

**Demo with SAS GF 1976-1980 abstracts**



## Text Parsing and Filtering

---

### Case Study: Analyzing SAS Global Forum Paper Abstracts

The data used in this demonstration are part of a case study included in a SAS Press book titled *Text Analytics and Sentiment Mining*. The data are abstracts of all SAS conference papers published each year at SUGI/SAS Global Forum from 1976 to 2011. Initially, we considered three types of textual data for text mining:

- title of the paper
- abstract of the paper
- complete body of the paper

The title of the paper might not be a good input because it is often restricted in length and might not fully reflect the theme of the paper semantically. Considering the complete paper for analysis will likely add a lot of noise because a full paper can include tables, images, references, SAS programs, and so on, which are problematic and might not add much value in text mining for topic extraction. We felt that analyzing the abstract of a paper to be most appropriate because it captures the detailed objective of a paper and does not contain extraneous items such as tables and images. We are thankful to SAS for making the SUGI/Global Forum Proceedings available for all the years the conference was held. We faced many challenges, starting from downloading papers in .pdf format from the SAS website to preparing a final data set that contains only the abstract for each paper. We used the %TMFILTER macro for preparing SAS data sets from a repository of SAS papers in .pdf and .txt format. We had to make some strategic choices to prepare the data sets. For complete details on the data preparation process, refer to the 2012 SAS Global Forum paper “SAS Since 1976: An application of text mining to reveal trends.”

For this case study, we start with the same four data sets used in the 2012 Global Forum paper with only one exception: the data set for the latest decade includes papers from the 2012 SAS Global Forum as well.

There are four data sets used in the case study published in the book. *But, for this demonstration, we will only use the first data set:*

- **sas1976\_1980**
- **sas1981\_1990**
- **sas1991\_2000**
- **sas2000\_2012**

The following sections take you through the detailed step-by-step text analytics process for text parsing and working with interactive filter.

1. Start a new SAS Enterprise Miner project or you can continue with an existing project.
2. Right-click **Diagrams** in the project panel and select **New Diagram**. Give it a suitable name (such as **SAS GF 76-80**). Click **OK**.
3. From the top menu, select **File** ⇒ **New** ⇒ **Library**.
4. In Step 1 of 3 Select Action, click **Next**.
5. In Step 2 of 3 Create or Modify, enter a library name, such as **Course**. Click **Browse** and point to the folder where the data are located. Click **Next**. Then click **Finish**.

**Note:** The path should point to a folder where course data are located.

6. Right-click the **Data Sources** icon in the project panel and select **Create Data Source**.
7. Click **Next>**.
8. In Step 2 of the Data Source Wizard, click **Browse**.
9. Double-click on the library **Course**.
10. Select the SAS data set **Sas1976\_1980**.
11. Click **OK**.
12. In Step 2 of the Data Source Wizard, click **Next**.

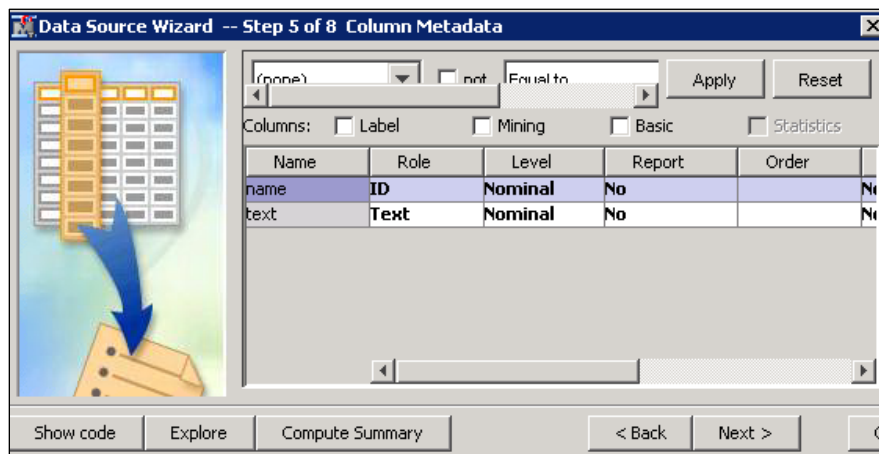
**Note:** The box shows that the two-part naming convention (*library-name.data-set-name*) is used in all SAS environments.

13. In Step 3 of the Data Source Wizard, click **Next**.
14. In Step 4 of the Data Source Wizard, click **Next**.



You should now see a window named Data Source Wizard-Step 5 of 8 Column Metadata.

In this step, you assign analysis roles to different fields/variables in the data. The field **Name** needs to be changed to a role of **ID** for this analysis.

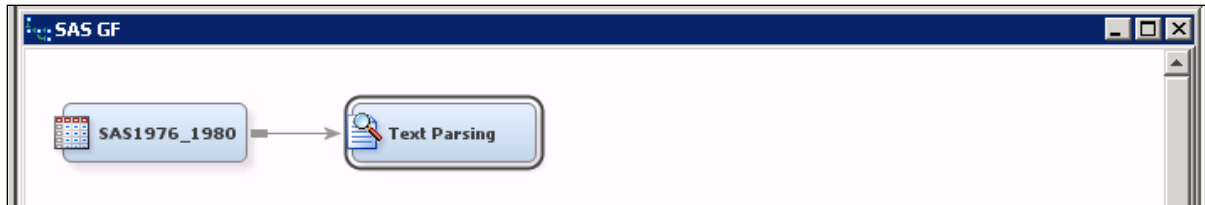
15. Click on the **Role** column for the field **Name** and select **ID** from the drop-down menu. Then click **Next**.



16. Click **Next** in Step 6.
17. Step 7 of the Data Source Wizard appears. Click **Next**.
18. Click **Finish** to complete Step 8 of the Data Source Wizard.

If needed, click  beside Data Sources in the project panel to change it to . You will see the newly added data under the Data sources.

19. Drag the data source **SAS1976\_1980** to the diagram workspace.
20. Connect a **Text Parsing** node to the **Data Source** node as shown below.



21. Right-click and rename the Text Parsing node as **Parsing (Default)** node. Then right-click and run the **Parsing (Default)** node with default options. Examine the results.

Note that the Text Parsing node has been run with all default options. (You should take a look at those default options in the properties panel.) The Results window has six panes with different types of information and graphs. Each pane has useful information. In this demonstration, we selectively explain what is in some of these panes.

22. Maximize the Terms window.

This window contains a list of all terms extracted with term properties such as role, attribute, frequency, Keep/Drop status, parent/child status, and the rank assigned to the term. A great deal of understanding about what is happening for the settings being used in the nodes can be made by looking at the Terms table. When you have a sense of which terms are being kept or dropped and their role, you can then adjust the settings accordingly.

Let us explore how we can make use of the entity extraction property in the Parsing node. This property has a default value of *None* (except that **SAS Institute**, by default, is always identified as an entity). Let's see what happens if we enable entity detection.

23. Close the Results window.
24. Connect another **Text Parsing** node to the **Data Source** node
25. In the properties panel of the newly connected Text Parsing node, change the Find Entities value to **Standard** as shown below.

Train	
Variables	...
Parse	
Parse Variable	text
Language	English ...
Detect	
Different Parts of Speech	Yes
Noun Groups	Yes
Multi-word Terms	SASHELP.ENG_MULTI ...
Find Entities	Standard
Custom Entities	

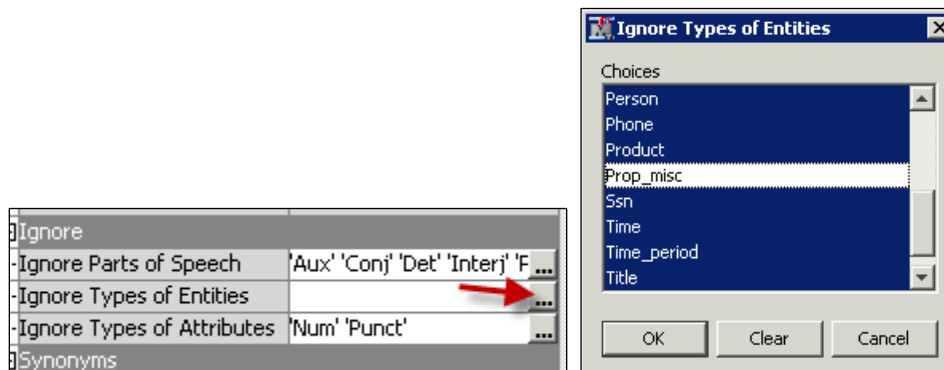
26. Then right-click rename it as **Text Parsing (custom)** node and right-click and run this node. Examine the results.

Compare the different entities shown in the *Role by Freq plot* compared to the same plot from the Parsing (default) node.

In this case study, we find names of authors, location, company, and address appearing in the text as entities due to limitations of the data preparation task. Most of these entities were parsed

as proper nouns when the entity property was disabled. These terms can be generally considered as noise in our analysis because our interest is in ***the trend of topics***.

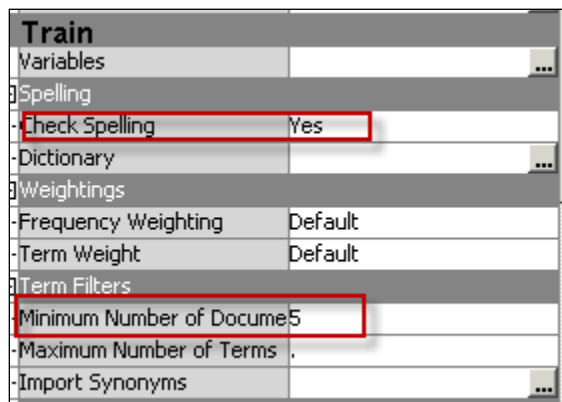
27. **Close** the Results window.
28. Then click the ellipsis button next to **Ignore Types of Entities** and select all terms except **Miscellaneous Proper Noun (Prop\_misc)** as shown below. With this setting, all entities except **Prop\_misc** will be ignored in the analysis.



29. Some parts of speech are automatically ignored in the default settings. You can modify those by clicking the **Ignore Parts of Speech** property ellipsis button. For this analysis, control-click and add **Abbr**, **Pref** and **Num** in the Ignore Parts of Speech list as shown below.

*When a synonyms list is already available, you can use it with the Text Parsing node. Otherwise, these lists can be custom generated for these data using the Text Filter node. If start or stop lists are available, they can be used in text parsing as well.*

30. Right-click and run the **Text Parsing** node. Do **not** view the results.
31. Connect a **Text Filter** node to the **Text Parsing** node. Enable the spell-checking property and set the Minimum Number of Documents property to **5** as shown below.



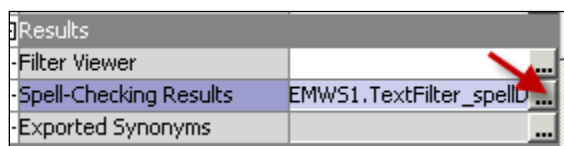
With this setting, SAS checks for spelling errors and corrects them using the default dictionary. SAS also ignores all terms that have appeared in fewer than five documents.

32. Right-click the **Text Filter** node and select **Run**. Examine the results.

This Results window has a lot of information, but often the interactive filter viewer provides a more convenient way to find specific information or to understand the terms.

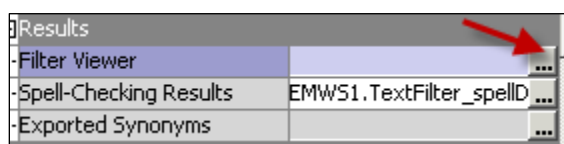
33. Close the Results window.

34. Click the ellipsis button for the **Spell Checking Results** property.



This shows the results for spell-checking. *As an analyst, you will often have to dig into these spell-checking results and then modify your dictionary and/or create stop lists or synonyms by considering commonly misspelled terms to improve your analysis.*

35. **Close** the spell-checking results. Click the ellipsis button for the **Filter Viewer** property. This accesses the interactive filter viewer.



36. In the Terms window, **click** the **TERM** column heading. *This sorts the term table by the **TERM** value and makes it easier to scroll and find terms. This makes it easy to scroll and find terms that may be of interest to you.*

37. You can also find specific terms that you might be interested in. Right-click anywhere in the Terms table. Select **Find**.

38. Type **regression** as the term to find.

The table should jump to the portion of the table that contains the term **regression** (if needed, use the scroll bar to navigate through the table).

39. **Click** the plus sign in front of the term regression to see all of the terms that are deemed equivalent to regression in this analysis. **Click** the minus sign in front of regression to collapse all equivalent forms of regression.

40. Right-click **regression** and select **Add Term to Search Expression**.

The term is added to the search box at top of the window.

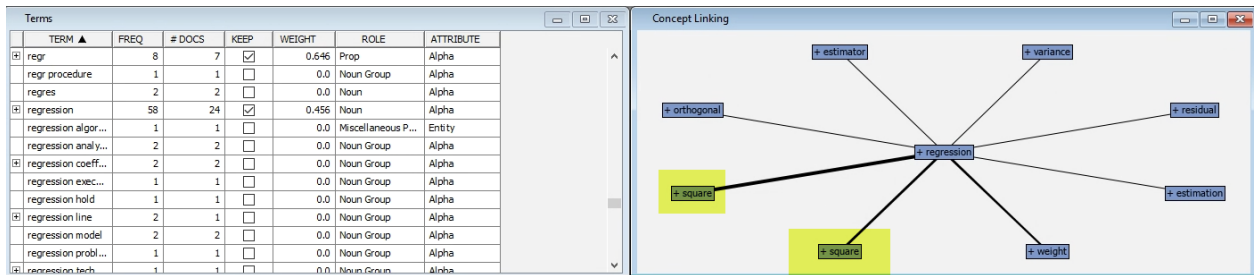
41. Click **Apply**. You will find the results in the viewer window:

A number of documents are returned by the filter. It seems the term **regression (or, its variants)** appeared in each of those documents. It is possible that sometimes documents are returned that do not contain the search terms because that document contains terms that are likely to co-occur with the search term. A relevance score is assigned to each document representing the similarity of the document to the query. Documents with relevance scores above an appropriate cutoff value are returned. This means that a document that does not contain the term in the search query might still be selected.

42. To see the entire document, right-click the first entry and select **Toggle Show Full Text**.

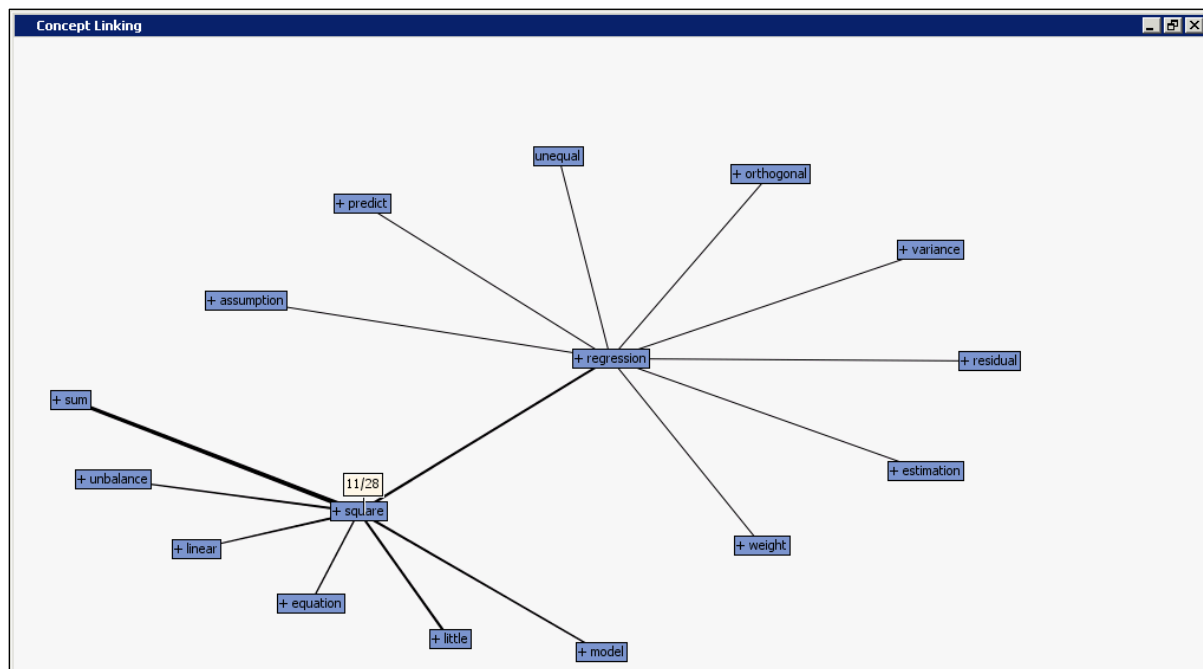
43. Right-click and select **Toggle Show Full Text** again to get back to the smaller view of documents.

44. Right-click the term **+regression** and select **view concept links**. The concept link plot appears.



The plot shows that regression is strongly associated with square and estimation in these documents (as indicated by thicker lines). If you hover your mouse over the term **(+regression)**, you will find it has appeared in 24 documents. Note also that the term square appeared twice in this plot because **Different Parts of Speech** was enabled. This might not be what you want in the analysis, in which case you should change the properties settings.

45. Close the interactive filter results.
46. Change **Different Parts of Speech** to **No** in the Text Parsing node.
47. Run the flow from the Text Filter node.
48. Use the interactive filter viewer and look at a concept link for the term **+regression** again.
49. Double-click the term **+square** to expand the concept links to show all terms associated with **square**. Positioning the cursor over a term shows the number of documents in which the term **+square** co-appears with the central term **+regression**. In this case, the term **square** and its variants have appeared in 28 documents, and 11 of those documents also contain the term **regression** and its variants.



The simplest way to filter words is often to select a number of terms from the Terms window and clear the Keep column. From the interactive filter window, try sorting the Terms table multiple times for each different column to understand groups of related terms. For example,

sorting the TERM column helps in identifying synonyms that are very close in spelling, such as the terms **airline**, **aviation**, and **aircraft**. In the next few steps, we demonstrate how we can drop some high-frequency terms and then create a synonym list from terms that are similar in meaning in the context of this analysis.

50. Sort the Terms table by the KEEP column by clicking the column heading. If needed, click the column heading twice so the sorted sign appears as **KEEP ▼**. Use the scroll bar to go to the top of this sorted Terms table.

You will see all the terms retained in the analysis in the order of document frequency.

Clearly, you will find certain terms that exist across almost all the documents that provide no value in discriminating among the documents. In this case, high-frequency terms such as **data** and **paper** may be dropped. Other high-frequency terms such as **set**, **type**, **ed**, **Carolina**, **north**, **introduction** can be dropped because they might be deemed not useful in this analysis.

51. Select all these terms as shown below. Right-click and select **Drop Terms**. This clears the KEEP column for all these terms, and they will be dropped from analysis. (Although not shown in the diagram due to lack of space, the term **SAS institute**, which appeared in 55 documents, is also dropped from this analysis. Make sure you do this by clearing the KEEP box for that term.)

Interactive Filter Viewer

File Edit View Window

Search :

TERM	FREQ	# DOCS	KEEP ▼	WEIGHT	ROLE	ATTRIBUTE
data	772	178	<input checked="" type="checkbox"/>	0.095		Alpha
analysis	247	113	<input checked="" type="checkbox"/>	0.164		Alpha
user	174	104	<input checked="" type="checkbox"/>	0.165		Alpha
procedure	184	93	<input checked="" type="checkbox"/>	0.196		Alpha
system	265	92	<input checked="" type="checkbox"/>	0.22		Alpha
paper	108	90	<input checked="" type="checkbox"/>	0.176		Alpha
program	194	88	<input checked="" type="checkbox"/>	0.207		Alpha
set	194	85	<input checked="" type="checkbox"/>	0.22		Alpha
university	125	80	<input checked="" type="checkbox"/>	0.218		Alpha
statistical	166	79	<input checked="" type="checkbox"/>	0.241		Alpha
group	101	76	<input checked="" type="checkbox"/>	0.221		Alpha
variable	169	75	<input checked="" type="checkbox"/>	0.241		Alpha
number	138	72	<input checked="" type="checkbox"/>	0.252		Alpha
problem	100	68	<input checked="" type="checkbox"/>	0.24		Alpha
model	173	67	<input checked="" type="checkbox"/>	0.291		Alpha
time	125	65	<input checked="" type="checkbox"/>	0.283		Alpha
north	69	65	<input checked="" type="checkbox"/>	0.229		Alpha
information	122	64	<input checked="" type="checkbox"/>	0.268		Alpha
type	101	63	<input checked="" type="checkbox"/>	0.262		Alpha
ed	68	61	<input checked="" type="checkbox"/>	0.247		Alpha
present	86	61	<input checked="" type="checkbox"/>	0.258		Alpha
introduction	60	59	<input checked="" type="checkbox"/>	0.245		Alpha
design	118	58	<input checked="" type="checkbox"/>	0.293		Alpha
carolina	70	57	<input checked="" type="checkbox"/>	0.275		Alpha
research	109	56	<input checked="" type="checkbox"/>	0.319		Alpha
observation	103	56	<input checked="" type="checkbox"/>	0.299		Alpha
form	84	55	<input checked="" type="checkbox"/>	0.287		Alpha
institute	57	55	<input checked="" type="checkbox"/>	0.259		Alpha
file	145	55	<input checked="" type="checkbox"/>	0.296		Alpha
computer	98	55	<input checked="" type="checkbox"/>	0.298		Alpha
value	111	55	<input checked="" type="checkbox"/>	0.316		Alpha
base	109	53	<input checked="" type="checkbox"/>	0.312		Alpha
report	92	53	<input checked="" type="checkbox"/>	0.292		Alpha



Suppose you want to treat a lot of related terms such as **bio**, **biometrics**, and **biological** as synonyms. First, you need to find the terms that you want to treat as synonyms and then add them to a synonym list.

52. First sort the Terms table by the TERM column.

53. Right-click to activate **Find** and type **bio**. Then select all of the terms shown below and right-click and select **Treat as synonyms**.

TERM ▲	FREQ	# DOCS	KEEP	WEIGHT	ROLE	ATTRIBUTE
blitynata	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
bill	1	1	<input type="checkbox"/>	0.0	Alpha	
bill gertsen	1	1	<input type="checkbox"/>	0.0	Miscellaneous P...	Entity
binulation	1	1	<input type="checkbox"/>	0.0	Miscellaneous P...	Entity
bination	1	1	<input type="checkbox"/>	0.0	Alpha	
bine	1	1	<input type="checkbox"/>	0.0	Alpha	
<b>bio</b>	1	1	<input checked="" type="checkbox"/>	0.0	Alpha	
bioassay	3	3	<input type="checkbox"/>	0.0	Alpha	
bioav	1	1	<input type="checkbox"/>	0.0	Alpha	
biochemical	1	1	<input checked="" type="checkbox"/>	0.0	Alpha	
biochemical cell	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
bioisay	1	1	<input type="checkbox"/>	0.0	Miscellaneous P...	Entity
biol ill	1	1	<input type="checkbox"/>	0.0	Miscellaneous P...	Entity
<b>biological</b>	5	4	<input checked="" type="checkbox"/>	0.0	Alpha	
biological data	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
biological experiment	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
biological model	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
biologist	1	1	<input type="checkbox"/>	0.0	Alpha	
biologist wish	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
biome	2	1	<input type="checkbox"/>	0.0	Alpha	
<b>biomedical</b>	2	2	<input checked="" type="checkbox"/>	0.0	Alpha	
biomedical computer laborato	1	1	<input type="checkbox"/>	0.0	Miscellaneous P...	Entity
biomedical computer programs	1	1	<input type="checkbox"/>	0.0	Miscellaneous P...	Entity
biomeo	1	1	<input type="checkbox"/>	0.0	Miscellaneous P...	Entity
<b>biometrics</b>	1	1	<input checked="" type="checkbox"/>	0.0	Miscellaneous P...	Entity
biometries	1	1	<input type="checkbox"/>	0.0	Alpha	
biometrika	1	1	<input type="checkbox"/>	0.0	Alpha	
biometrika article	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
biometry	1	1	<input checked="" type="checkbox"/>	0.0	Miscellaneous P...	Entity
biometry	2	1	<input checked="" type="checkbox"/>	0.0	Alpha	
<b>biostatistics</b>	5	5	<input checked="" type="checkbox"/>	0.701	Alpha	
biostatistics senic	1	1	<input type="checkbox"/>	0.0	Miscellaneous P...	Entity
birth	6	4	<input type="checkbox"/>	0.0	Alpha	
birth date	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
births t	1	1	<input type="checkbox"/>	0.0	Noun Group	Alpha
brison	1	1	<input type="checkbox"/>	0.0	Miscellaneous P...	Entity
brity	1	1	<input type="checkbox"/>	0.0	Alpha	
brivariate	1	1	<input type="checkbox"/>	0.0	Alpha	

54. In the pop-up box, **bio** is selected as the keyword to represent this group of terms. Click **OK**.

When each of these terms was considered individually, all the terms were dropped from analysis due to their low document frequency (less than 5 as in the setting). With the creation of a synonym group, the term **bio** will now be considered in the analysis because the document frequency will meet the cutoff value of 5. *However, you will have to re-run the node to see this term included in your analysis.*

55. Close the interactive filter. Make sure that you select **Yes** when asked to save results.

56. Run the Filter node again.

57. View the interactive filter and find the term **bio**. You see that the term is now being kept in the analysis. You can click the plus sign in front of the term to check that all terms in the synonym list are included.

Typically, you will have to browse and iterate through the complete list of drop/keep terms and make appropriate changes based on the objective of the analysis. This is *the most important and time-consuming task* of text mining analysis. This task is purely subjective and completely depends on the problem and the analyst's level of domain expertise.

At the end, you can save your custom synonyms by clicking **File** ⇒ **Export Synonyms**. Select the library where you intend to save your work and give it a meaningful name.