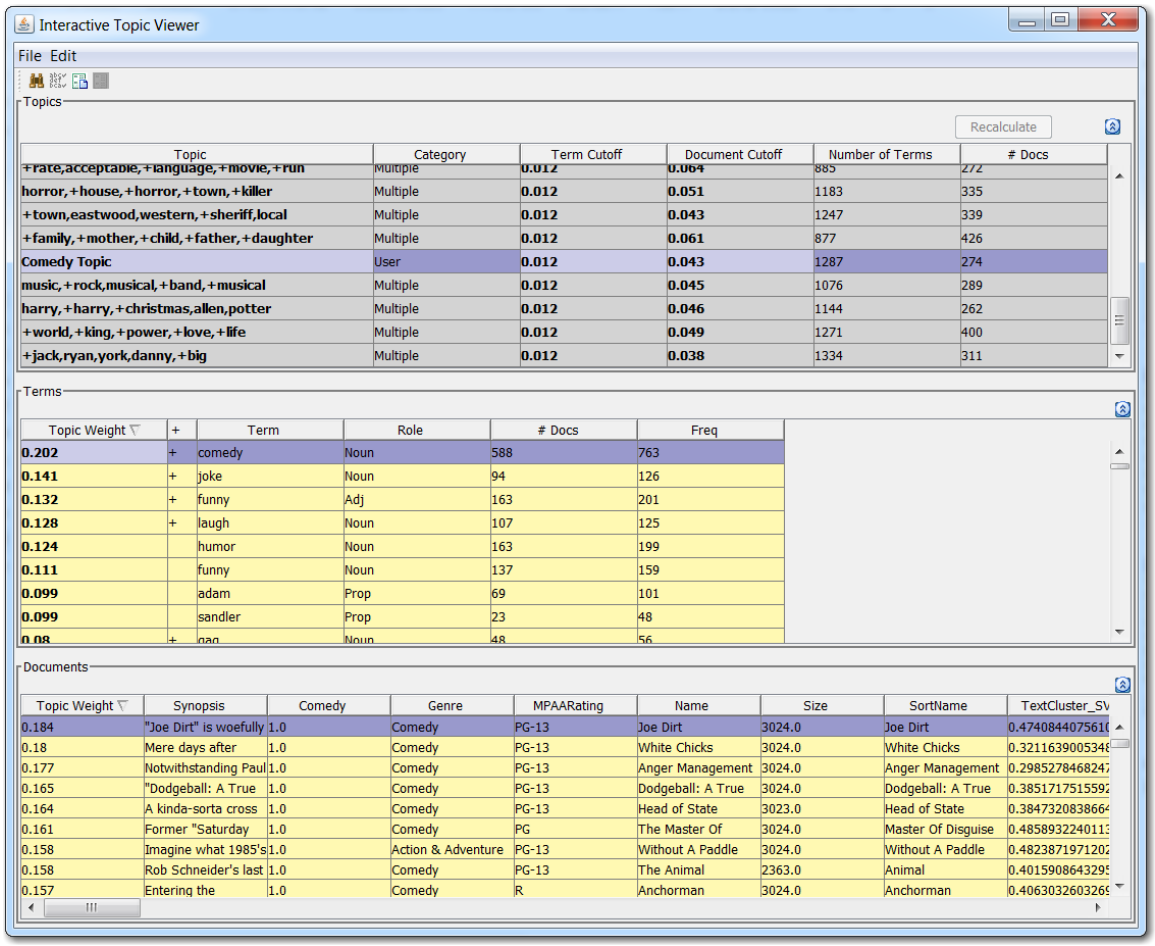
9. Open the Topic Viewer.



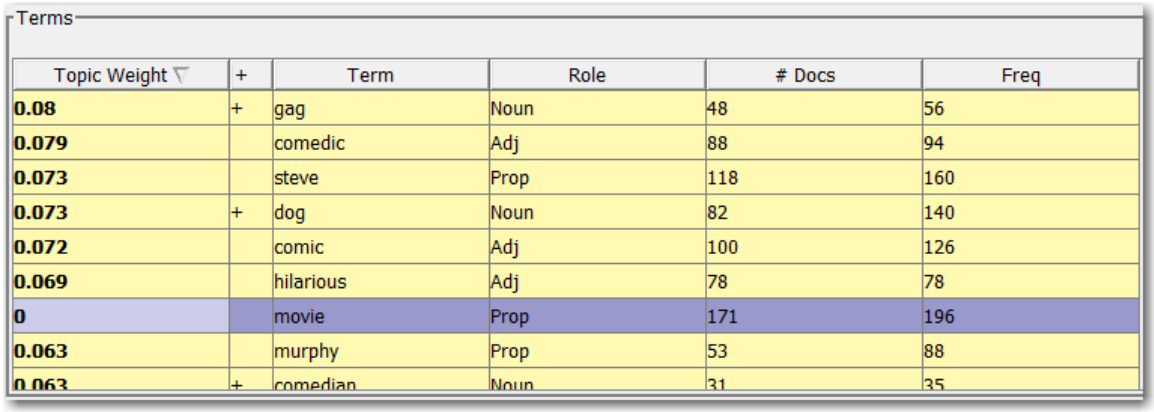
The two custom topics, Negative and Positive, are at the top of the Topic table. Scroll down in the Topic table to the last five topics.



The derived topic with keywords +comedy, +joke, +funny, +laugh, and humor appears to be a comedy topic. This possibly represents a solution to the document categorization problem without using a target variable, also known as unsupervised learning. Select the topic by right-clicking and selecting Select Current Topic. Change the name to Comedy Topic.

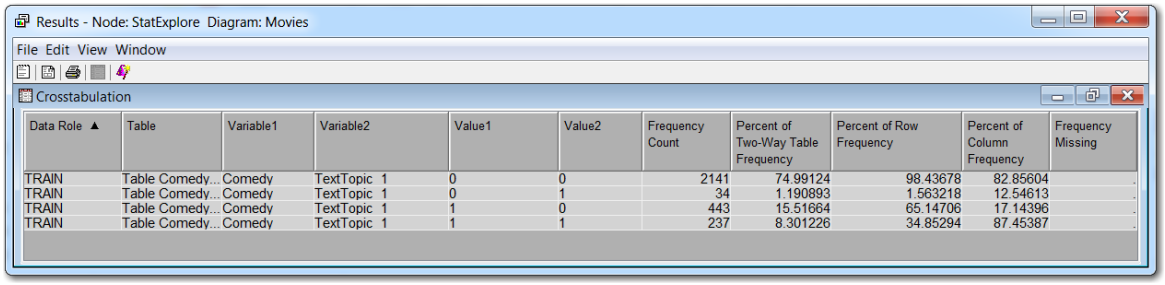


Note that any editing of a topic causes the Category to change from Multiple to User. Scroll down to the term movie and change the Topic Weight to 0 (zero).



Click the Recalculate button. The Document weights will have changed very slightly. Note that for the Comedy target variable, all values listed in the table are equal to one (1). This suggests that perhaps the Comedy Topic is correctly categorizing the movies with respect to being a comedy movie. To validate this, you should calculate precision and recall for the Comedy Topic. Doing so is beyond the scope of this presentation, but a SAS program is included that can perform the calculations. To calculate the values by hand, use the Stat Explore node to get a crosstabulation table.

10. Close the Topic Viewer and save the results. Run the node again to update the topics based on the new added custom topic. You can check the Custom Topic table to verify that Comedy Topic has been added.
11. Attach a Metadata node and change the roles of TextTopic_1 and Comedy to Input. The Stat Explore node only performs crosstabulations for input variables. Run the Metadata node.
12. Attach a Stat Explore node to the Metadata node. Select **Cross-Tabulation** and choose variables Comedy and TextTopic_1. Run the node.



Data Role ▲	Table	Variable1	Variable2	Value1	Value2	Frequency Count	Percent of Two-Way Table Frequency	Percent of Row Frequency	Percent of Column Frequency	Frequency Missing
TRAIN	Table Comedy...	Comedy	TextTopic_1	0	0	2141	74.99124	98.43678	82.85604	
TRAIN	Table Comedy...	Comedy	TextTopic_1	0	1	34	1.190893	1.563218	12.54613	
TRAIN	Table Comedy...	Comedy	TextTopic_1	1	0	443	15.51664	65.14706	17.14396	
TRAIN	Table Comedy...	Comedy	TextTopic_1	1	1	237	8.301226	34.85294	87.45387	

Precision=237/(237+34)=87.5%; Recall=237/(237+443)=34.9%. The TextTopic_1 variable is derived using a cutoff of 0.043 for the TextTopic_raw1 variable. Changing the cutoff to 0.019 produces precision and recall values of about 62% with a misclassification rate of about 18%.

13. Attach a Decision Tree node to the Text Topic node. Change the Leaf Size to 25 and the Assessment Measure to Average Squared Error. Select the Variables property and set the Use status of Genre to "No." Run the node.
14. Open the results window. The fit statistics reveal a misclassification rate of 14.3%. Using a breakeven point that yields approximately the same precision and recall, the misclassification rate is about 17% with precision and recall around 66%. [These values can be obtained using a SAS code node and a precision/recall break even program.]

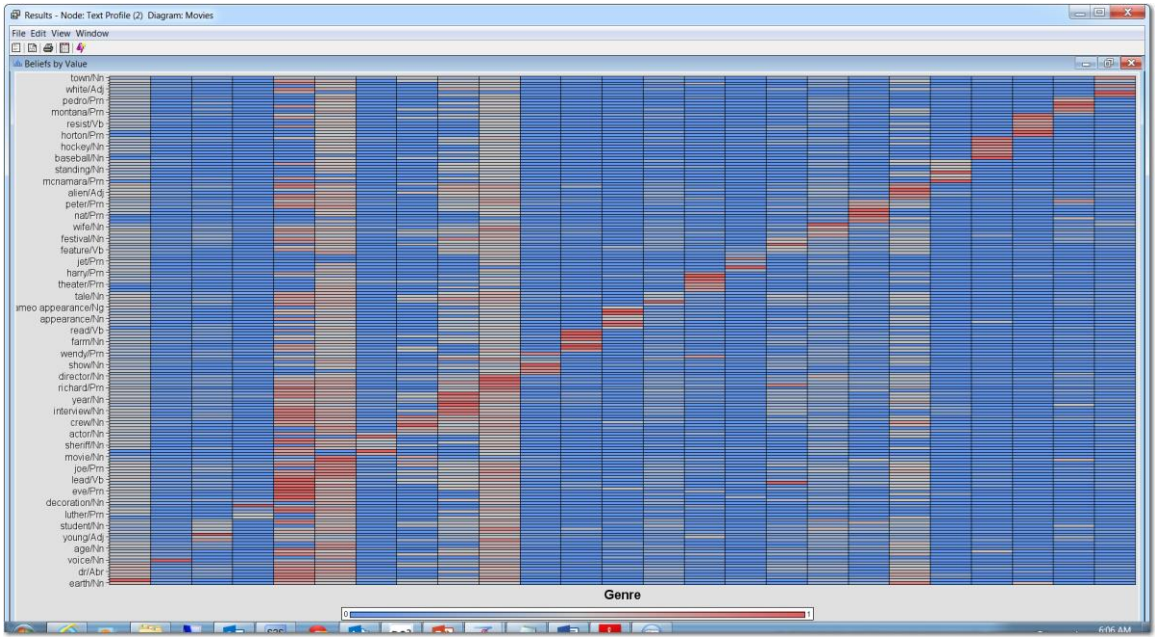
Suppose you want to profile every genre based on the synopsis of each movie.

1. Attach a Metadata node to the Text Filter node. Change the role of Genre to target and the role of Comedy to rejected. Run the node.
2. Attach a Text Profile node to the Metadata node. Select the Variables property and make the following changes.

Name	Use	Report	Role	Level
Genre	**Yes**	**No**	Target	Nominal
Name	**No**	**No**	Text	Nominal
Synopsis	**Yes**	**No**	Text	Nominal
Title	**No**	**No**	Text	Nominal

Run the node.

3. Open the results. Expand the Beliefs by Value window.



Examine the Profiled Variables window.

Profiled Variables

Name	Value	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8	Freq	Corpus	Genre	Var/Var Level ID	Terms
Corpus		mcnama...	likely/Adj	abyss/Nn	standing/...	secretary...	cameo a...	taylor/Prn	close/Nn	2855	1		0	mcnama...
Genre	Action & ...	action/Nn	scott/Prn	john/Prn	earth/Nn	alien/Nn	know/Vb	dr/Abr	james/Prn	451		Action & ...	1	action/Nn...
Genre	Animation	voice/Nn	dr/Abr	jonah/Nn	age/Nn	pal/Nn	summer/...	human/Adj	dream/Nn	70		Animation	2	voice/Nn,...
Genre	Art Hous...	boy/Nn	drama/Nn	student/Nn	brother/Nn	young/Adj	begin/Vb	woman/Nn	sex/Nn	37		Art Hous...	3	boy/Nn, d...
Genre	Christmas	christma...	luther/Prn	nora/Prn	know/Vb	decoratio...	blair/Prn	holiday/Nn	receive/Vb	1		Christmas	4	christma...
Genre	Classic	hollywoo...	kong/Prn	eve/Prn	lead/Vb	music/Nn	win/Vb	blair/Prn	spectacl...	38		Classic	5	hollywoo...
Genre	Comedy	comedy/Nn	movie/Nn	david/Prn	character...	joe/Prn	scene/Nn	mom/Nn	thing/Nn	586		Comedy	6	comedy/...
Genre	Crime	corleone/...	stallone/...	vito/Prn	godfather...	sheriff/Nn	actor/Nn	cop/Nn	pacino/Prn	3		Crime	7	corleone/...
Genre	Cult	jack/Nn	john/Nn	crew/Nn	movie/Nn	work/Nn	scene/Nn	proceed/Vb	research...	7		Cult	8	jack/Nn, j...
Genre	Docume...	look/Nn	filmmake...	people/Nn	interview/...	footage/Nn	history/Nn	year/Nn	america/...	138		Docume...	9	look/Nn, f...
Genre	Drama	star/Vb	mother/Nn	girl/Nn	director/Nn	drama/Nn	husband/...	woman/Nn	richard/Prn	1035		Drama	10	star/Vb, ...
Genre	Family	show/Nn	wendy/Prn	garfield/Prn	andre/Prn	ryan/Prn	dad/Nn	family/Nn	child/Nn	5		Family	11	show/Nn,...
Genre	Foreign	sun/Nn	erik/Prn	ellen/Prn	book/Nn	american...	read/Vb	novel/Nn	farm/Nn	3		Foreign	12	sun/Nn, e...
Genre	Gay & Le...	taylor/Prn	guy/Vb	four/Num	vegas/Prn	cameo a...	cameo/Nn	gay/Adj	appearan...	2		Gay & Le...	13	taylor/Prn...
Genre	Horror	house/Nn	home/Nn	begin/Vb	family/Nn	student/Nn	group/Nn	popular/Adj	tale/Nn	73		Horror	14	house/Nn...
Genre	Kids & F...	series/Nn	cartoon/Nn	original/Nn	child/Nn	dickens/...	theater/Prn	harry/Prn	guthrie/Prn	16		Kids & F...	15	series/Nn...
Genre	Martial-Arts	chan/Prn	li/Prn	jackie/Prn	kong/Prn	eddie/Prn	lethal/Prn	jet/Prn	jet/Nn	2		Martial-Arts	16	chan/Prn...
Genre	Musical ...	music/Nn	video/Nn	girl/Nn	festival/Nn	title/Nn	feature/Vb	star/Nn	tale/Nn	30		Musical ...	17	music/Nn...
Genre	Mystery ...	man/Nn	daughter/...	find/Vb	begin/Vb	director/Nn	woman/Nn	wife/Nn	return/Vb	205		Mystery ...	18	man/Nn, ...
Genre	Romance	soccer/Nn	nat/Prn	keith/Prn	eli/Prn	quot/Nn	allen/Prn	peter/Prn	love/Nn	17		Romance	19	soccer/N...
Genre	Science ...	world/Nn	earth/Nn	alien/Adj	crew/Nn	back/Adv	planet/Nn	productio...	war/Nn	69		Science ...	20	world/Nn,...
Genre	Special I...	mcnama...	likely/Adj	secretary...	war/Prn	abyss/Nn	standing/...	cameron/...	close/Nn	2		Special I...	21	mcnama...
Genre	Sports & ...	baseball/...	team/Nn	fan/Nn	mandela/...	hockey/Nn	doug/Prn	yearn/Vb	jimmy/Nn	6		Sports & ...	22	baseball/...
Genre	Television	mama/Nn	help/Nn	horton/Prn	terry/Prn	evans/Prn	alien/Vb	resist/Vb	intention/...	6		Television	23	mama/N...
Genre	Unknown	germany/...	marie/Prn	truth/Nn	love/Nn	montana/...	jane/Nn	comedy/Nn	pedro/Prn	18		Unknown	24	germany/...
Genre	Western	western/Nn	wayne/Prn	western/...	town/Nn	hero/Nn	man/Nn	van/Prn	white/Adj	35		Western	25	western/...

The 8 term profile for Comedy is: comedy/Nn, movie/Nn, david/Prn, character/Nn, joe/Prn, scene/Nn, mom/Nn, and thing/Nn.

The Text Profile node can profile any nominal target variable by deriving keywords that are associated with each nominal category.

You can experiment with different properties and settings to try to improve the profiles.

1.6 References

Text Analytics Using SAS Text Miner Course Notes. Cary, North Carolina: SAS Institute Inc. 2014.

E. Allan, M. Horvath, C. Kopek, B. Lamb, T. Whaples and M. Berry. Anomaly detection using nonnegative matrix factorization. In M. Berry and M. Castellanos, Editors, *Survey of Text Mining II: Clustering, Classification, and Retrieval,* pages 203-218. Springer-Verlag London Limited, 2008.

Chakraborty, Goutam, Murali Pagolu, and Satish Garla. 2013. Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS. Cary, North Carolina: SAS Institute Inc.

Evangelopoulos, Nicholas, Xiaoni Zhang, and Victor R. Prybutok. 2010. "Latent Semantic Analysis: Five methodological recommendations." European Journal of Information Systems. 1-17.