Ebb and Flow in Words of US Presidents Inaugural Speeches

For each term of presidency, the Inaugural addressed by each US president always becomes one of the most heated discussion at the time. Since it not only reflects the personality, the style of presidency of the chair-leader, but also reveals the policies they promised to take–corresponding to the policy demands under a unique historical background. The words used in each inaugural speech changes along the time and carries different information, which is quite interesting to take look at.

In this essay I scrapped some inaugural speeches from relevant website, and analyze the text of address from three angles:I firstly conduct an overview on the general inaugural speeches in US history, visualizing the features of sentiment and speech length on a 2-dimensional range. Then I compare the Sentence length over-time; For the third part I picked out several president in American history, explore the text structure in their addresses. At the end of the essay I summerized the changes of inaugural speech based on a historical angle.

Before text mining: Web Scrapping and Data Processing

Step(1): Load package I need for this project

```r
packages.used=c("rvest", "tibble","qdap",
                "sentimentr", "gplots", "dplyr",
                "tm", "syuzhet", "factoextra",
                "beeswarm", "scales", "RColorBrewer",
                "RANN", "tm", "topicmodels","tidytext", "rio",
                "wordcloud","shiny","ggrepel",
                "stringr","grid","testthat","magrittr")

# check packages that need to be installed.
packages.needed=setdiff(packages.used,
                        intersect(installed.packages()[,1],
                                  packages.used))

#Install additional packages
if(length(packages.needed)>0){
  install.packages(packages.needed, dependencies = TRUE,
                  repos='http://cran.us.r-project.org')
}
library(xml2)
library(rJava)
library("rvest")
library("tibble")
#dyn.load('/Library/Java/JavaVirtualMachines/jdk1.8.0_66.jdk/Contents/Home/jre/lib/server/libjvm.dylib'
#sudo ln -f -s $(/usr/libexec/java_home)/jre/lib/server/libjvm.dylib /usr/local/lib
library(qdap)
library(sentimentr)
library(gplots)
library(syuzhet)
library(factoextra)
library(beeswarm)
library(scales)
library(RANN)
library(topicmodels)
library(tm)
library(wordcloud)
library(RColorBrewer)
library(dplyr)
library(tidytext)
library(rio)
library(wordcloud)
library(shiny)
library(ggrepel)
library(stringr)
library(grid)
library(testthat)
library(magrittr)

source("/Users/YKP1993/Desktop/Fall\ 2017/ADS/pro1/wk2-TextMining/lib/speechFuncs.R")
source("/Users/YKP1993/Desktop/Fall\ 2017/ADS/pro1/wk2-TextMining/lib/plotstacked.R")
```

Step(2): Print my R version

```r
print(R.version)
```

```
##                 _
## platform       x86_64-apple-darwin13.4.0
## arch           x86_64
## os             darwin13.4.0
## system         x86_64, darwin13.4.0
## status
## major          3
## minor          3.3
## year           2017
## month          03
## day            06
## svn rev        72310
## language       R
## version.string R version 3.3.3 (2017-03-06)
## nickname       Another Canoe
```

Step(3): Extracting links from a list of speeches

```r
#Inauguaral speeches
main.page <- read_html(x = "http://www.presidency.ucsb.edu/inaugurals.php")
#Get link URLs
#f.speechlinks is a function for extracting links from the list of speeches.
inaug=f.speechlinks(main.page)
as.Date(inaug[,1], format="%B %e, %Y")
```

```
##  [1] "1789-04-30" "1793-03-04" "1797-03-04" "1801-03-04" "1805-03-04"
##  [6] "1809-03-04" "1813-03-04" "1817-03-04" "1821-03-04" "1825-03-04"
## [11] "1829-03-04" "1833-03-04" "1837-03-04" "1841-03-04" "1845-03-04"
## [16] "1849-03-05" "1853-03-04" "1857-03-04" "1861-03-04" "1865-03-04"
## [21] "1869-03-04" "1873-03-04" "1877-03-05" "1881-03-04" "1885-03-04"
## [26] "1889-03-04" "1893-03-04" "1897-03-04" "1901-03-04" "1905-03-04"
## [31] "1909-03-04" "1913-03-04" "1917-03-04" "1921-03-04" "1925-03-04"
## [36] "1929-03-04" "1933-03-04" "1937-01-20" "1941-01-20" "1945-01-20"
## [41] "1949-01-20" "1953-01-20" "1957-01-21" "1961-01-20" "1965-01-20"
## [46] "1969-01-20" "1973-01-20" "1977-01-20" "1981-01-20" "1985-01-21"
## [51] "1989-01-20" "1993-01-20" "1997-01-20" "2001-01-20" "2005-01-20"
## [56] "2009-01-20" "2013-01-21" "2017-01-20" NA
```

```
inaug=inaug[-nrow(inaug),] # remove the last line, irrelevant due to error.
```

Step(4): reshape the dataset

```
#Bulid president inaugural speech information index and corresponding url index
inaug.list=read.csv("/Users/YKP1993/Desktop/Fall\ 2017/ADS/pro1/wk2-TextMining/data/inauglist.csv", stri
speech.list=inaug.list
speech.list$type=c(rep("inaug", nrow(inaug.list)))
speech.url=inaug
speech.list=cbind(speech.list, speech.url)

#Loop over each row in speech.list
speech.list$fulltext=NA
for(i in seq(nrow(speech.list))) {
  text <- read_html(speech.list$urls[i]) %>% # load the page
    html_nodes(".displaytext") %>% # isloate the text
    html_text() # get the text
  speech.list$fulltext[i]=text
#Create the file name
  filename <- paste0("/Users/YKP1993/Desktop/Fall\ 2017/ADS/pro1/wk2-TextMining/data/fulltext/",
                     speech.list$type[i],
                     speech.list$File[i], "-",
                     speech.list$Term[i], ".txt")
  sink(file = filename) %>%
  cat(text)
  sink()
}
```

Now I've scrapped the inaugural speeches of all 58 presidents in US history! Let's start with text mining on their speeches!

Before we start, for the convinience of text mining, I extracted every sentence from full text of inauguaral speech, generate them as individual samples in a column called 'sentence', and extract some relevant information for each sentence, such as, 'word.count': refers to the number of words a sentence contains, 'sent.id': the position of each sentense in the complete speech. All these features elements are included in a datasets 'sentence.list'.

```
#Generate the list of sentenses
sentence.list=NULL
for(i in 1:nrow(speech.list)){
  sentences=sent_detect(speech.list$fulltext[i],
                        endmarks = c("?", ".", "!", "|",";"))
  if(length(sentences)>0){
    emotions=get_nrc_sentiment(sentences)
    word.count=word_count(sentences)
    # colnames(emotions)=paste0("emo.", colnames(emotions))
    # in case the word counts are zeros?
    emotions=diag(1/(word.count+0.01))%*%as.matrix(emotions)
    sentence.list=rbind(sentence.list,
                        cbind(speech.list[i,-ncol(speech.list)],
                              sentences=as.character(sentences),
                              word.count,
                              emotions,
                              sent.id=1:length(sentences)
                              )
    )
  }
}

sentence.list=
  sentence.list%>%
  filter(!is.na(word.count))
```

Considered there might be presidents who ran a second term. I wrote a function to create a new index 'President_Year' so as to distinct the speech delivered by the same president during different terms.

```
#build a function "getdate" to generate a new column join the string "president" and "year"
getdate<-function(x)
{
  substr(x, nchar(x)-3,nchar(x))
}
sentence.list$Year<-sapply(sentence.list$links,getdate)
sentence.list$President_Year<-paste(sentence.list$President," ",sentence.list$Year)
sentence.list$President_Year<-factor(sentence.list$President_Year)
sentence.list$President_YearOrdered=reorder(sentence.list$President_Year,
                                  sentence.list$word.count, mean, order=T)
```

Overview on Historical Inaugural Speech: Word Length and Sentiment

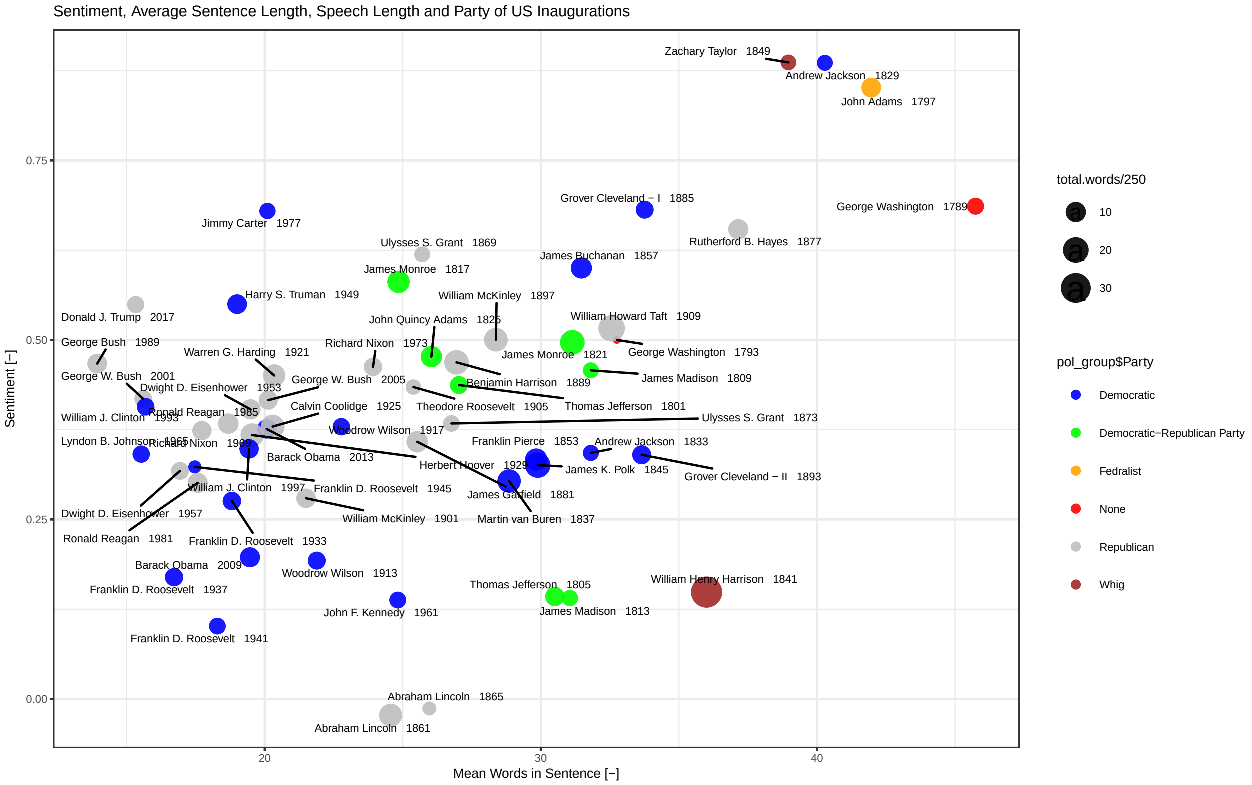
I first came up with an idea of overviewing the general features of inaugural speeches on a time axis, so that I can detect in which aspect the changes were made by the difference of presidency. Therefore, I presented all US president inauguration speeches in terms of sentiment, length of sentences, party of the president and length of speech in one graph.

Step(1): Reshape the data

```
pol <- polarity(sentence.list$sentences, sentence.list$President_Year)
pol_df <- pol$all
pol_df <- pol_df %>% dplyr::filter(!is.na(President_Year))
pol_df$pos.words <- NULL
pol_df$neg.words <- NULL
pol_group <- pol$group

speech.list$Year<-sapply(speech.list$links,getdate)
speech.list$President_Year<-paste(speech.list$President," ",speech.list$Year)
pres_party<-as.data.frame(cbind(speech.list$Party, speech.list$President_Year))
colnames(pres_party)<-c("Party", "President_Year")

#Get party information
pol_group <-merge(pol_group, pres_party, by = "President_Year")
pol_group$President_Year <- gsub("_", " ", pol_group$President_Year)
pol_group$Party <-as.character(pol_group$Party)
pol_group$Party[is.na(pol_group$Party)] <-"None"
pol_group$Party <-as.factor(pol_group$Party)
```

Step(2): Draw scatter plot

```
color_party <- c("blue", "green", "orange", "red", "grey", "brown")
g <-ggplot(pol_group, aes(x =total.words / total.sentences,
               y = stan.mean.polarity))
g <-g + geom_point(aes(color = pol_group$Party,
            size = total.words/250),
            alpha = .9)
g <-g + geom_text_repel(aes(x =total.words / total.sentences,
                y = stan.mean.polarity,
                label = factor(President_Year),cex=1.6))
g <-g + scale_color_manual(values = color_party)
g <-g + xlab ("Mean Words in Sentence [-]")
g <-g + ylab ("Sentiment [-]")
g <-g + ggtitle ("Sentiment, Average Sentence Length, Speech Length and Party of US Inaugurations")
g <-g+ theme(plot.title = element_text(size=rel(2)))
g <-g + theme_bw(base_size = 6)
g
```

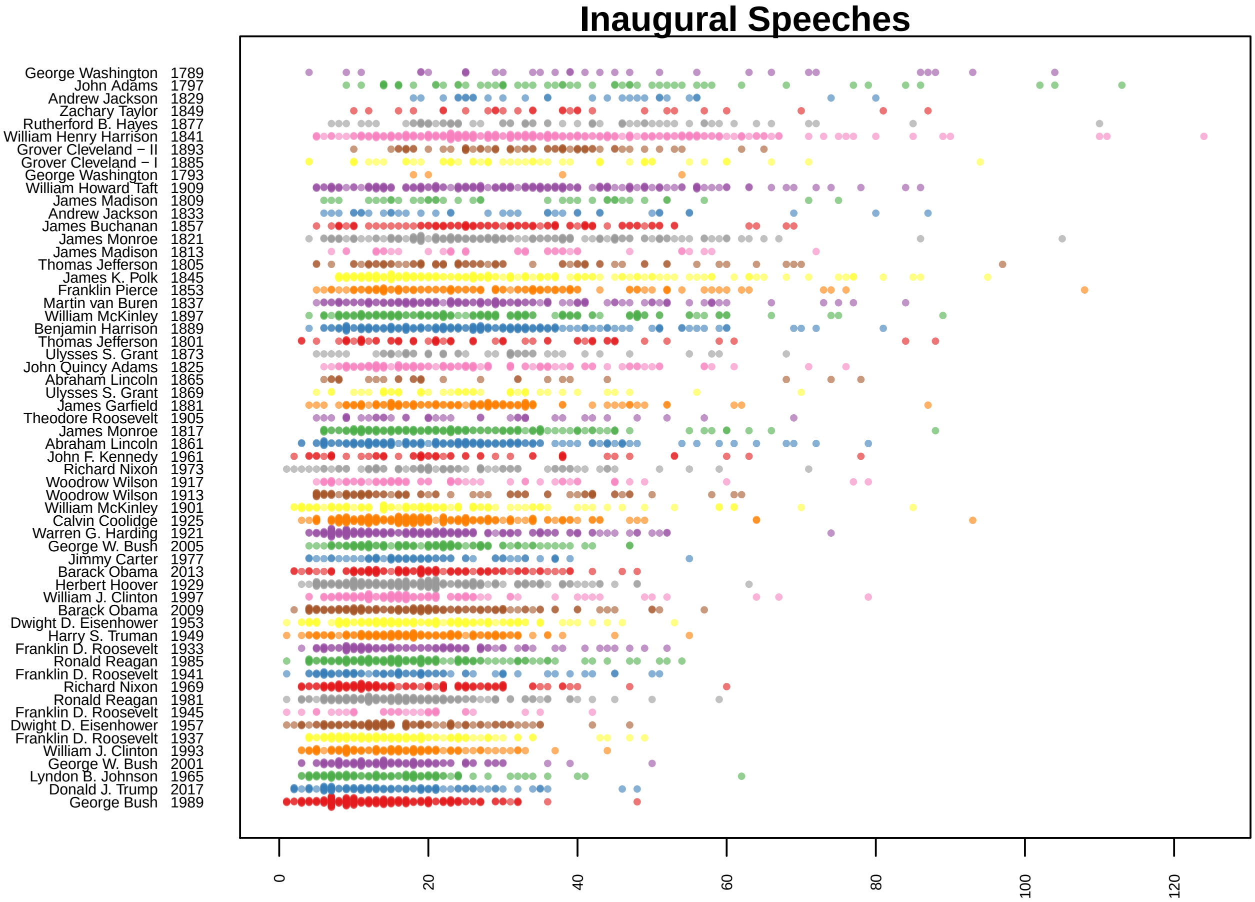


The outcome contains a large number of informations: The size of sentence is various when the speech giver changes. it is obvious that Trump uses the shortest sentences of all presidents. His speech is comparably short, but not the shortest speech. The first presidents used the longest sentences. This might mean that sentences got shorter with time. The sentiments also differ significantly in different speeches, but it has no apparent relations to which party the presidency belongs to.Such difference might just rely on the personality of the president. For example, Donald Trump, who's famous for having aggressive views on politics, has a higher numerical sentiment compared to Obama's speeches. Based on this observation I decided to look at the structure of inaugural speeches in both macro and micro angle, and explore the sentiment as an individual character, picked out some presidents from all and compare sentiment.

The Size of Speech: Shortened From Pass to Now

The size of speech refers to the length of speech and the length of each sentence. To explore the text structure of inaugural speech I mainly conduct comparison of the 'size' based on the later definition. Which presidents like to use long sentenses? And how about the distribution of long- short sentenses in the whole speech? To answer this question I visualized the sum of words in each sentences for every inaugural speech.

```r
par(mar=c(1.8, 11, 1.0, 1))
beeswarm(word.count~President_YearOrdered,
         data=sentence.list,
         horizontal = TRUE,
         pch=16, col=alpha(brewer.pal(9, "Set1"), 0.6),
         cex=0.55, cex.axis=0.5, cex.lab=0.6,
         spacing=5/nlevels(sentence.list$President_YearOrdered),
         las=2, ylab="", xlab="Number of words in a sentence.",
         main="Inaugural Speeches")
```



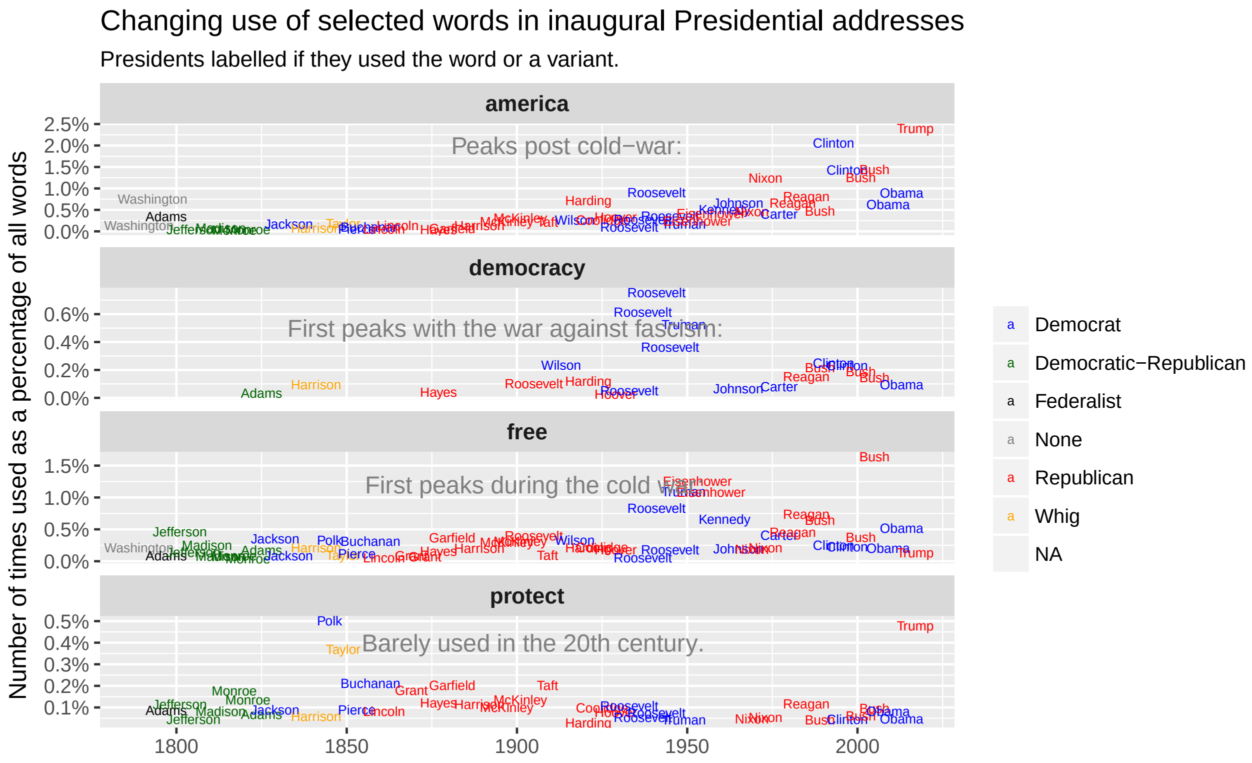
Based on R function beeswarm(), the plot "Inaugural Speeches" demonstrates the length of presidents' speech, for each ylabel(each inaugural speech), every individual point sums the number of words of a sentence in a speech. The ylabel was ranked according to the mean value of sentence length during the entire speech. The president with highest ylabel value delivered the longest average length of speech. The larger x value of the point, the more words in this sentence. As we can see from the plot, presidents William.Henry.Harrison got longest sentence among all historical inaugural speeches. President George.Washington, in his first term, delivered inaugural with longest mean value of sentence length in entire history.Presidents Lyndon.B.Johnson, George.Bush, George.W.Bush as well as the present president Donald.J.Trump, used much shorter sentences in their speeches. President George.Bush used shortest average length of sentence among all these president.

Keywords Overtime: Changing Resonances in US Politics

Here I took four words that have particular resonances in US politics: freedom, democracy, protection and America. I wanted to see how the popuplarity of these words changes in US inaugural speech.

```
parties <- c(None = "grey50", Federalist = "black", `Democratic-Republican` = "darkgreen",
             Whig = "orange", Republican = "red", Democrat = "blue")
namelist<-strsplit(as.character(speech.list$President)," ")
inauguralname<-rep(0,n=58)
for (i in 1:58)
{
  inauguralname[i]<-namelist[[i]][length(namelist[[i]])]
}
speech.list$inaugural<-paste0(as.character(speech.list$Year),"-",inauguralname)
inaugural<-data.frame(speech.list$fulltext, speech.list$inaugural)
colnames(inaugural)<-c("fulltext","inauguration")
inaugural$fulltext<-as.character(inaugural$fulltext)
inaugural$inauguration<-as.character(inaugural$inauguration)

#Combine with all the other speeches, and break down into tokens (words),
inaugural <-inaugural%>%
  mutate(year = as.numeric(str_sub(inauguration, 1, 4)),
         president = str_sub(inauguration, start = 6))%>%
  unnest_tokens(word, fulltext, token = "words") %>%
  group_by(inauguration) %>%
  mutate(sequence = 1:n())

#Aggregate / count how many occurances of word in each speech:
words <- inaugural %>%
   group_by(inauguration, word, year, president) %>%
   summarise(count = n()) %>%
   bind_tf_idf(word, inauguration, count) %>%
   ungroup()
expect_equal(nrow(inaugural), sum(words$count))

#Combine with the total count each word used in all speeches:
all_usage <- words %>%
   group_by(word) %>%
   summarise(total_count = sum(count)) %>%
   arrange(desc(total_count))

expect_equal(sum(all_usage$total_count), sum(words$count))

words <- words %>%
   left_join(all_usage, by = "word")

#Vector of all inaugurations (eg '1961-Kennedy'), use later for looping through:
inaugs <- unique(inaugural$inauguration)

#Time series chart
presidents <- read.csv("https://raw.githubusercontent.com/ellisp/ellisp.github.io/source/data/presidents
   filter(!is.na(year)) %>%
   select(inauguration, party)
```

```r
presidents$party[is.na(presidents$party)]<-"None"
annotations <- data_frame(word = c("america", "democracy", "protect", "free"),
                          lab = c("Peaks post cold-war:",
                                  "First peaks with the war against fascism:",
                                  "Barely used in the 20th century.",
                                  "First peaks during the cold war:"),
                          y = c(2, .5, 0.4, 1.2) / 100
)

words$inauguration<-as.character(words$inauguration)
par(mar=c(1,1,4,1))
words %>%
  mutate(word = ifelse(grepl("americ", word), "america", word),
         word = ifelse(grepl("democra", word), "democracy", word),
         word = ifelse(grepl("protect", word), "protect", word),
         word = ifelse(grepl("free", word), "free", word)) %>%
  group_by(inauguration, president, year, word) %>%
  summarise(count = sum(count)) %>%
  group_by(inauguration, president, year) %>%
  mutate(relative_count = count / sum(count)) %>%
  filter(word %in% c("america", "free", "democracy", "protect")) %>%
  left_join(presidents, by = "inauguration") %>%
  ggplot(aes(x = year, y = relative_count, label = president)) +
  geom_text(size = 2, aes(colour = party)) +
  facet_wrap(~word, ncol = 1, scales = "free_y") +
  ggtitle("Changing use of selected words in inaugural Presidential addresses",
          "Presidents labelled if they used the word or a variant.") +
  labs(x = "", y = "Number of times used as a percentage of all words") +
  scale_colour_manual("", values = parties) +
  scale_y_continuous(label = percent) +
  geom_text(data = annotations, x = 1935, aes(y = y, label = lab), colour = "grey50", hjust = 0.8) +
  theme(strip.text = element_text(size = 10, face = "bold"))
```



From this time-series chart, the keynotes of US politics passed back and forth between these four concepts. the keywords "America" first appeared in first president George.Washington's speech, to emphasize the concept of 'the nation'. This word peaks its usage frequency before the cold-war, at the presidency of Roosevelt, when the "America First" slogan appeared to keep the USA out of the European war against Nazism. The keynote featured very prominently in president Trump's address. Used 35 times out of 1,455 words, fully one word in 40 of Trump's was "America" or a variant. Bill Clinton, George W.Bush and Richard Nixon also had more than 1% of their words as "America" or variant. It is obvious that in the second sub-chart that 'democracy' is prominent in several of F.D.R Roosevelt's speeches during the war with Germany and Japan. The word experienced a revival at the end of the twentieth century in the period between the Cold War and the "War on Terror". The usage of "protect" first mounted its frequency in the speech of President Polk in 1845. The figure reveals the fact that protectionism was a hot topic at that time, and within twenty years it was an important contributing cause of the Civil War. "Free" as a word was most popular in Cold War speeches, with a second spike in George W. Bush's 2005 inauguration.

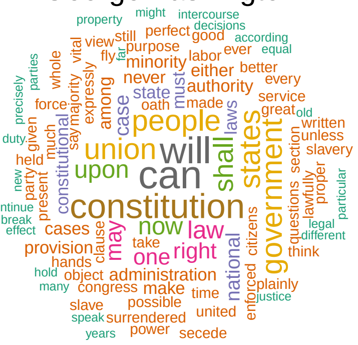
Unique Words of Particular Presidents:Some Interesting Snippets

Now I want to know the distinctive Words that are most frequently used in a particular inaugurals speech.So I pick out some president who were famous for their strong personality as well as their presidency policies: George.Washington, Abraham.Lincoln, Herbert.Hoover, JamesMadison, William.McKinley, Franklin.D.Roosevelt, Barack.Obama, Donald.J.Trump. For each president I plotted a wordclouds which present the most frequently appeared words in their speech. For the convience of observation, the larger size of a word, the higher frequency it was used in the speech.

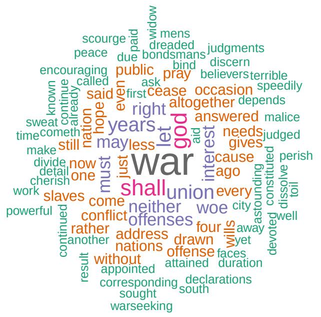
```r
#Draw the wordclouds for presidents
folder.path="/Users/YKP1993/Desktop/Fall\ 2017/ADS/pro1/wk2-TextMining/data/inaugurals/"
speeches=list.files(path = folder.path, pattern = "*.txt")
prex.out=substr(speeches, 6, nchar(speeches)-4)
ff.all<-Corpus(DirSource(folder.path))
```

```r
speaker<-c(speeches[18], speeches[30], speeches[1], speeches[54], speeches[25], speeches[12], speeches[
individual_folder<-rep(0,length(speaker))
par(mar=c(0.5, 8, 0.5, 1))
par(mfrow=c(4,2))
for(i in 1:length(speaker))
  {
    individual_folder[i]<-paste0(folder.path,'/',speeches[i])
     text<-readLines(individual_folder[i])
     docs<-Corpus(VectorSource(text))
     toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
     docs <- tm_map(docs, toSpace, "/")
     docs <- tm_map(docs, toSpace, "@")
     docs <- tm_map(docs, toSpace, "\\|")
     docs <- tm_map(docs, content_transformer(tolower))
     docs <- tm_map(docs, removeNumbers)
     docs <- tm_map(docs, removeWords, stopwords("english"))
     docs <- tm_map(docs, removeWords, c("blabla1", "blabla2"))
     docs <- tm_map(docs, removePunctuation)
     docs <- tm_map(docs, stripWhitespace)
      dtm <- TermDocumentMatrix(docs)
        m <- as.matrix(dtm)
        v <- sort(rowSums(m),decreasing=TRUE)
        d <- data.frame(word = names(v),freq=v)
      set.seed(1234)
      wordcloud(words = d$word, scale=c(2,0.5), freq = d$freq, min.freq = 1,
      max.words=100, random.order=FALSE, rot.per=0.3, use.r.layout=T,
      colors=brewer.pal(8, "Dark2"))
      text(x=0.5, y=1.0, substr(speaker[i], start=6, stop=nchar(speaker[i])-6),cex=1.5)
      text=NULL
      docs=NULL
      dtm=NULL
      m=NULL
      v=NULL
      d=NULL
    }
```

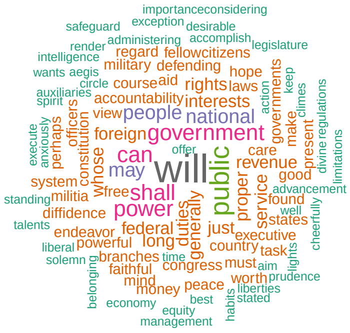
GeorgeWashington



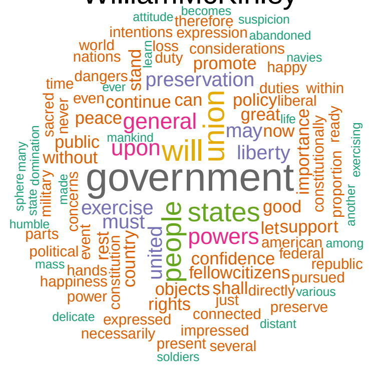
JamesMadison



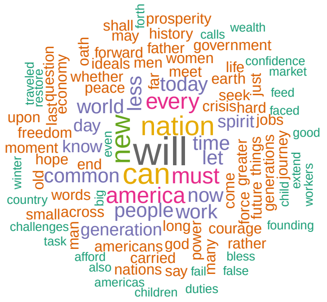
AbrahamLincoln



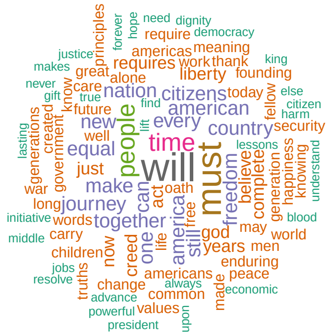
WilliamMcKinley



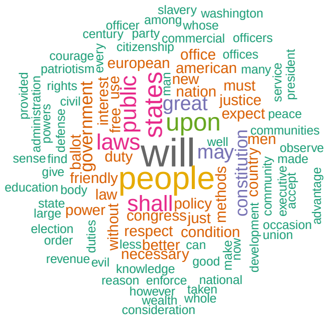
HerbertHoover



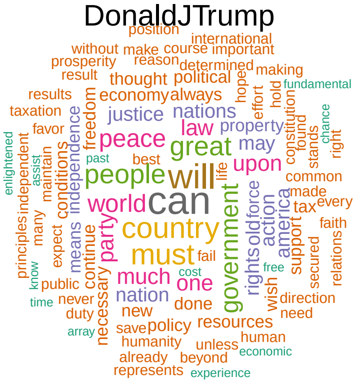
FranklinDRoosevelt



BarackObama



DonaldJTrump



Each wordcloud, to some extent, demonstrates words that draw resonances during the period of time.

In the first wordcloud, the first president delivered his speech by using a lot determinative words as ‘people’, ‘will’, ‘can’, to show his confidence of building up the well-being United States. Those imperative words has become the most commonly used tones in the later presidents' speeches. It is also important to note that, at this very start of US, he also used keynotes as 'constitution', 'law', 'union', 'goverment' so as to show his call for designing laws and regulations to ensure a strongly united presidency.

In the second wordcloud, Madison made considerable mention of the 1812 war with Britain, part of the great global confrontation of the Napoleonic wars which led to both Moscow and Washington in flames.

President Abraham.Lincoln, Herbert.Hoover emphasized more on the power of goverment and the importance of bringing benifits to citizens, as words like 'revenue', 'service', 'money','promote', appears relatively frequent in their speeches than others.

President Donald.J.Trump mentioned 'country' in his speech more frequently, which demonstrates the revival of 'America First' that first came up in 1940s to keep the USA out of the European war against Nazism.

Summary

The above analysis mainly shows us what information words contain in inaugural speeches. The length of the words tends to be shortern as we reviewing from past to now. This not only related to the personality of a certain president but also can be explained by the upcoming conciseness of the US literature.

From a two-demonsional map one remarkable fact is that the president donald Trump used more emotionally negetive words in his speech, which is commented by the press as breaking the traditional presidency.It to some extent expressed his own anger on sterotype of the US politics, which raised resonance from those not well-educated.

The keywords time-series chart, as well as wordcloud, demonstrates some latent political hotspot in US histories, and reveals what sort of resonances that public expected to get from president inaugurals: such as the wish of getting relief from war, from poverty and getting stable social status. The inaugural speeches helps government to show their determination of making a difference by repeating these words.