Visual Representation of Text Data Sets Using the R *tm* and *wordcloud* Packages: Part One

Douglas M. Wiig

April 1, 2021

Abstract

This paper is the next installment in series that examines the use of R scripts to present and analyze complex data sets using various types of visual representations. Previous papers have discussed data sets containing a small number of cases and many variable, and data sets with a large number of cases and many variables. The previous tutorials have focused on data sets that were numeric. In this tutorial I will discuss some uses of the R packages *tm* and *wordcloud* to meaningfully display and analyze data sets that are composed of text. These types of data sets include formal addresses, speeches, web site content, Twitter posts, and many other forms of text based communication.

1 Introduction

I will present basic R script to process a text file and display the frequency of significant words contained in the text. The results include a visual display of the words using the size of the font to indicate the relative frequency of the word. This approach displays increasing font size as specific word frequency increases. This type of visualization of data is generally referred to as a "wordcloud." To illustrate the use of this approach I will produce a wordcloud that contains the text from the 2017 Presidental State of the Union Address.

There are generally four steps involved in creating a wordcloud. The first step involves loading the selected text file and required packages into the R environment. In the second step the text file is converted into a corpus file type and is cleaned of unwanted text, punctuation and other non-text characters. The third step involves processing the cleaned file to determine word frequencies, and in the fourth step the wordcloud graphic is created and displayed.

2 *Installing Required Packages*

As discussed in previous tutorials I would highly recommend the use of an IDE such as RStudio when composing R scripts. While it is possible to use the

basic editor and package loader that is part of the R distribution, an IDE will give you a wealth of tools for entering, editing, running, and debugging script. While using RStudio to its fullest potential has a fairly steep learning curve, it is relatively easy to successfully navigate and produce less complex R projects such as this one.

Before moving to the specific code for this project run a list of all of the packages that are loaded when R is started. If you are using RStudio click on the "Packages" tab in the lower right quadrant of the screen and look through the list of packages. If you are using the basic R script editor and package loader, at the command prompt use the following command:

```
################################################################################
>installed.packages()
################################################################################
```

The command produces a list of all currently installed packages. Depending on the specific R version that you are using the packages for this project may or may not be loaded and available. I will assume that they will need to be installed. The packages to be loaded are *tm*, *wordcloud*, *tidyverse*, *readr*, and *RColorBrewer*. Use the following code:

```
################################################################################
#Load required packages
################################################################################
install.packages("tm")   #processes data
install.packages("wordcloud")   #creates visual plot
install.packages("tidyverse")   #graphics utilities
install.packages("readr")       #to load text files
install.packages("RColorBrewer") #for color graphics
################################################################################
```

Once the packages are installed the raw text file can be loaded. The complete text of Presidential State of the Union Addresses can be readily accessed on the government web site `https://www.govinfo.gov/features/state-of-the-union`. The site has sets of complete text for various years that can be downloaded in several formats. For this project I used the 2017 State of the Union downloaded in text format. To load and view the raw text file in the R environment use the "Import Dataset" tab in the upper right quadrant of RStudio or the code below:

```
################################################################################
library(readr)
yourdatasetname <- read_table2("path to your data file", col_names = FALSE)
View(dataset)
################################################################################
```

3 *Processing The Data*

The goal of this step is to produce the word frequencies that will be used by wordcloud to create the wordcloud graphic display. This process entails converting the raw text file into a corpus format, cleaning the file of unwanted text, converting the cleaned file to a text matrix format, and producing the word frequency counts to be graphed. The code below accomplishes these tasks. Follow the comments for a description of each step involved.

```
################################################################################
#Take raw text file statu17 and convert to corpus format named docs17
################################################################################
library(tm)
docs17 <- Corpus(VectorSource(statu17))
################################################################################
################################################################################
#Clean punctuation, stopwords, white space
#Three passes create corpus vector source from original file
#A corpus is a collection of text
################################################################################
library(tm)
library(wordcloud)
data(docs17)
docs17 <- tm_map(docs17,removePunctuation) #remove punctuation
docs17 <- tm_map(docs17,removeWords,stopwords("english"))  #remove stopwords
docs17 <- tm_map(docs17,stripWhitespace)    #remove white space
################################################################################
#Cleaned corpus is now formatted into text document matrix
#Then frequency count done for each word in matrix
#dmat <-create matrix; dval <-sort; dframe <-count word frequencies
#docmat <- converts cleaned corpus to text matrix for processing
################################################################################
docmat <- TermDocumentMatrix(docs17)
dmat <- as.matrix(docmat)
dval <- sort(rowSums(dmat),decreasing=TRUE)
dframe <- data.frame(word=names(dval),freq=dval)
################################################################################
```

Once these steps have been completed the data frame "dframe" will now be used by the *wordcloud* package to produce the graphic.

4 *Producing the Wordcloud Graphic*

We are now ready to produce the graphic plot of word frequencies. The resulting display can be manipulated using a number of settings including color schemes,

number of words displayed, size of the wordcloud, minimum word frequency of words to display, and many other factors. Refer to *Appendix B* for additional information.

For this project I have chosen to use a white background and a multi-colored word display. The display is medium size, with 150 words maximum, and a minimum word frequency of two. The resulting graphic is shown in Figure One on page 6 of this document. Use the code below to produce and display the wordcloud:

```
#############################################################################################
#Final step is to use wordcloud to generate graphics
#There are a number of options that can be set
#See Appendix for details
#Use RColorBrewer to generate a color wordcloud
#RColorBrewer has many options, see Appendix for details
#############################################################################################
library(RColorBrewer)
set.seed(1234)  #use if random.color=TRUE
par(bg="white")     #background color
wordcloud(dframe$word,dframe$freq,colors=brewer.pal(8,"Set1"),random.order=FALSE,
scale=c(2.75,0.20),min.freq=2,max.words=150,rot.per=0.35)
#############################################################################################
```

As seen in Figure 1, the *wordcloud* display is arranged in a manner with the most frequently used words in the largest font at the center of the graph. As word frequency drops there are somewhat concentric rings of words in smaller and smaller fonts with the smallest font outer rings set by the *wordcloud* parameter min.freq=2. At this point I will leave an analysis of the *wordcloud* to the interpretation of the reader.

In part two of this tutorial I will discuss further use of the *wordcloud* package to produce comparison worddclouds using SOTU text files from 2017, 2018, 2019, and 2020. I will also introduce part three of the tutorial which will discuss using wordcloud with very large text data sets such as Twitter posts.

5 *Appendix A: Resources and References*

This section contains links and references to resources used in this project. For further information on specific R packages see the links below.

Package *tm*:

`https://cran.r-project.org/web/packages/tm/tm.pdf`

Package *RColorBrewer*:

`https://cran.r-project.org/web/packages/RColorBrewer/RColorBrewer.pdf`

Package *readr*:

`https://cran.r-project.org/web/packages/readr/readr.pdf`

Package *wordcloud*:
`https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf`
Package *tidyverse*
`https://cran.r-project.org/web/packages/tidyverse/tidyverse.pdf`

To download the RStudio IDE: `https://www.rstudio.com/products/rstudio/download/`

General works relating to R programming:

Robert Kabacoff, *R in Action: Data Analysis and Graphics With R*, Sheleter Island, NY: Manning Publications, 2011.

N.D. Lewis, *Visualizing Complex Data in R*, N.D. Lewis, 2013.

The text data for the 2017 State of the Union Address was downloaded from:
`https://www.govinfo.gov/features/state-of-the-union`

6 *Appendix B: R Functions Syntx Usage*

This appendix contains the syntax usage for the main R functions used in this paper. See the links in *Appendix A* for more detail on each function.

readr:

```
read_table2(file,col_names = TRUE,col_types = NULL,locale = default_locale(),na = "NA",
skip = 0,n_max = Inf,guess_max = min(n_max, 1000),progress = show_progress(),comment = "",
skip_empty_rows = TRUE)
```

wordcloud:

```
wordcloud(words,freq,scale=c(4,.5),min.freq=3,max.words=Inf,
random.order=TRUE, random.color=FALSE, rot.per=.1,
colors="black",ordered.colors=FALSE,use.r.layout=FALSE,
fixed.asp=TRUE, ...)
```

rcolorbrewer:

```
brewer.pal(n, name)
display.brewer.pal(n, name)
display.brewer.all(n=NULL, type="all", select=NULL, exact.n=TRUE,colorblindFriendly=FALSE)
brewer.pal.info
```

tm:

```
tm_map(x, FUN, ...)
```

All R programming for this project was done using RStudio Version 1.2.5033
This document was produced using TeXstudio 2.12.6
Author: Douglas M. Wiig 4/01/2021

Web Site: http://dmwiig.net

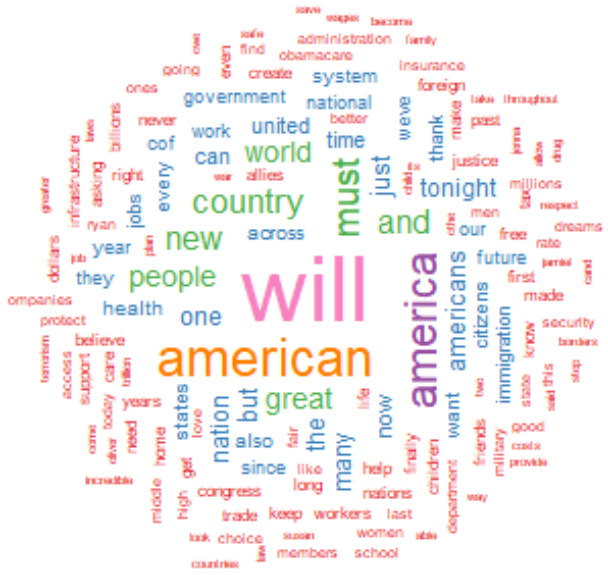

Figure 1: 2017 State of the Union Address Wordcloud