Text Mining

statgraphics 18®
centurion

Revised: 10/9/2017

Summary

The *Text Mining* procedure analyzes one or more text columns or documents to determine how frequently various words are used.

The calculations are performed by the "tm" package in R. To run the procedure, R must be installed on your computer together with the *tm, wordcloud, and RColorBrewer* packages. For information on downloading and installing R, refer to the document titled "R – Installation and Configuration".

The main output of this procedure is an identification of those words that occur most frequently. Both tabular and graphical summaries are provided.

Sample StatFolios: *textmining1.sgp* and *textmining2.sgp*

Sample Data

The data for this procedure may be in either of 2 formats:

1. A set of text documents external to the program.
2. A collection of text columns loaded into the Statgraphics DataBook

As an example of analyzing external documents, we will analyze the following speeches using the text mining procedure:

1. *Give Me Liberty or Give Me Death* (1775): Patrick Henry.
2. *The Hypocrisy of American Slavery* (1852): Frederick Douglas
3. *Gettysburg Address* (1863): Abraham Lincoln
4. *Women's Right to Suffrage* (1873): Susan B. Anthony
5. *Blood, Toil, Tears and Sweat* (1940): Winston Churchill
6. *Inaugural Address* (1961): John F. Kennedy
7. *I Have a Dream* (1963): Martin Luther King, Jr.
8. *The American Promise* (1965): Lyndon B. Johnson
9. *Remarks at the Brandenburg Gate* (1987): Ronald Reagan

Each speech is saved in a separate *txt* file using the author's name.

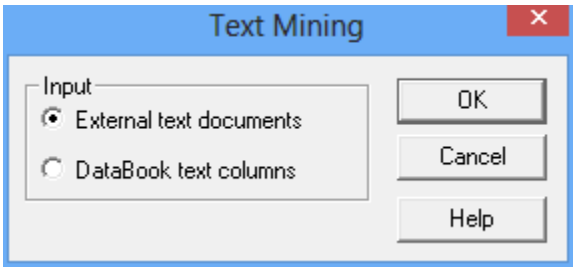
A second example uses the data contained in the file *job openings.sgd*, a portion of which is shown below:

Business Title
Deputy Director, QA & Strategic Statistical Analysis
Deputy Director, QA & Strategic Statistical Analysis
HIGH PRESSURE PLANT TENDER
HIGH PRESSURE PLANT TENDER
Project Coordinator, District Public Health Brooklyn
REPOST - Per Diem - Remote Learning Team Specialist
REPOST - Per Diem - Remote Learning Team Specialist
Research Analyst, Family and Child Health Administration
Graphic Artist
Deputy Commissioner, Wastewater Treatment
Graphic Artist
Account Manager
Deputy Director, QA & Strategic Statistical Analysis
…

The file contains a single column of text showing the available job openings in the City of New York on July 11, 2017.

Data Input

When the *Text Mining* procedure is selecting from the Statgraphics menu, the first dialog box displayed requests the type of format in which the data are stored:

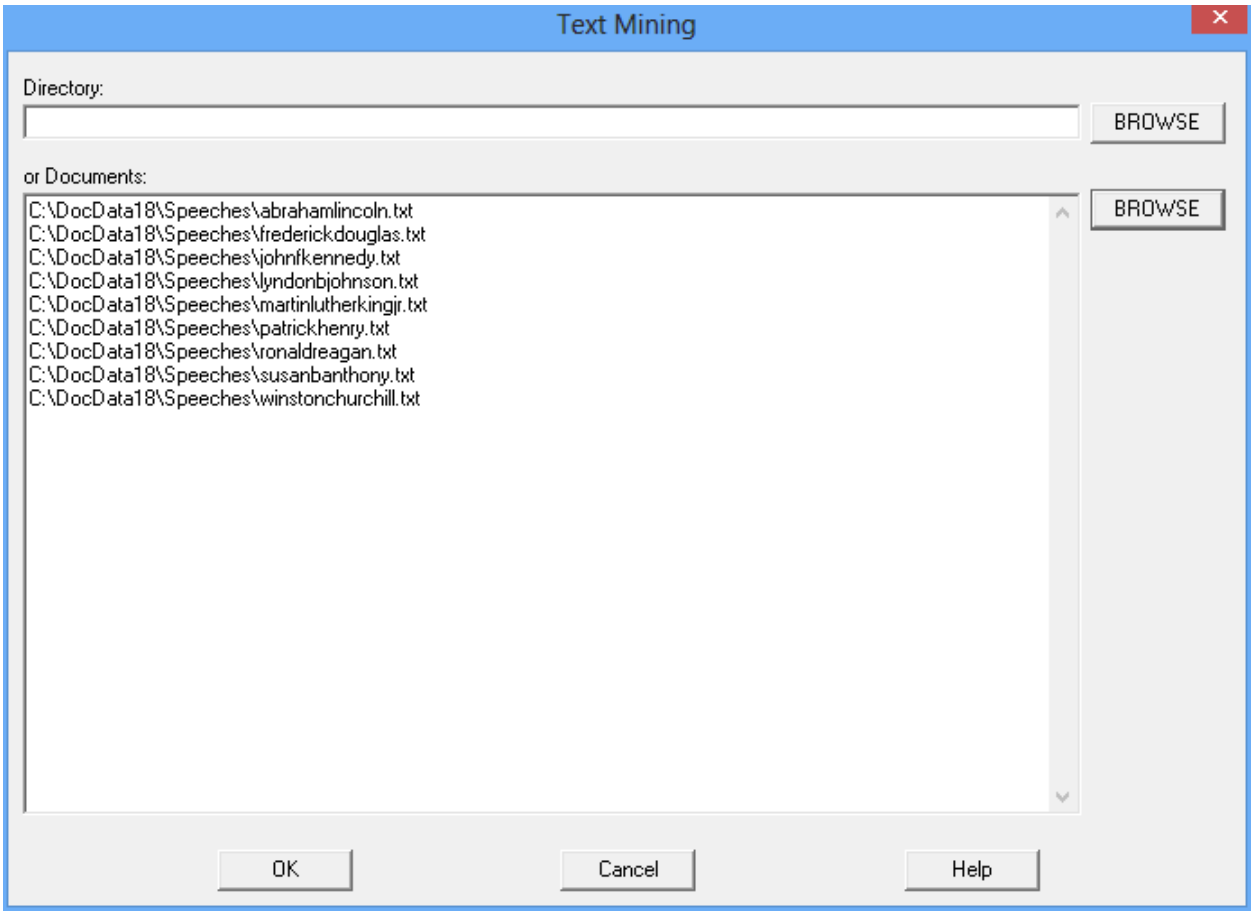


The data may consist of either a set of external text documents or a collection of character columns already entered into the Statgraphics DataBook.

External text documents

If the text to be mined is contained in a set of external documents, those documents must be in the form of plain text. Such documents typically have the extension TXT and are easily opened in programs such as Notepad or Word.

To select the documents to be analyzed, the data input dialog box shown below is used. If a directory name is provided, all documents in that directory will be analyzed. Otherwise, each document should be listed separately.

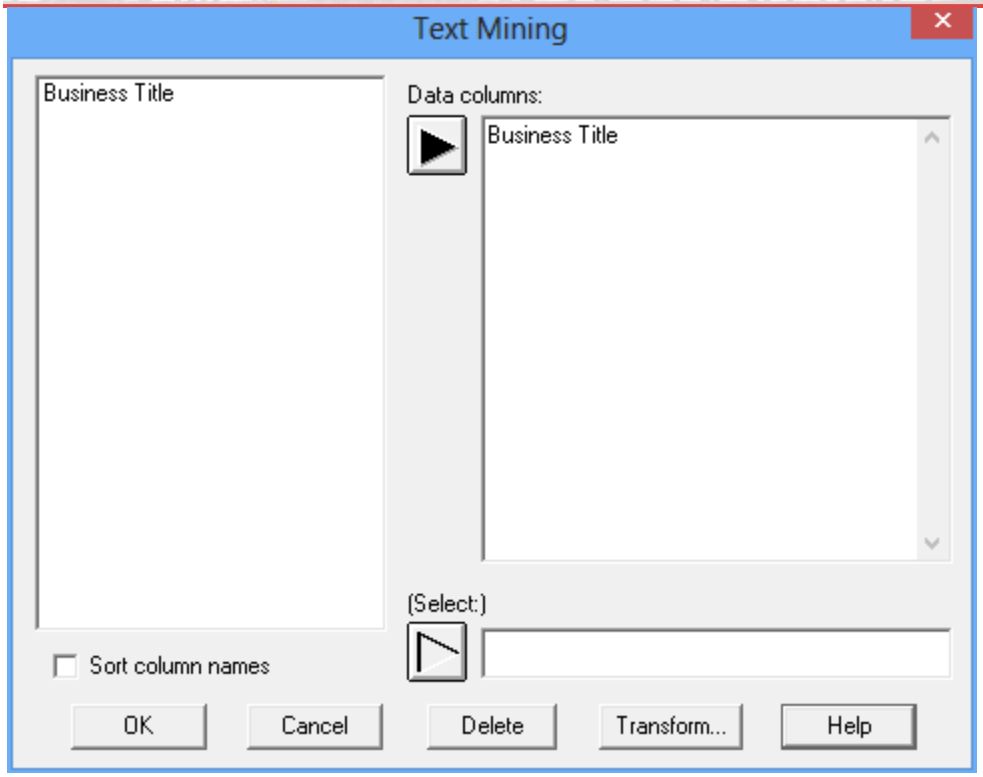


- **Directory:** name of a directory containing the text documents to be analyzed. All files ending with the extension TXT will be analyzed.

- **Documents:** if the *Directory* field is blank, then the files listed in this field will be analyzed. Only plain text files may be specified.

DataBook text columns

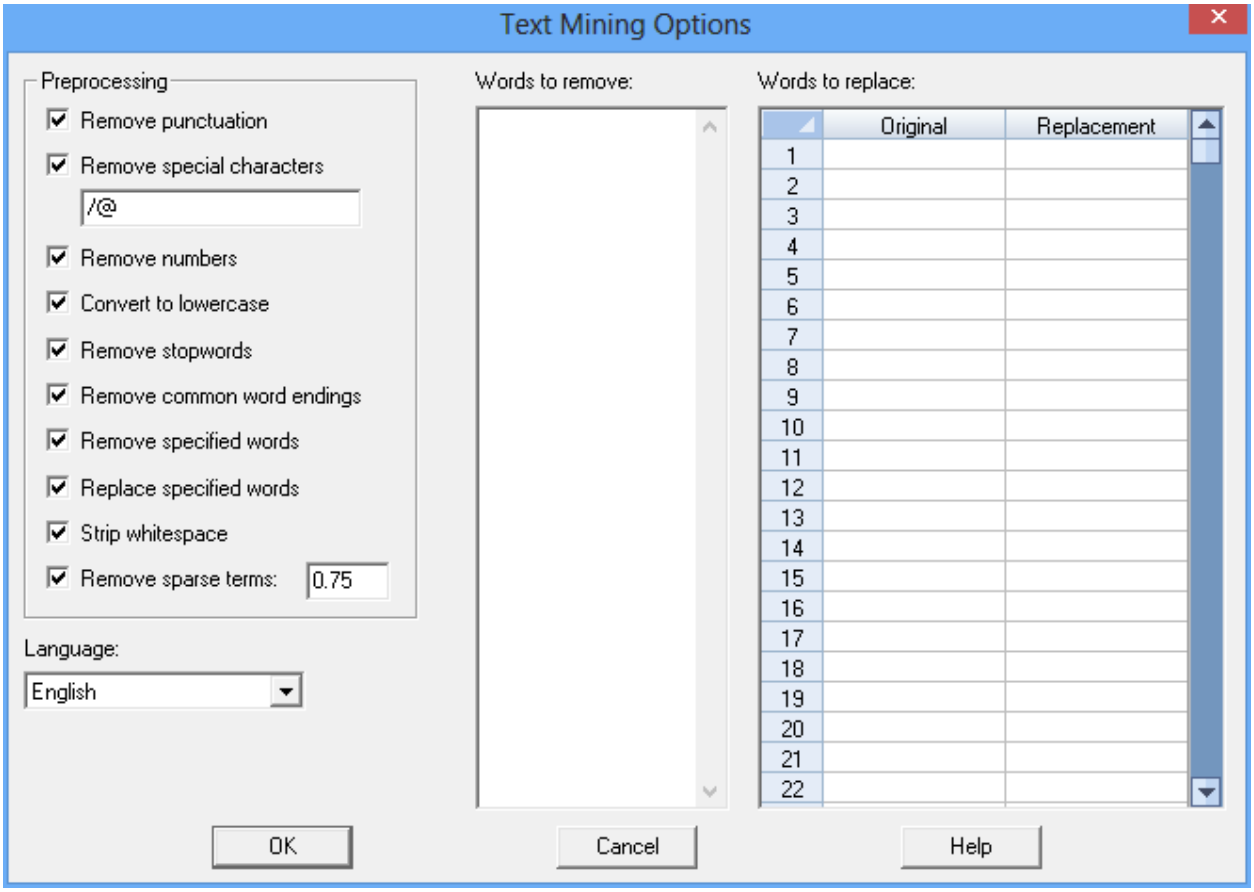
If the text to be mined has been loaded into the Statgraphics DataBook, the following dialog box is displayed:



- **Data columns:** one or more columns of text to be analyzed.
- **Select:** row subset selection.

Analysis Options

The *Analysis Options* dialog box sets various options for processing the documents:



- **Preprocessing**: specifies the steps to be performed on each document before words are counted:
    - **Remove punctuation**: removes all punctuation.
    - **Remove special characters**: removes any characters entered in the field below.
    - **Remove numbers**: removes all numbers.
    - **Convert to lowercase**: converts all words to lowercase.
    - **Remove stopwords**: removes common words such as "a", "the", "and" and "or". This option is affected by the **Language** setting. In the case of English, there are 2 options: "English" and "English (SMART list)". The latter setting removes a much larger set of common words.
    - **Remove common word endings**: removes endings such as "ing" and "ed", leaving the roots. This prevents words such as "working" and "worked" from appearing as separate terms. This option is also affected by the **Language** setting.
    - **Remove specified words**: removes any words listed in the **Words to remove** field.
    - **Replace specified words**: replaces any words listed in the **Words to replace** field with the specified replacements.
    - **Strip whitespace:** removes any unnecessary empty space between words.

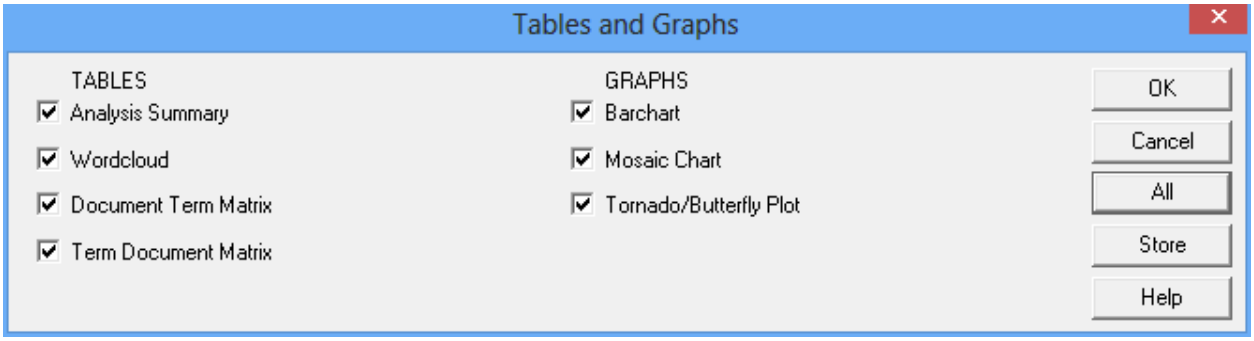
- **Remove sparse terms:** when analyzing multiple documents, removes terms that are missing from many documents. The numeric value specified is the maximum allowable proportion of documents that may be missing the word for that word to be retained. By default, words will be retained if they are missing in no more than 75% of the documents.

- **Words to remove:** a list of words to be removed from the documents. List each word on a separate line.

- **Words to replace:** a list of words to be replaced with other words before analyzing the documents. List each word in a separate row with its replacement.

Tables and Graphs

The following tables and graphs may be created:



Analysis Summary

The *Analysis Summary* begins with a list of the R commands that were executed. The first section lists the documents that were analyzed, which are placed into a "corpus":

Text Mining

```
setwd("c:\\temp")
library("tm")

## Loading required package: NLP

input <-c("C:\\DocData18\\Speeches\\abrahamlincoln.txt","C:\\DocData18\\Sp...
corpus <- Corpus(source, readerControl=list(reader=readPlain))
summary(corpus)

##                          Length Class             Mode
## abrahamlincoln.txt      2      PlainTextDocument list
## frederickdouglas.txt    2      PlainTextDocument list
## johnfkennedy.txt        2      PlainTextDocument list
## lyndonbjohnson.txt      2      PlainTextDocument list
## martinlutherkingjr.txt 2      PlainTextDocument list
## patrickhenry.txt        2      PlainTextDocument list
## ronaldreagan.txt        2      PlainTextDocument list
## susanbanthony.txt       2      PlainTextDocument list
## winstonchurchill.txt    2      PlainTextDocument list
```

The second section shows the transformations made to the documents in the corpus:

```
corpus <- tm_map(corpus, removePunctuation)
toSpace <- content_transformer(function(x, pattern) {return (gsub(pattern, "
", x))})
corpus <- tm_map(corpus, toSpace, "/")
corpus <- tm_map(corpus, toSpace, "@")
corpus <- tm_map(corpus, removeNumbers)
corpus <- tm_map(corpus, content_transformer(tolower))
corpus <- tm_map(corpus, removeWords, stopwords("en"))
corpus <- tm_map(corpus, stemDocument, language = "en")
corpus <- tm_map(corpus, stripWhitespace)
```

Two additional commands are then used to create a "document-term matrix", which tabulates the number of times each word is included in each document:

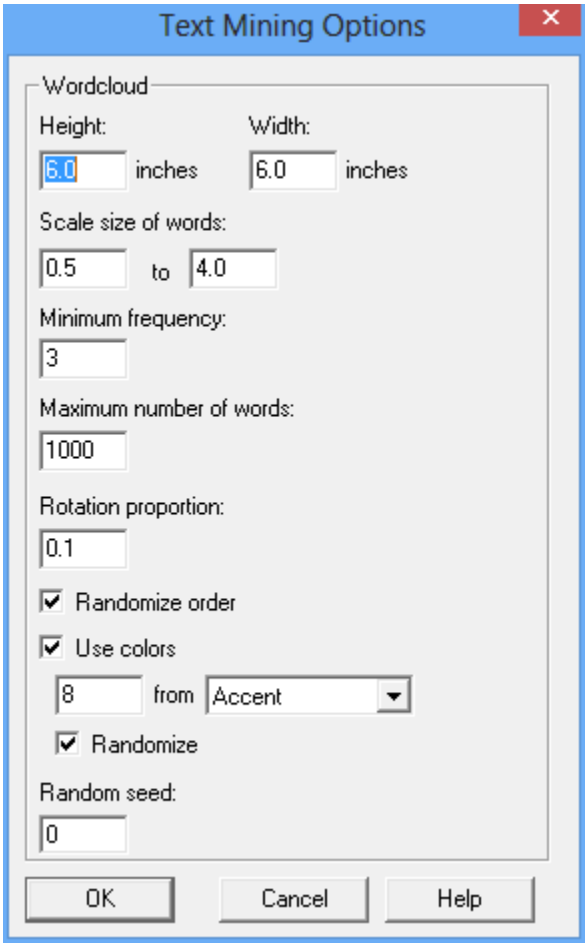
```
dtm <- DocumentTermMatrix(corpus)
dtm <- removeSparseTerms(dtm, 0.75)
```

Wordcloud

The document-term matrix is used to create a wordcloud:

The size of the words in the cloud is proportional to the total number of times each word occurs across all documents. The above wordcloud shows that the most frequently used word in the 9 speeches combined (other than those removed during preprocessing) is the word “will”. Other frequently used words are “right”, “freedom”, “nation”, “must”, “one”, and “american”. Note the result of removing word endings, so that “peac” represents both “peace” and “peaceful”.

Pane Options



- **Height:** height of the wordcloud in inches.
- **Width:** width of the wordcloud in inches.
- **Scale size of words:** range of allowable sizes for the words.
- **Minimum frequency:** words will only be included if they occur at least this often when summing the counts across all of the documents.
- **Maximum number of words:** the maximum number of words that will be displayed.
- **Rotation proportion:** the approximate proportion of words that will be oriented vertically rather than horizontally.
- **Randomize order:** if checked, words will be added to the cloud in random order.
- **Use colors:** whether to plot the words in color rather than using only black. If checked, the fields below specify the number of colors to use and the RColorBrewer palette from which

they will be selected. Palettes may be viewed at https://www.nceas.ucsb.edu/~frazier/RSpatialGuides/colorPaletteCheatsheet.pdf.

- **Randomize:** whether to randomize the order of colors used.

- **Random seed:** seed used by the random number generator. Using the same random seed insures that the same wordcloud will be created if the analysis is rerun.

Document Term Matrix

A key component used in text mining is the document-term matrix. Given *d* documents and *t* terms, the document-term matrix *F* is a *d* by *t* matrix with elements

$f_{i,j}$ = number of times term *j* occurs in document *i*.

The table below shows part of the matrix *F* calculated for the sample data, where $d = 9$ and $t = 367$:

Document Term Matrix
Number of documents: 9
Number of terms: 367

	act	age	ago	air	allow	almighti	alon	alreadi	also	america	american	among	answer
abrahamlincoln.txt	0	0	1	0	0	0	0	0	0	0	0	0	0
frederickdouglas.txt	2	0	0	1	1	1	0	2	0	4	4	1	2
johnfkennedy.txt	0	1	1	0	0	1	0	0	0	3	4	0	1
lyndonbjohnson.txt	2	0	3	0	1	0	0	1	2	5	25	4	2
martinlutherkingjr.txt	0	0	1	0	2	1	1	0	1	5	4	0	0
patrickhenry.txt	0	0	0	0	0	1	2	2	0	0	0	0	0
ronaldreagan.txt	0	1	8	2	0	0	2	0	1	2	3	1	0
susanbanthony.txt	0	0	0	0	0	0	0	0	0	1	0	0	0
winstonchurchill.txt	1	1	0	2	2	0	0	0	1	0	0	1	1
TOTAL	5	3	14	5	6	4	5	5	5	20	40	7	6

	argument	arm	ask	back	basic	battl	becom	begin	belief	believ	beyond	black	bless	bodi
abrahamlincoln.txt	0	0	0	0	0	0	0	0	0	0	0	0	0	0
frederickdouglas.txt	4	0	2	0	0	0	0	0	0	0	1	1	2	1
johnfkennedy.txt	0	4	6	2	0	1	2	4	2	1	2	0	1	0
lyndonbjohnson.txt	1	1	3	2	2	3	0	1	2	4	1	2	1	2
martinlutherkingjr.txt	0	0	1	9	1	0	1	1	0	3	0	4	0	1
patrickhenry.txt	2	3	2	2	0	3	0	0	0	0	0	0	0	0
ronaldreagan.txt	0	7	1	1	1	0	5	2	2	2	0	0	1	0
susanbanthony.txt	0	0	0	0	0	0	0	0	0	1	1	0	4	0
winstonchurchill.txt	0	0	2	0	0	2	0	0	0	0	0	0	0	0
TOTAL	7	15	17	16	4	9	8	8	6	11	5	7	9	4

The total number of times each word occurred in all documents combined

$$f_j = \sum_{i=1}^{d} f_{i,j} \quad j=1, 2, \ldots, t \qquad (1)$$

is also shown.

Term Document Matrix

This table shows the term-document matrix, which is a transposed version of the document-term matrix:

Term Document Matrix
Number of terms: 367
Number of documents: 9

	abrahamlincoln.txt	frederickdouglas.txt	johnfkennedy.txt	lyndonbjohnson.txt
act	0	2	0	2
age	0	0	1	0
ago	1	0	1	3
air	0	1	0	0
allow	0	1	0	1
almighti	0	1	1	0
alon	0	0	0	0
alreadi	0	2	0	1
also	0	0	0	2
america	0	4	3	5
american	0	4	4	25
among	0	1	0	4
answer	0	2	1	2
argument	0	4	0	1
arm	0	0	4	1
ask	0	2	6	3
back	0	0	2	2
basic	0	0	0	2
battl	0	0	1	3
becom	0	0	2	0
begin	0	0	4	1
belief	0	0	2	2
believ	0	0	1	4
beyond	0	1	2	1
black	0	1	0	2
bless	0	2	1	1
bodi	0	1	0	2
bring	0	1	2	0
british	0	0	0	0
brother	0	1	0	1
brought	1	2	0	1
brutal	0	0	0	1
build	0	1	0	1
burden	0	1	3	0
busi	0	0	0	0

TOTAL	78	385	438	1002

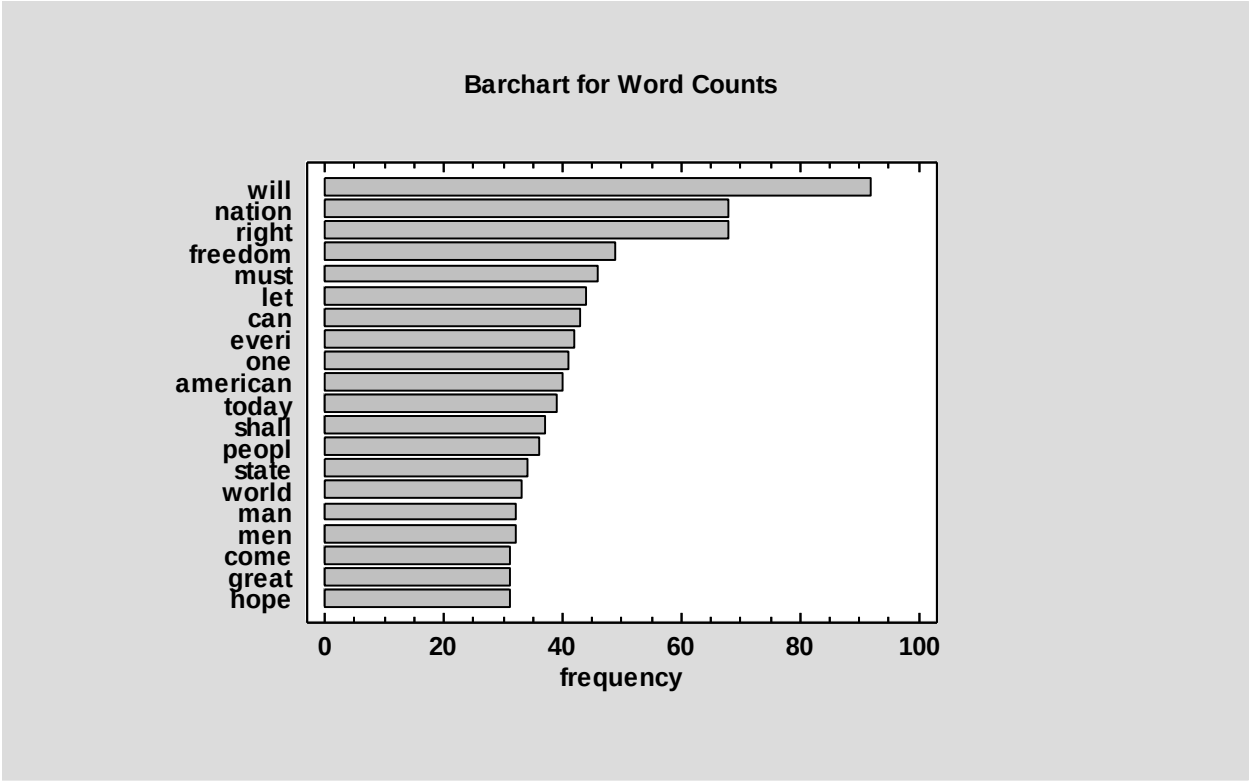
The total number of words in each document

$$f_i = \sum_{j=1}^{t} f_{i,j} \qquad i=1, 2, \ldots, d \tag{2}$$

is also shown.

Barchart

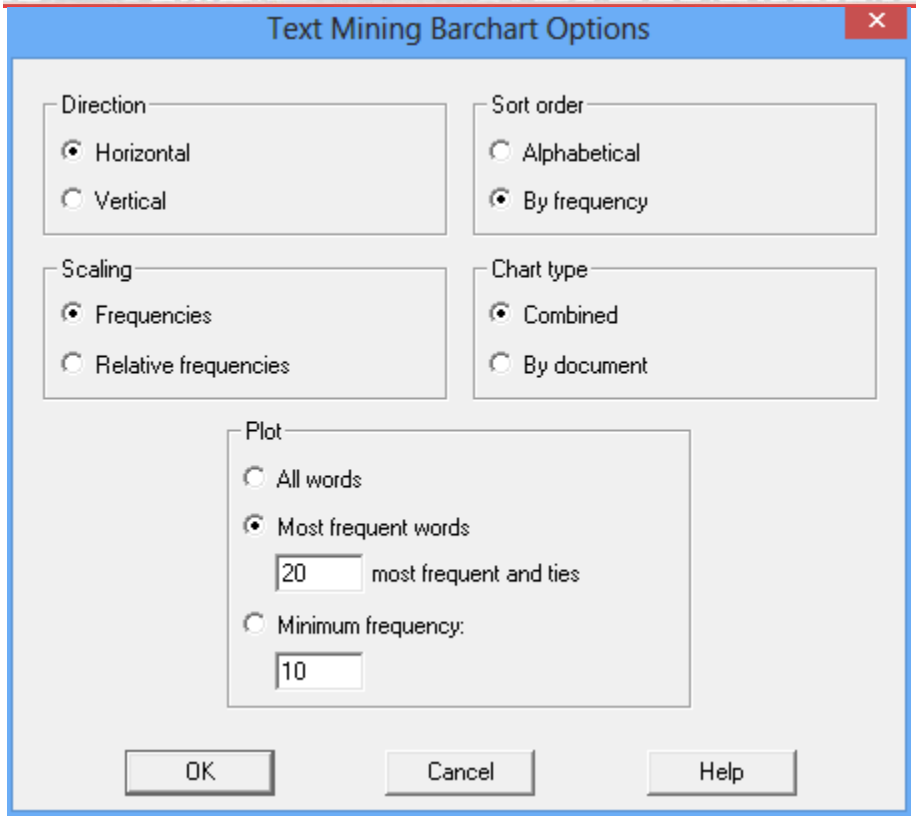
A barchart may be created illustrating how often each word occurs. The chart belows shows the 20 most frequent words, summed across all of the documents:



For example, the word “will” occurred more than 90 times throughout the 9 documents.

Pane Options

There are several options that control the format of the barchart:



- **Direction:** orientation of the bars.

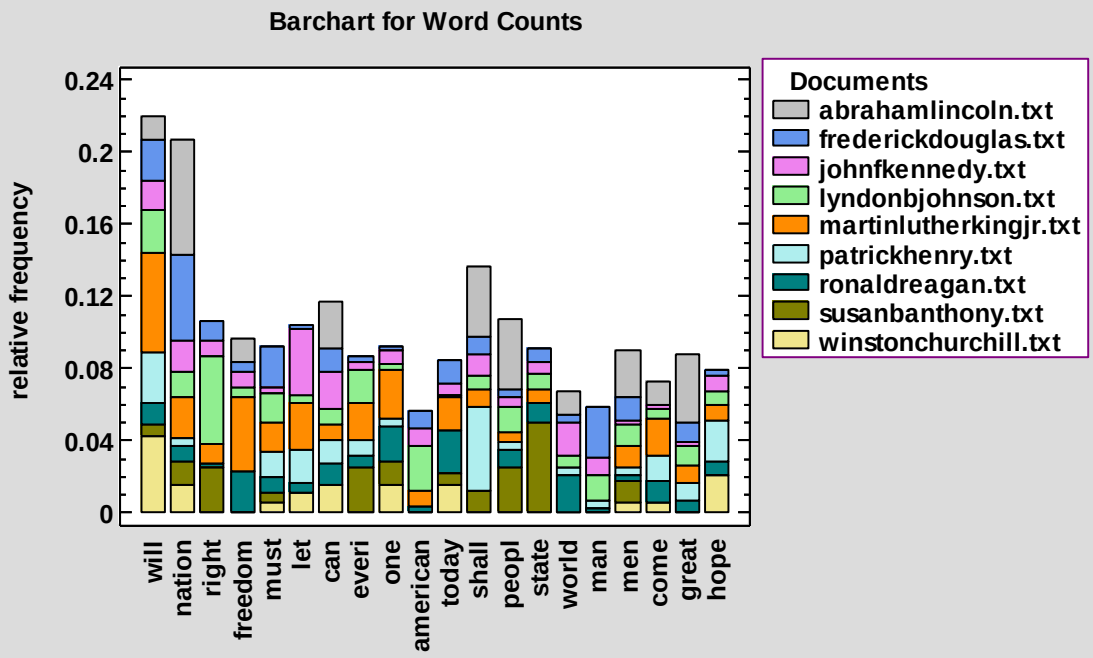
- **Sort order:** whether the terms should be sorted alphabetically or in order of decreasing frequency.

- **Scaling:** whether the size of the bars should be proportional to the frequencies $f_{i,j}$ or the relative frequencies $f_{i,j} / f_i$. Whereas the absolute frequencies give more weight to longer documents, the relative frequencies attempt to weight all documents equally by dividing the frequencies by the number of terms in the corresponding document.

- **Chart type:** whether the chart should show single bars for each term or split the bars by document.

- **Plot:** controls the number of terms plotted in the chart. The number of words may be restricted by specifying either the number of bars to be plotted or the minimum frequency f_j for a term to be included.

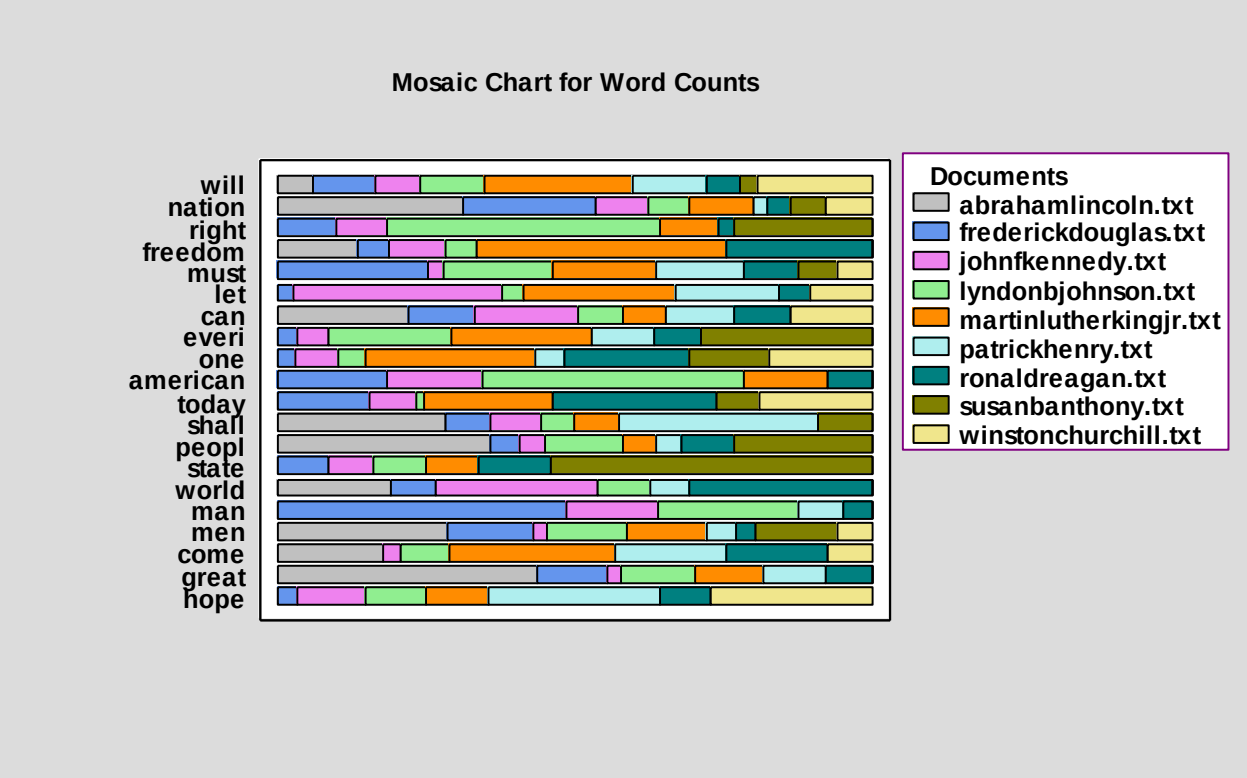
The chart below shows a vertical barchart containing relative frequencies by document:



Notice that the 2 speeches in which the word "nation" occupies a large proportion of the speech are those by Abraham Lincoln and Frederick Douglas.

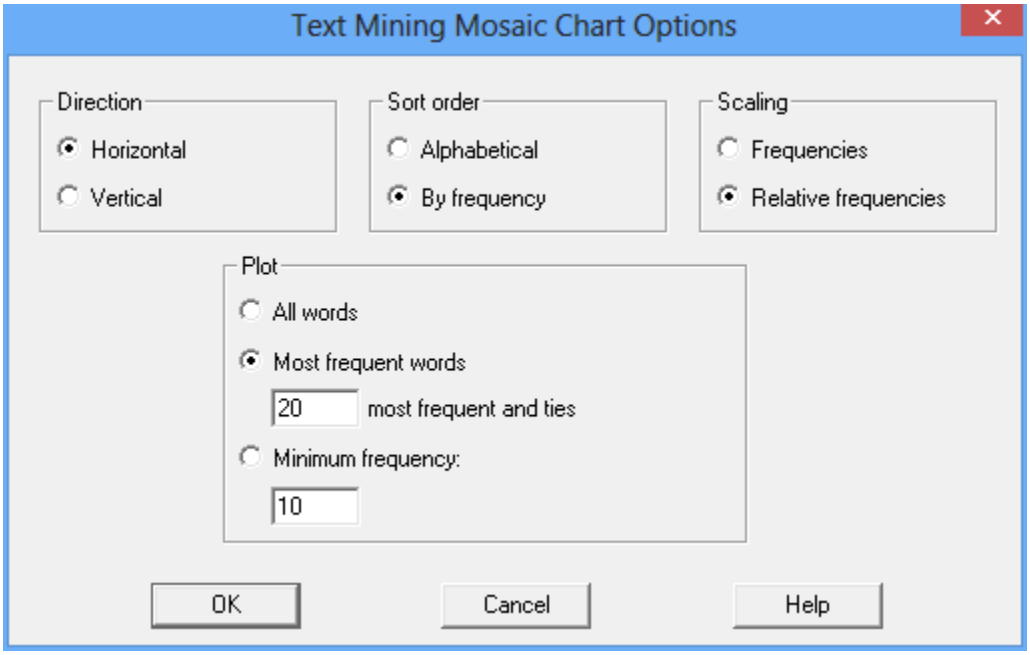
Mosaic Chart

This plot is similar to a barchart by document, except that each bar is expanded to fill the entire width or height of the graph:



As such, the width of each section of the bar is proportional to the conditional distribution of each term across the documents. In the above plot, the widest section is for the word “state” in the speech by Susan B. Anthony. (She used a variant of the word 8 times out of a total of 161 words.)

Pane Options



- **Direction:** orientation of the bars.

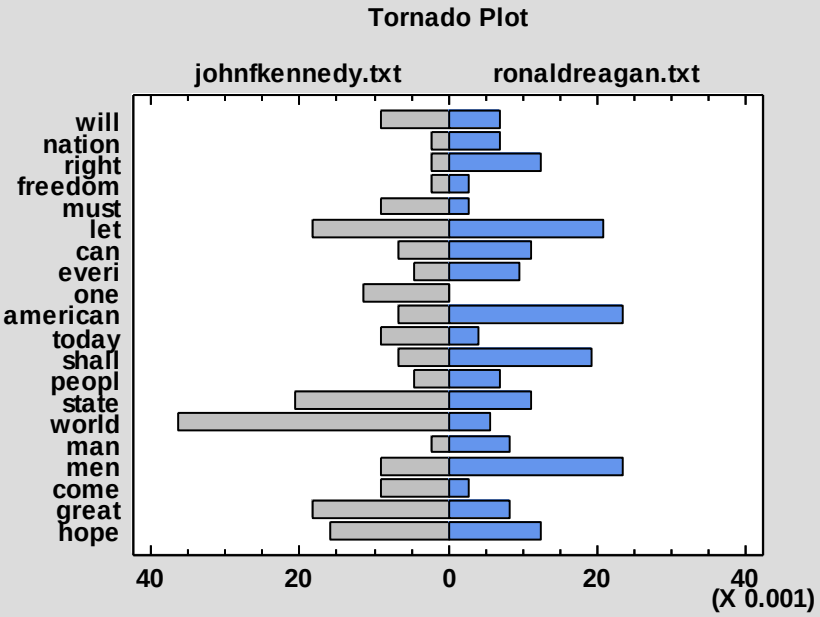
- **Sort order:** whether the terms should be sorted alphabetically or in order of decreasing frequency.

- **Scaling:** whether the size of the bars should be proportional to the frequencies $f_{i,j}$ or the relative frequencies $f_{i,j} / f_i$. Whereas the absolute frequencies give more weight to longer documents, the absolute frequencies attempt to weight all documents equally by dividing the frequencies by the number of terms in the corresponding document.

- **Plot:** controls the number of terms plotted in the chart. The number of words may be restricted by specifying either the number of bars to be plotted or the minimum frequency f_j for a term to be included.

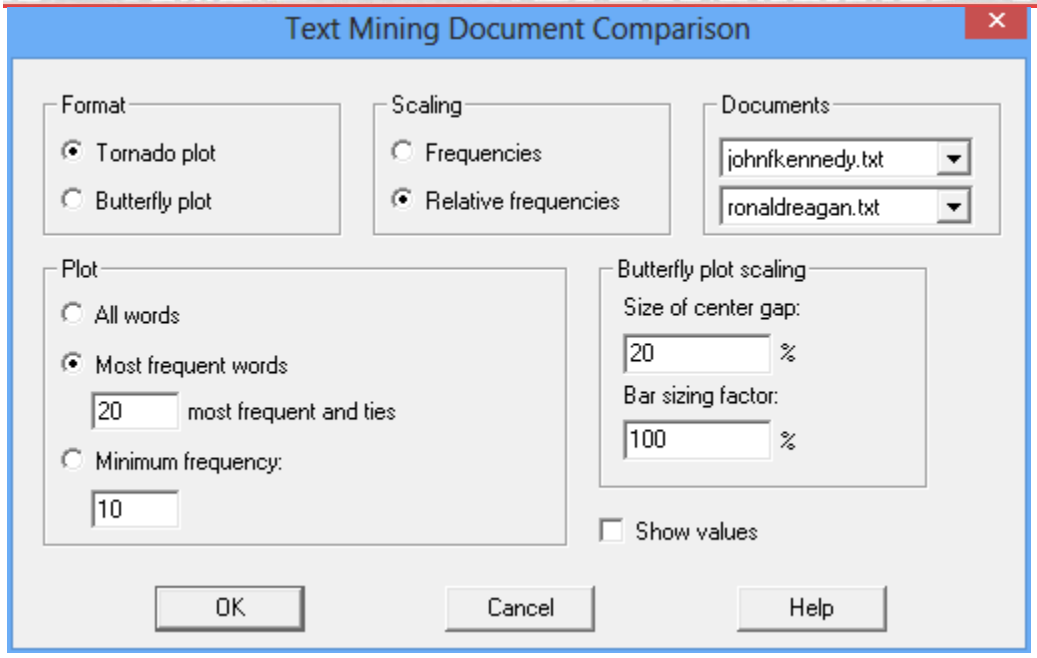
Tornado/Butterfly Plot

This plot is used to compare 2 documents. A typical tornado plot is shown below, comparing the speeches by John F. Kennedy and Ronald Reagan:



The most noticeable difference between the speeches is the frequent use of the word “world” by Kennedy compared to the frequent use of the word “american” by Reagan.

Pane Options



- **Format:** the chart can be plotted as either a tornado plot in which the bars are adjacent to each other, or a butterfly plot which has the terms in the middle.

- **Scaling:** whether the size of the bars should be proportional to the frequencies $f_{i,j}$ or the relative frequencies $f_{i,j} / f_{i\cdot}$. Whereas the absolute frequencies give more weight to longer documents, the absolute frequencies attempt to weight all documents equally by dividing the frequencies by the number of terms in the corresponding document.

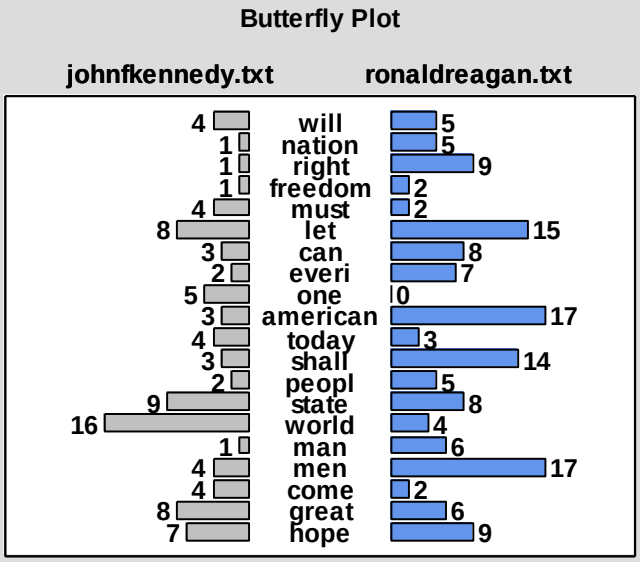
- **Documents:** the 2 documents to be compared.

- **Plot:** controls the number of terms plotted in the chart. The number of words may be restricted by specifying either the number of bars to be plotted or the minimum frequency f_{j} for a term to be included.

- **Butterfly plot scaling:** for a butterfly plot, controls the size of the gap in which the terms are displayed and the relative width of each bar.

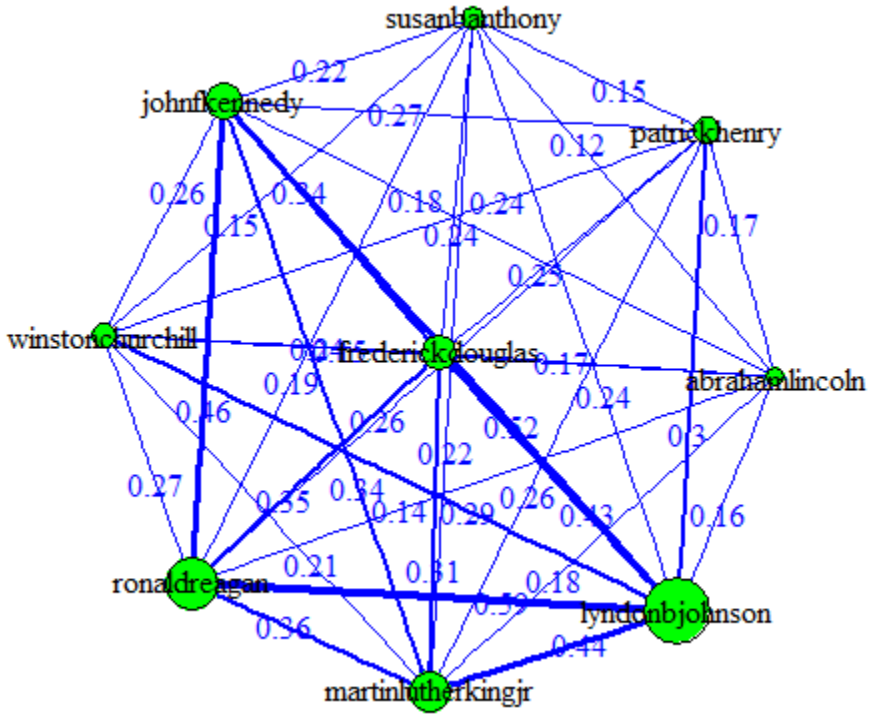
- **Show values:** if checked, the values associated with each bar are displayed.

The chart below is a butterfly plot displaying the frequencies of each word:



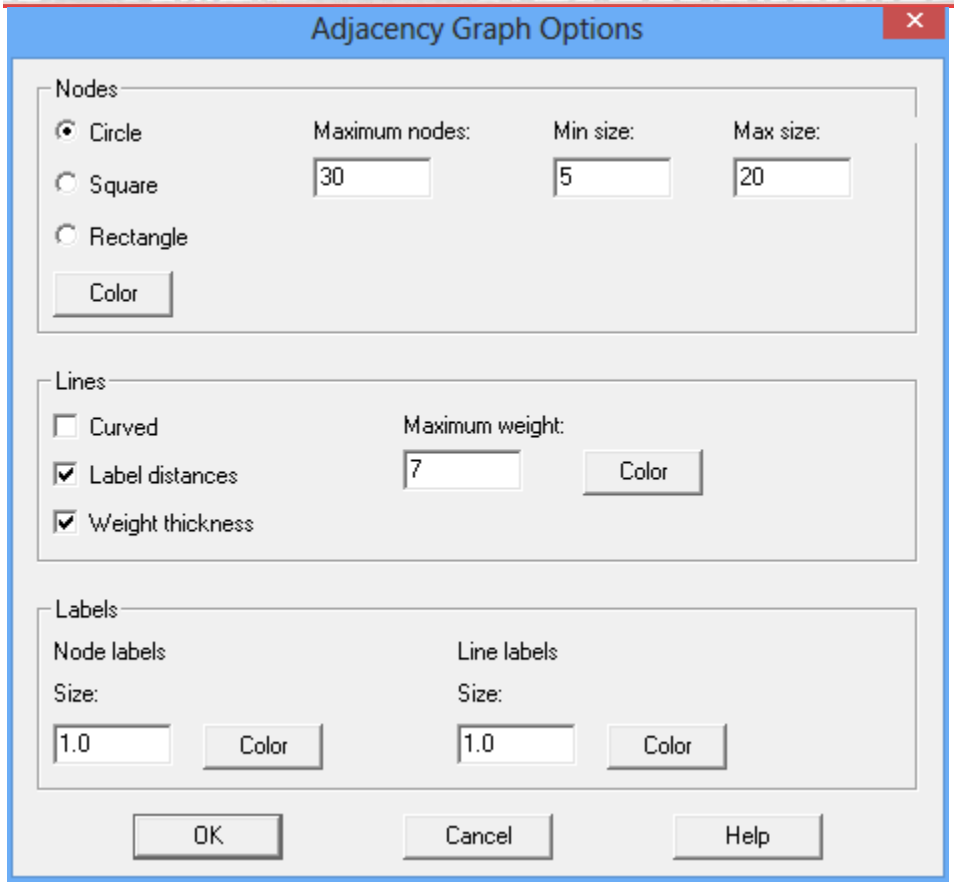
Document Adjacency Diagram

A diagram may be creating illustrating the similarity between the word distributions in different documents. It takes a form similar to that shown below:



Each pair of documents is connected by a line. The size of the nodes represents the number of words in the document-term matrix corresponding to each doument. The width of each connecting line and its label is a meaure of the similarity between the word distributions in the documents connected by the line. The larger the value and the thicker the line, the greater the similarity between the documents.

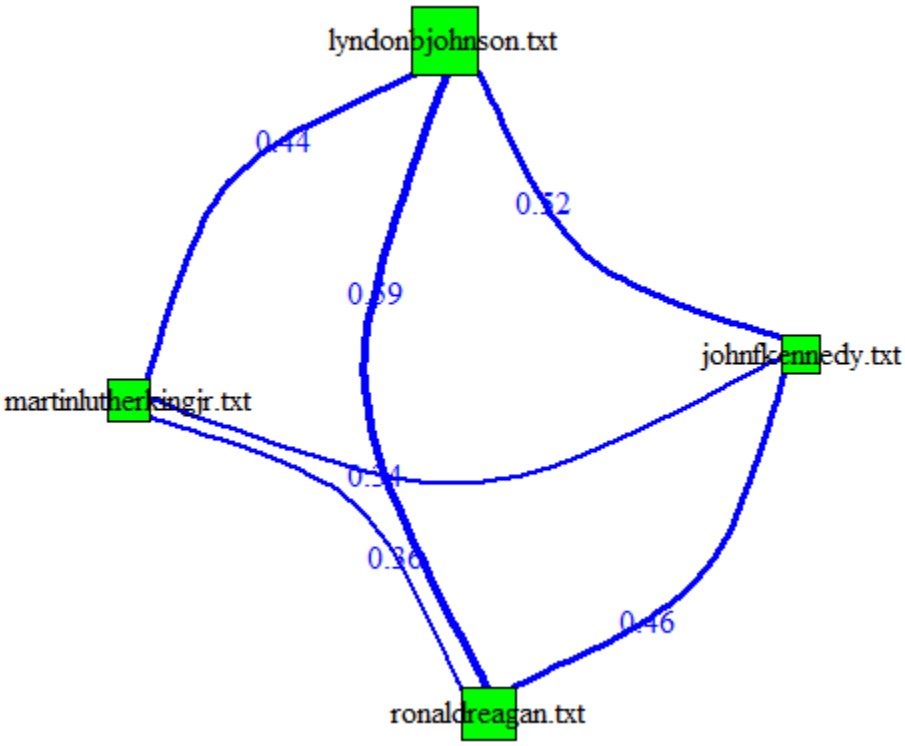
Pane Options



- **Node shape:** indicates the shape of the nodes.
- **Color:** Push the button to select a color for the nodes.
- **Maximum nodes:** maximum number of documents to be displayed on the diagram. If less than the total number of documents, the documents with the greatest number of terms will be displayed.
- **Min size:** minimum size of the nodes.
- **Max size:** maximum size of the nodes.
- **Curved:** if checked, the line connectors between the nodes will be curved.
- **Label distances:** if checked, each line connector will be labeled with a measure of similarity between the connected documents.
- **Weight thickness:** if checked, the thickness of the line connectors will be weighted by the similarities between the documents.
- **Maximum weight:** specifies the maximum thickness of the line connectors.

- **Nodes labels:** controls the size and color of the node labels.
- **Line labels:** controls the size and color of the labels on the line connectors.

The diagram below illustrates the similarities amongst the 4 largest documents:



Of the 4 speeches, the ones given by Lyndon Johnson and Ronald Reagan are the most similar.

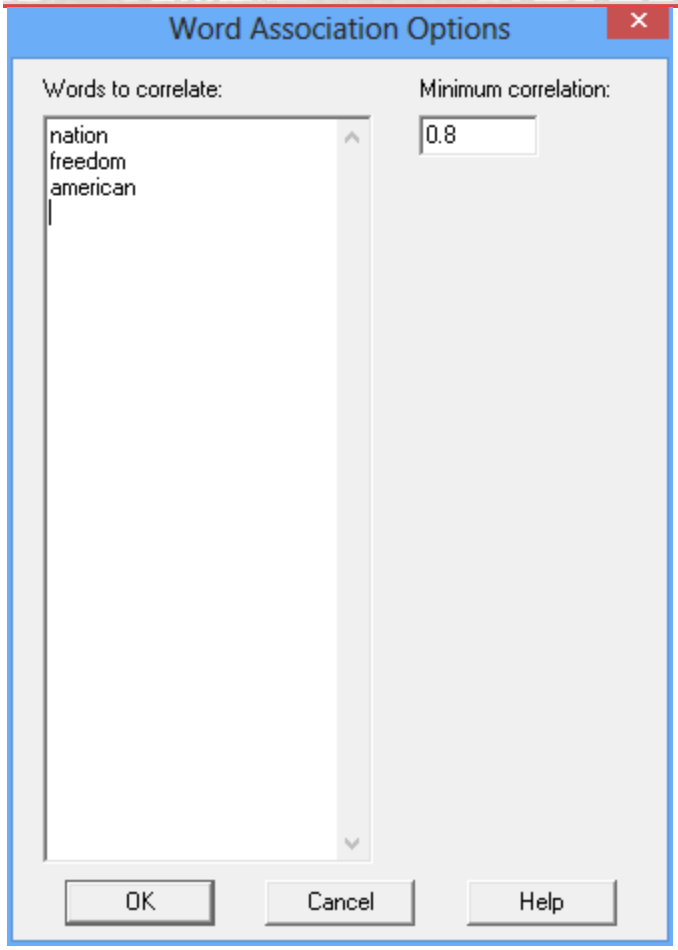
Word Associations

This pane shows the result of examining the distributions of frequently occurring words across multiple documents. For a set of specified words, those terms most highly correlated with them are displayed:

Word Associations

```
findAssocs(dtm, c("nation","freedom","american"), 0.8)
## $nation
##  whose   need   soul victim present america  wrong   rich command
##   0.90   0.88   0.88   0.88   0.87   0.86   0.86   0.85   0.84
##    man
##   0.81
##
## $freedom
##   one moment  come  still brutal  sign togeth  lead  refus  given
##  0.97  0.93  0.92  0.91  0.90  0.89  0.89  0.88  0.85  0.84
## today  note  year
##  0.84  0.82  0.81
##
## $american
##    right     issu     pass    share  countri    heart     race
##     0.99     0.98     0.98     0.97     0.96     0.96     0.96
##     came    civil    equal opportun  tonight     time     vote
##     0.95     0.95     0.95     0.95     0.95     0.94     0.94
##     hatr   histori   intend     root    among     caus democraci
##     0.93     0.93     0.93     0.93     0.92     0.92     0.92
##      men   poverti      use     just     help    peopl     serv
##     0.92     0.92     0.92     0.91     0.90     0.89     0.89
##    everi   fought    great     must    bodi conscienc    elect
##     0.88     0.88     0.88     0.88     0.87     0.87     0.87
##     live     mani  violenc   presid    basic   extend progress
##     0.87     0.87     0.87     0.86     0.85     0.85     0.85
##     real constitut  protest     view    found    give communiti
##     0.85     0.84     0.84     0.84     0.83     0.83     0.82
## injustic   leader    never  privileg      man     work
##     0.82     0.82     0.82     0.82     0.81     0.80
```

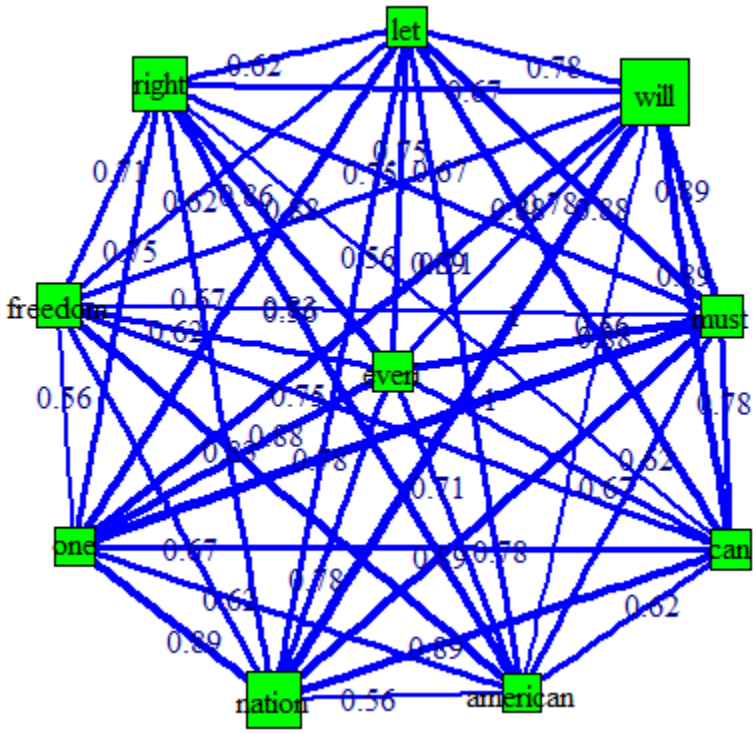
Pane Options



- **Words to correlate:** selected words for which correlations are to be examined.
- **Minimum correlation:** the minimum corrrelation between words for the pair to be included in the table.

Term Adjacency Diagram

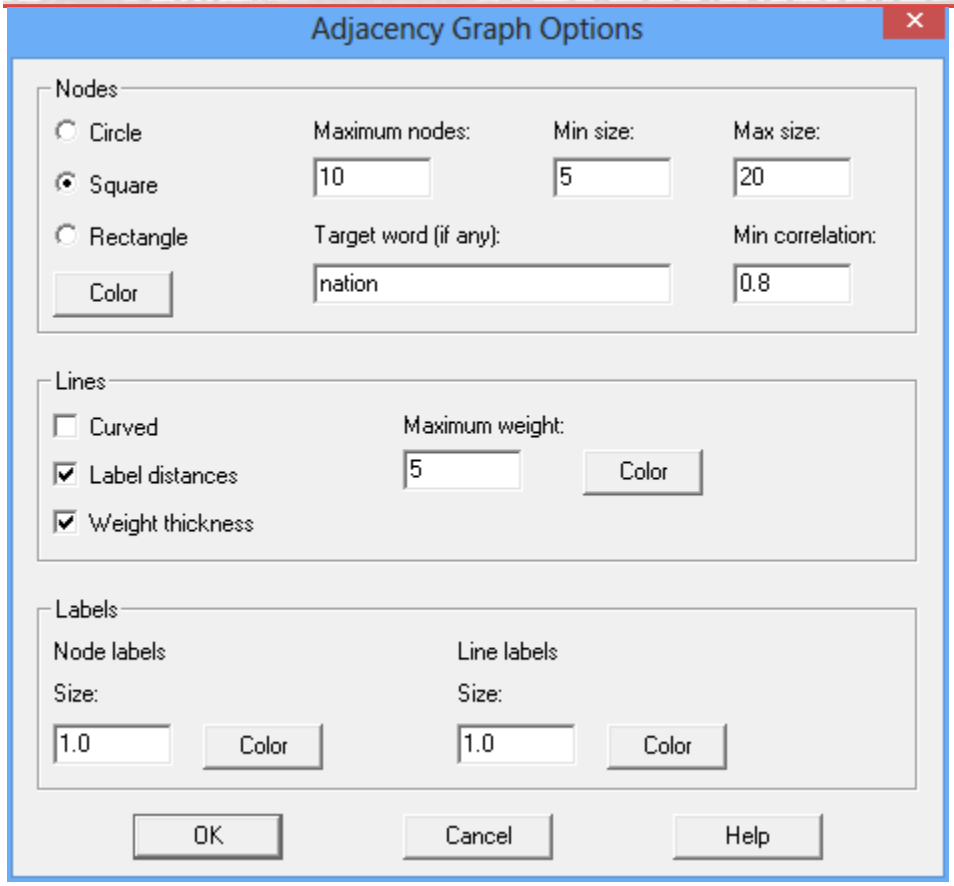
A diagram may be created illustrating the similarity between the distribution of pairs of words across different documents. It takes a form similar to that shown below:



Each pair of words is connected by a line. The size of the nodes represents the number of words in the term-document matrix corresponding to each term. The width of each connecting line and its label is a measure of the similarity between the distributions of the words connected by the line. The larger the value and the thicker the line, the greater the similarity between the words.

Note: The "similarity" between 2 terms is defined as the proportion of documents in which both terms appear. This is different than the correlations shown in the word assocations table.

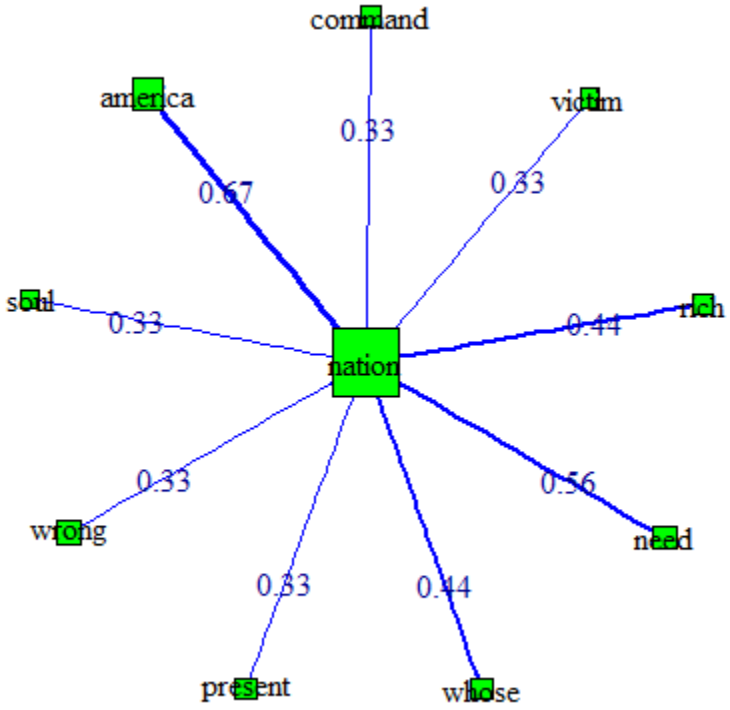
Pane Options



- **Node shape:** indicates the shape of the nodes.
- **Color:** Push the button to select a color for the nodes.
- **Maximum nodes:** maximum number of terms to be displayed on the diagram. If less than the total number of terms, the terms occurring most frequently will be displayed.
- **Target word:** If a word is entered in this field, the diagram will show connections between the indicated word and other terms, but will not contain connectors between the others. If the field is left blank, all pairs of terms will be connected.
- **Min correlation:** If a target word is specified, only words with a correlation of this magnitude or higher will be included in the diagram. If the number of such words is greater than "Maximum nodes" – 1, the words with the strongest correlation will be shown.
- **Min size:** minimum size of the nodes.
- **Max size:** maximum size of the nodes.
- **Curved:** if checked, the line connectors between the nodes will be curved.

- **Label distances:** if checked, each line connector will be labeled with a measure of similarity between the connected terms.
- **Weight thickness:** if checked, the thickness of the line connectors will be weighted by the similarities between the terms.
- **Maximum weight:** specifies the maximum thickness of the line connectors.
- **Nodes labels:** controls the size and color of the node labels.
- **Line labels:** controls the size and color of the labels on the line connectors.

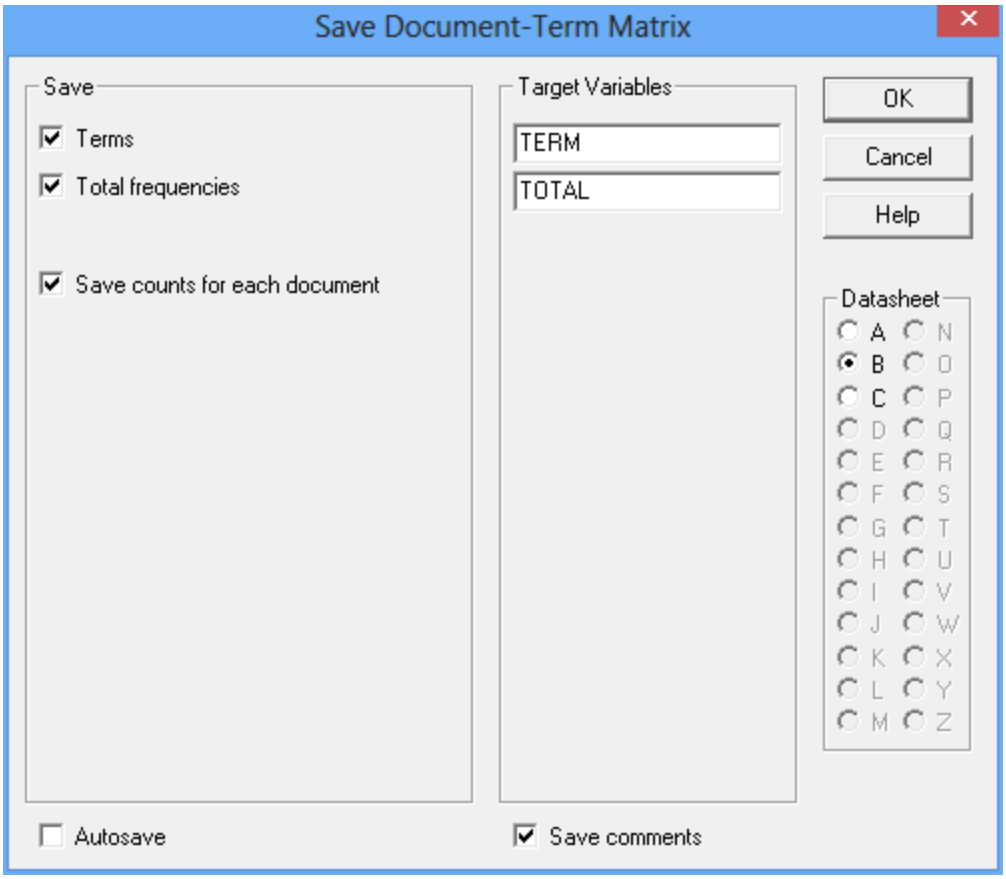
The diagram below illustrates the similarities amongst the word "nation" and other words:



The word with the distribution most similar to that of "nation' is "america".

Save Results

The document-term matrix may be saved in a Statgraphics datasheet by pressing the *Save Results* button on the analysis toolbar. The following dialog box will be presented:



To save the document-term matrix, select:

- **Save:** select the items to be saved. If "Save counts for each document" is checked, columns will be created containing the counts for each document in addition to the total count.

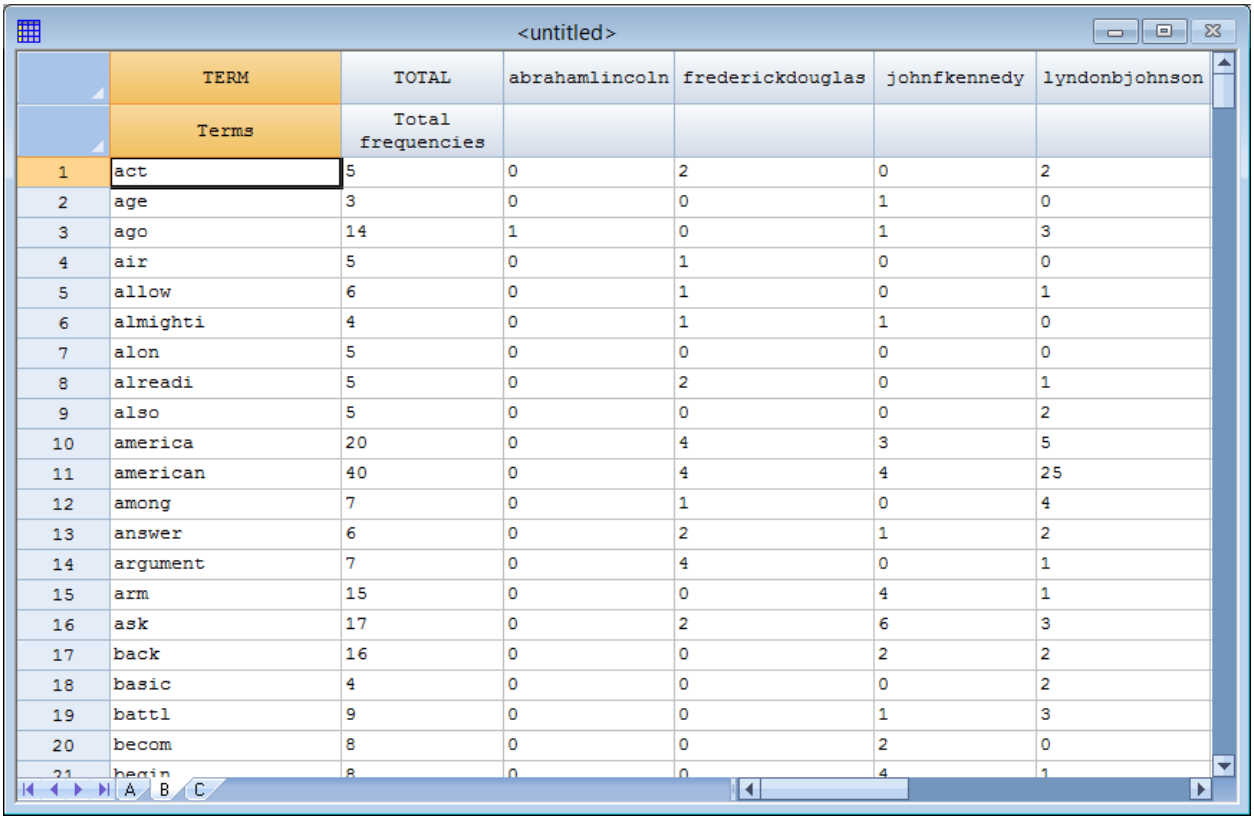
- **Target Variables:** enter names for the columns to be created.

- **Datasheet:** the datasheet into which the frequencies will be saved. Columns will be created containing each term in the matrix and the total count across all documents. If desired, separate counts may also be saved for each document. (Note: additional datasheets may be created if needed by selecting *Edit – Databook Properties* from the main menu.)

- **Autosave:** if checked, the document-term matrix will be saved automatically each time a saved StatFolio is loaded.

- **Save comments:** if checked, comments for each column will be saved in the second line of the datasheet header.

The datasheet below shows the saved document-term matrix for the 9 speeches:



<untitled>

	TERM	TOTAL	abrahamlincoln	frederickdouglas	johnfkennedy	lyndonbjohnson
	Terms	Total frequencies				
1	act	5	0	2	0	2
2	age	3	0	0	1	0
3	ago	14	1	0	1	3
4	air	5	0	1	0	0
5	allow	6	0	1	0	1
6	almighti	4	0	1	1	0
7	alon	5	0	0	0	0
8	alreadi	5	0	2	0	1
9	also	5	0	0	0	2
10	america	20	0	4	3	5
11	american	40	0	4	4	25
12	among	7	0	1	0	4
13	answer	6	0	2	1	2
14	argument	7	0	4	0	1
15	arm	15	0	0	4	1
16	ask	17	0	2	6	3
17	back	16	0	0	2	2
18	basic	4	0	0	0	2
19	battl	9	0	0	1	3
20	becom	8	0	0	2	0
21	begin	8	0	0	4	1

A B C

Example 2

The second example examines a list of 3,451 job openings in the City of New York. Using the default options, the wordcloud produced for this data is shown below:

The stem "bureau" occurs most often in the selected data column.

References

R Package “tm” (2015) https://cran.r-project.org/web/packages/tm/tm.pdf

R Package “wordcloud” (2015)
https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf

R Color Cheatsheet
https://www.nceas.ucsb.edu/~frazier/RSpatialGuides/colorPaletteCheatsheet.pdf.