Lots of fun projects in the works, so today’s post will be short – a demonstration on how to create wordclouds, both with and without sentiment analysis results. While I could use song lyrics again, I decided to use a different dataset that comes with the quanteda packages: all 58 Inaugural Addresses, from Washington’s first speech in 1789 to Trump’s in 2017.

library(quanteda) #install with install.packages("quanteda") if needed

data(data\_corpus\_inaugural)  
speeches <- data\_corpus\_inaugural$documents  
row.names(speeches) <- NULL

As you can see, this dataset has each Inaugural Address in a column called “texts,” with year and President’s name as additional variables. To do analysis on the words in speeches, and generate a wordcloud, we’ll want to unnest the words in the texts column.

library(tidytext)  
library(tidyverse)

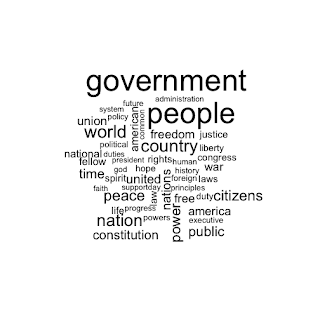
speeches\_tidy <- speeches %>%  
 unnest\_tokens(word, texts) %>%  
 anti\_join(stop\_words)

## Joining, by = "word"

For our first wordcloud, let’s see what are the most common words across all speeches.

library(wordcloud) #install.packages("wordcloud") if needed

speeches\_tidy %>%  
 count(word, sort = TRUE) %>%  
 with(wordcloud(word, n, max.words = 50))

[](https://i2.wp.com/3.bp.blogspot.com/-z7PnU0ZByRI/WxyIFAUVUdI/AAAAAAAALtg/2jXj3Nz5mxouFdWvgkayyMyXcVu0817UgCLcBGAs/s1600/unnamed-chunk-3-1.png?ssl=1)

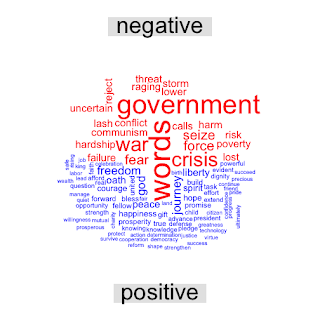
While the language used by Presidents certainly varies by time period and the national situation, these speeches refer often to the people and the government; in fact, most of the larger words directly reference the United States and Americans. The speeches address the role of “president” and likely the “duty” that role entails. The word “peace” is only slightly larger than “war,” and one could probably map out which speeches were given during wartime and which weren’t.

We could very easily create a wordcloud for one President specifically. For instance, let’s create one for Obama, since he provides us with two speeches worth of words. But to take things up a notch, let’s add sentiment information to our wordcloud. To do that, we’ll use the comparison.cloud function; we’ll also need the reshape2 library.

library(reshape2) #install.packages("reshape2") if needed

obama\_words <- speeches\_tidy %>%  
 filter(President == "Obama") %>%  
 count(word, sort = TRUE)  
  
obama\_words %>%  
 inner\_join(get\_sentiments("nrc") %>%  
 filter(sentiment %in% c("positive",  
 "negative"))) %>%  
 filter(n > 1) %>%  
 acast(word ~ sentiment, value.var = "n", fill = 0) %>%  
 comparison.cloud(colors = c("red","blue"))

## Joining, by = "word"

[](https://i2.wp.com/2.bp.blogspot.com/-K7S5kqCS2kc/WxyINVNEJZI/AAAAAAAALtk/qsKbECEwsn8KEsYDlYy4YqXWtjeR-RXcACLcBGAs/s1600/unnamed-chunk-4-1.png?ssl=1)

The acast statement reshapes the data, putting our sentiments of positive and negative as separate columns. Setting fill = 0 is important, since a negative word will be missing a value for the positive column and vice versa; without fill = 0, it would drop any row with NA in one of the columns (which would be every word in the set). As a sidenote, we could use the comparison cloud to compare words across two documents, such as comparing two Presidents. The columns would be counts for each President, as opposed to count by sentiment.

Interestingly, the NRC classifies “government” and “words” as negative. But even if we ignore those two words, which are Obama’s most frequent, the negatively-valenced words are much larger than most of his positively-valenced words. So while he uses many more positively-valenced words than negatively-valenced words – seen by the sheer number of blue words – he uses the negatively-valenced words more often. If you were so inclined, you could probably run a sentiment analysis on his speeches and see if they tend to be more positive or negative, and/or if they follow arcs of negativity and positivity. And feel free to generate your own wordcloud: all you’d need to do is change the filter(President == “”) to whatever President you’re interested in examining (or whatever text data you’d like to use, if President’s speeches aren’t your thing).