Project1-JingtianYao

Jingtian Yao

January 31, 2018

Project 1: The change of talking style over time and the difference between two Parties of American Presidents' inauguration speeches.

In this project, I want to figure out that whether there exists any significant change in the talking style of American Presidents' inauguration speeches over time. If it does, we are willing to give reasonable explanations to those phenomena. I also want to find difference between the talking style of Republican and Democratic.

The following is the abstract of this project.

Part I: Data Processing

In this part, we process the inauguration speeches text and make some data cleaning for further analysis.

Part II: Data Analysis

In theis part, we focus on the average lenth of sentences/ verbs/ sentiments in each speech and figure out whether there exists any significant changing pattern over time.

Part III: Conclusion

In this part, we state the conclusions obtained from the previous analysis.

```
packages.used <- c("ggplot2", "knitr", "plyr", "dplyr",
                   "ngram", "factoextra",  "readxl", "tidytext",
                   "wordcloud", "qdapDictionaries")

# 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(ggplot2)
library(knitr)
library(plyr)
library(dplyr)
library(ngram)
library(factoextra)
library(readxl)
library(tidytext)
library(wordcloud)
library(qdapDictionaries)
```

```
source("../libs/sentence.count.R")
source("../libs/extract.verb.R")
source("../libs/compute.DP.R")
```

This notebook was prepared with the following environmental settings.

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

```
##                _
## platform       x86_64-w64-mingw32
## arch           x86_64
## os             mingw32
## system         x86_64, mingw32
## status
## major          3
## minor          4.3
## year           2017
## month          11
## day            30
## svn rev        73796
## language       R
## version.string R version 3.4.3 (2017-11-30)
## nickname       Kite-Eating Tree
```

Part I: Data Processing

Above all, we process the text of each speech and store those information in the column `fulltext` in the dataset `speech.list`.

```
speech.list <- read_xlsx("../data/InaugurationInfo.xlsx")
speech.list$File <- paste(speech.list$File, speech.list$Term,
                          sep = "-")
speech.list$File<- paste("inaug", speech.list$File, ".txt",
                         sep = "")

folder.path="../data/InauguralSpeeches/"
speeches=list.files(path = folder.path, pattern = "*.txt")

temp <- list.files(path = folder.path, pattern="*.txt")
speech.list$fulltext <- NA
for(i in 1:length(temp)){
speech.list$fulltext[i]<-paste(readLines(paste(folder.path,
                               speech.list$File[i], sep = "")),
                               collapse = " ")
}
```

Part II: Data Analysis

1. Analysis – Average length of sentences

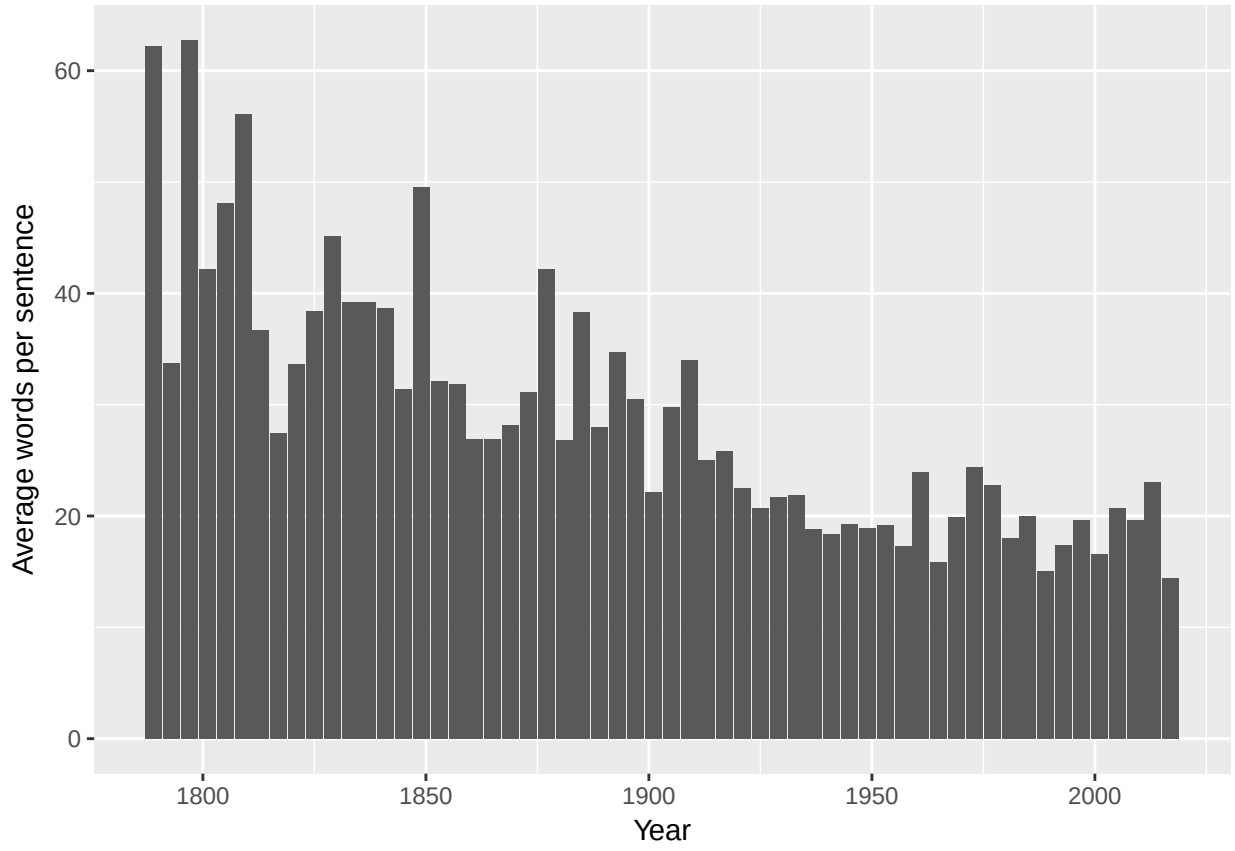
Firstly, we want to analyze the average number of words per sentence.

The number of words in Donald Trump's inauguration speech is missed out, so we need to count the words in his speech first. Then, we could compute the number of punctuations *"."*, *"?"* and *"!"* to figure out how many sentences there are in each text. We do this by the function `sentence.count` in *libs* folder.

```
# count the number of words in Trump's speech
speech.list$Words[nrow(speech.list)] <- wordcount(speech.list$fulltext[nrow(speech.list)])

# count the number of sentences in each speech
speech.list$Sentences <- aaply(speech.list$fulltext,1,sentence.count)

speech.list$Year <- seq(1789,2017,4) # label the years of inauguration

# count the average words per sentence in different years
avg.words.per.sentence <- as.numeric(speech.list$Words)/speech.list$Sentences
names(avg.words.per.sentence) <- speech.list$Year
```

We could make a plot to visualize the result.

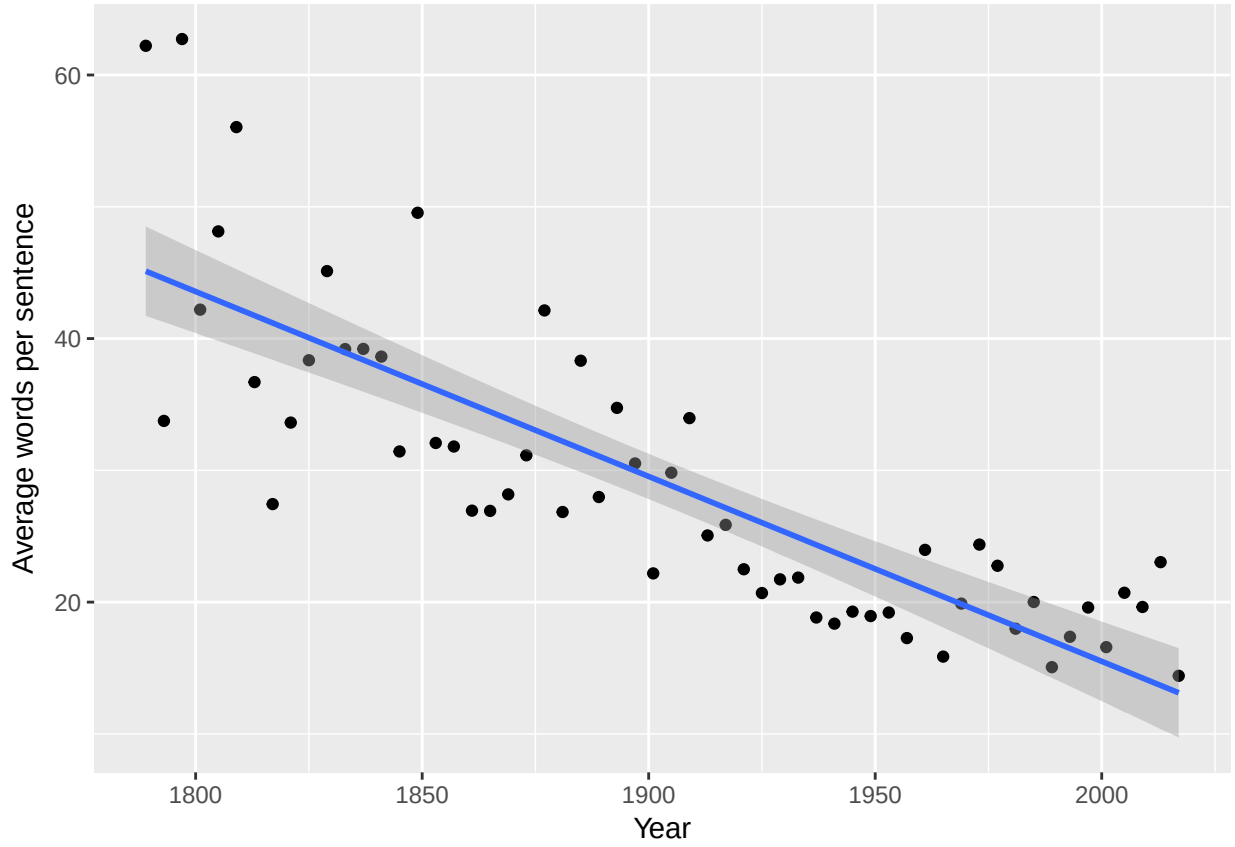
```
ggplot(speech.list,aes(x = Year, y = avg.words.per.sentence))+
  geom_bar(aes(x = Year, y = avg.words.per.sentence),stat = "identity") +
  labs(x = "Year",y = "Average words per sentence")
```



It's obvious that there exists a decreasing tendency for the average number of words per sentence over year. We could also draw a scatter plot to do further analysis.

```
ggplot(speech.list,aes(x = Year, y = avg.words.per.sentence)) +
  geom_point() +
  labs(x = "Year", y = "Average words per sentence") +
```

```
  geom_smooth(method = "lm",formula = y~x)
```



The blue line is the regression line.

```
fit <- lm(avg.words.per.sentence~speech.list$Year)
summary(fit)
```

```
##
## Call:
## lm(formula = avg.words.per.sentence ~ speech.list$Year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.744  -4.224  -1.260   3.937  18.740
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      296.14502   24.36259   12.16  < 2e-16 ***
## speech.list$Year  -0.14032    0.01279  -10.97 1.45e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.525 on 56 degrees of freedom
## Multiple R-squared:  0.6823, Adjusted R-squared:  0.6767
## F-statistic: 120.3 on 1 and 56 DF,  p-value: 1.448e-15
```

The p-value is extremely small, which means that the decreasing tendency of number of words per sentence

in the presidents' inauguration speech is confirmed.

This indicates that American Presidents are making the speeches more and more concise. One reason for this phenomenon may due to the fact that shorter sentences could express emotions stronger than longer ones. There is a trend that presidents are trying to motivate people with shorter sentences.

2. Analysis – Verbs

The `qdapDictionaries` package offers us the commonly used English verbs in a dataset named `action.verbs`. There are several verbs like "be", "make", "let", etc, which are widely-used but not informative, we do not consider those verbs. The dataset `stop_words` contain those meaningless verbs.

```
head(action.verbs)
```

```
## [1] "abduct"     "abide"      "abolish"    "abscond"    "abuse"
## [6] "accelerate"
```

```
action.verbs <- anti_join(data.frame(word = action.verbs),stop_words,by = "word")$word
```

We need to extract the verbs in each text. We do this by the function `extract.verb` in *libs* folder. The following is the example of the verbs used by George Washington.

```
head(extract.verb(speech.list[1,]))
```

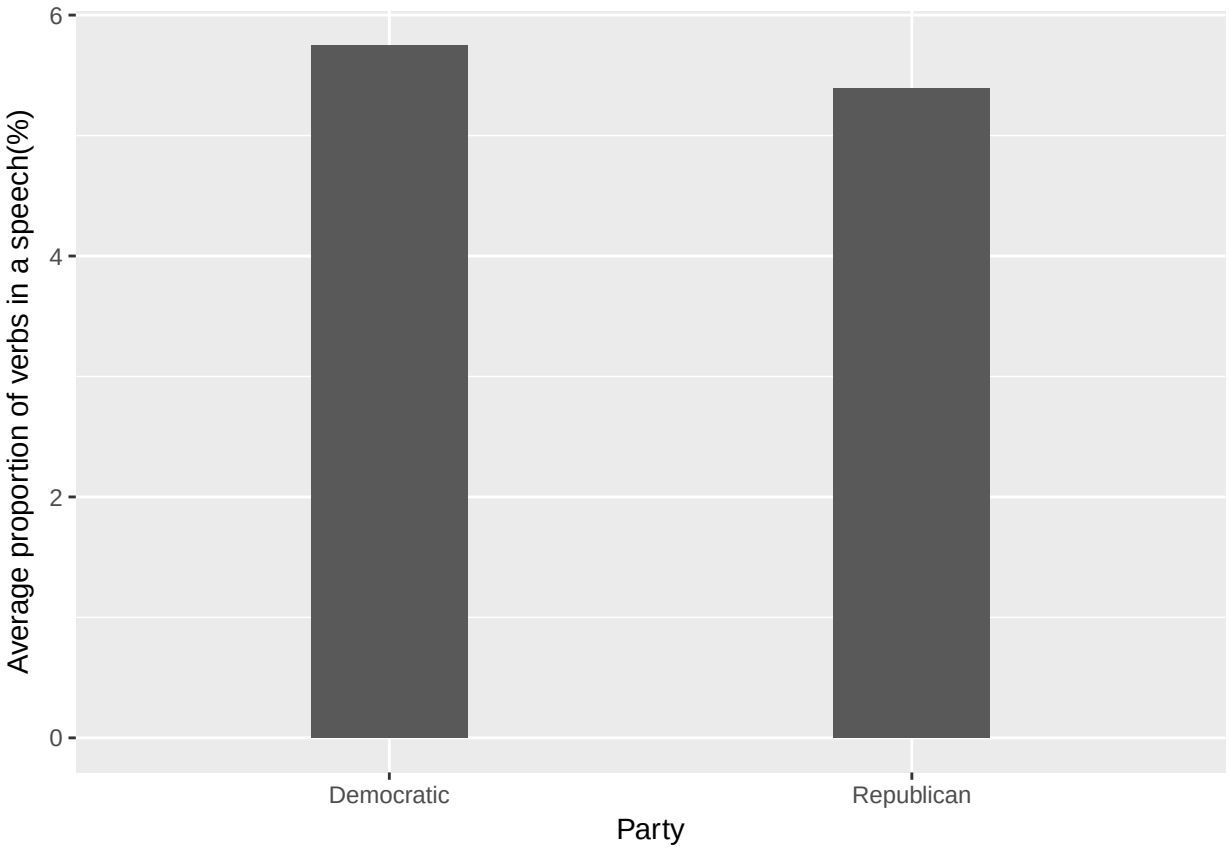
```
## [1] "act"    "add"    "adopt"  "assure" "assure" "author"
```

Next, we are intersted in whether we could find some difference between the verbs used by Democratic and Republican Party, to have an insight of their different policies. We consider the data after or in the year 1853.

```
Republican <- speech.list[speech.list$Party == "Republican" & speech.list$Year>=1853,]
Democratic <- speech.list[speech.list$Party == "Democratic" & speech.list$Year>=1853,]

Republican.verb <- extract.verb(Republican)
Democratic.verb <- extract.verb(Democratic)
```

We can visualize the average proportion of verb used in each speech by Republican and Democratic Presidents.

```
ggplot()+
  geom_bar(aes(x = c("Republican","Democratic"),
               y = 100*c(length(Republican.verb)/sum(as.numeric(Republican$Words)),
                     length(Democratic.verb)/sum(as.numeric(Democratic$Words))))
               ,stat = "identity",width = 0.3) +
  labs(x = "Party",y = "Average proportion of verbs in a speech(%)")
```



The average proportion of verb used in each speech by Democratic is larger than that by Republican, which may indicate that Democratic focuses more on actions while making speeches than Republican.

The verb word clouds for Republican and Democratic are as follow.

```
Republican.verb.freq <- count(data.frame(verb = Republican.verb),verb)
Democratic.verb.freq <- count(data.frame(verb = Democratic.verb),verb)

# Set word cloud
wordcloud(Republican.verb.freq$verb, Republican.verb.freq$n,
          max.words = 50, color = c("purple4", "red4", "black"))
```


```
wordcloud(Democratic.verb.freq$verb, Democratic.verb.freq$n,
          max.words = 50, color = c("purple4", "red4", "black"))
```

To have an insight of the verbs, we have a look at the most frequently used top 20 verbs by 2 parties. Then we could find some unique words for both parties.

```
Republican.verb.top20 <- Republican.verb.freq[order(Republican.verb.freq$n,
                                                    decreasing = TRUE),]$verb[1:20]
Democratic.verb.top20 <- Democratic.verb.freq[order(Democratic.verb.freq$n,
                                                    decreasing = TRUE),]$verb[1:20]
Top20.verb <- data.frame("Republican" = Republican.verb.top20,
                         "Democratic" = Democratic.verb.top20)

kable(Top20.verb)
```

Republican	Democratic
time	time
hope	change
progress	hope
party	land
secure	stand
service	service
force	question
seek	live
support	seek
question	progress
trade	promise
act	support
race	trade

Republican	Democratic
live	trust
love	act
meet	fear
hold	form
promise	meet
continue	set
bring	call

```
# verbs used commonly by 2 parties
same.verbs <- intersect(Top20.verb$Republican,
                        Top20.verb$Democratic)
kable(data.frame(same.verbs = same.verbs),
      col.names = "Same verbs used by 2 Parties")
```

Same verbs used by 2 Parties
time
hope
progress
service
seek
support
question
trade
act
live
meet
promise

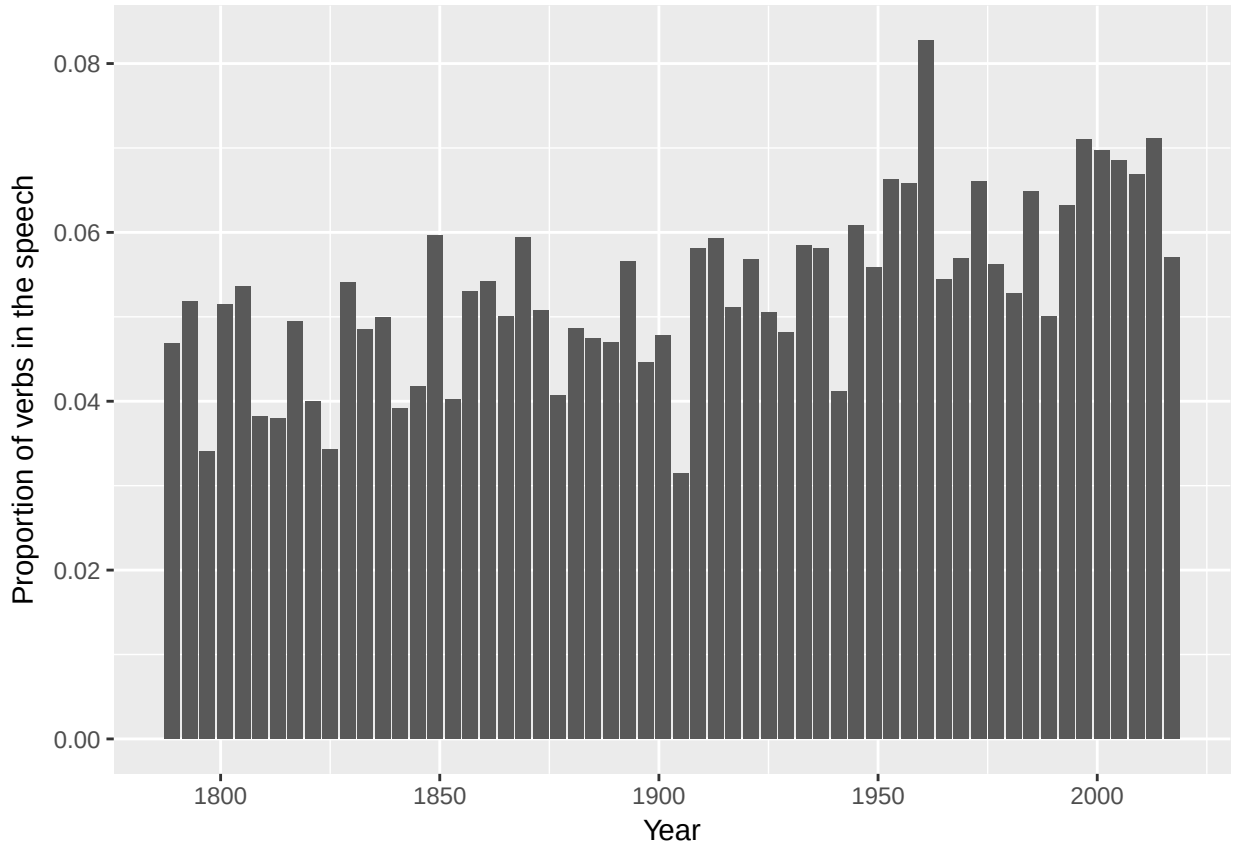
```
# unique verbs for 2 parties
Rep.uniq <- setdiff(Top20.verb$Republican,
                    Top20.verb$Democratic)
Dem.uniq <- setdiff(Top20.verb$Democratic,
                    Top20.verb$Republican)

kable(data.frame(Rep.uniq,Dem.uniq),
      col.names = c("Unique Words for Rep","Unique Words for Dem"),
      method = "markdown")
```

Unique Words for Rep	Unique Words for Dem
party	change
secure	land
force	stand
race	trust
love	fear
hold	form
continue	set
bring	call

Although some nouns are treated as verbs in the analysis, we can still find something interesting. It is shown that Republican always stresses words like "hold", "continue" and "force". It seems that Republican is more willing to stick to a plan. Meanwhile, Democratic prefers words like "change", "form" and "set", indicating that Democratic is more likely to making changes and trying something new.

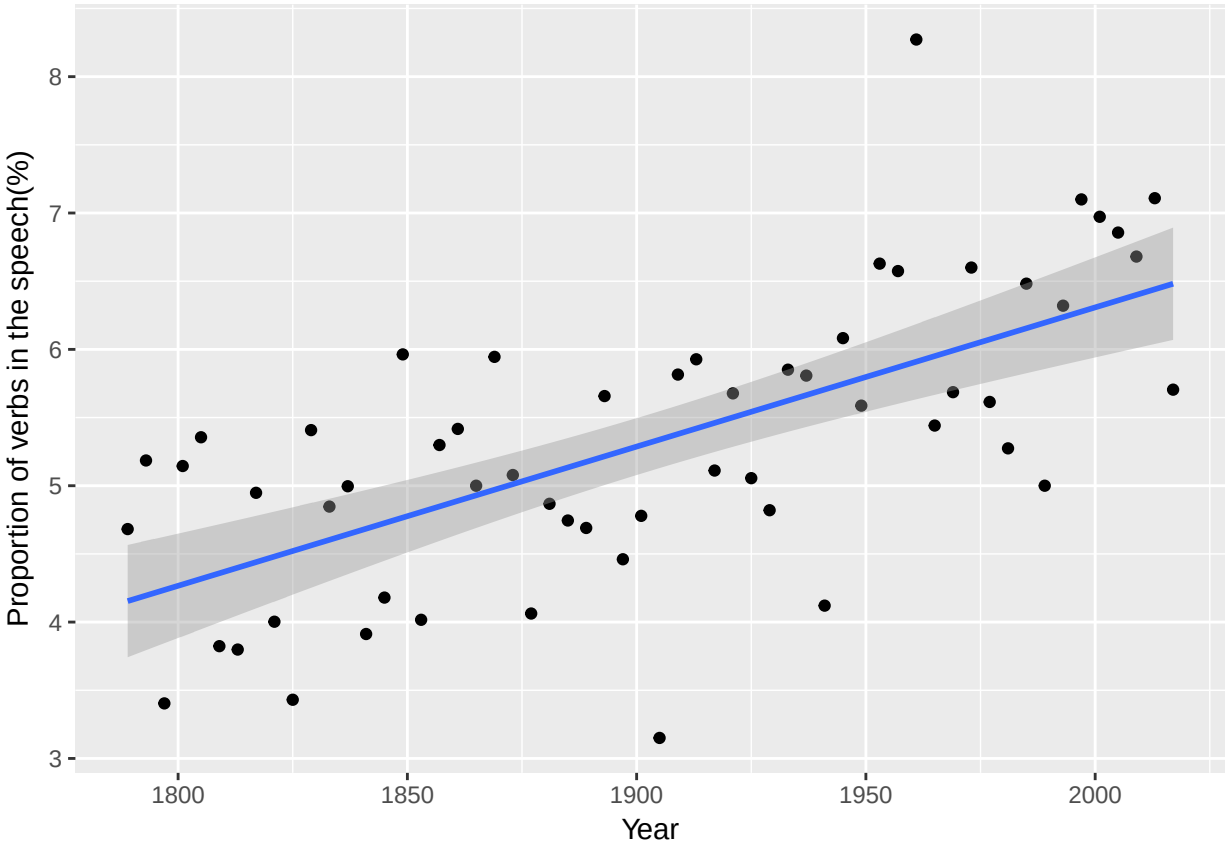
We are also interested in whether there exists a trend for the proportion of verbs changing over time.

```
# Store all verbs by years in a list
lst.verb.over.time <- dlply(speech.list,.(Year),extract.verb)

# count number of verbs of each year
count.verb.over.time <- laply(lst.verb.over.time,length)

# compute the proportion of verbs in each speech
prop.verb.over.time <- count.verb.over.time/as.numeric(speech.list$Words)

ggplot(data.frame("Year" = speech.list$Year, "Proportion" = prop.verb.over.time)) +
  geom_bar(aes(x = Year,y = Proportion),stat = "identity") +
  labs(x = "Year",y = "Proportion of verbs in the speech")
```



It seems that the proportion of verbs in the inauguration speech has an increasing tendency over time. To be more convincing, we make a scatter plot and try to build a regression model.

```
ggplot(data.frame("Year" = speech.list$Year, "Proportion" = prop.verb.over.time),
       aes(x = Year,y = 100*Proportion)) +
  geom_point() +
  labs(x = "Year", y = "Proportion of verbs in the speech(%)") +
  geom_smooth(method = "lm",formula = y~x)
```




```
fit <- lm(prop.verb.over.time~speech.list$Year)
summary(fit)
```

```
##
## Call:
## lm(formula = prop.verb.over.time ~ speech.list$Year)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.021876 -0.005306  0.001135  0.005229  0.023627
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.411e-01  2.960e-02  -4.767 1.37e-05 ***
## speech.list$Year  1.021e-04  1.554e-05   6.567 1.79e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.007927 on 56 degrees of freedom
## Multiple R-squared:  0.4351, Adjusted R-squared:  0.425
## F-statistic: 43.13 on 1 and 56 DF,  p-value: 1.791e-08
```

The p-value is extremely small, which means that the increasing tendency of proportion of verbs in the presidents' inauguration speech over time is confirmed.

This indicates that American Presidents are more and more willing to use verbs in their speeches. This may indicate that presidents focus more on what actions they are going to take and express their ideas more

directly without polishing their language too much.

3. Analysis – Sentiment

We can also analyze the sentiment in the presidents' inauguration speeches, to figure out whether the sentiment expressed in the text has some changing pattern. For each sentence, we apply sentiment analysis using NRC sentiment lexion.

```
head(get_sentiments('nrc'))
```

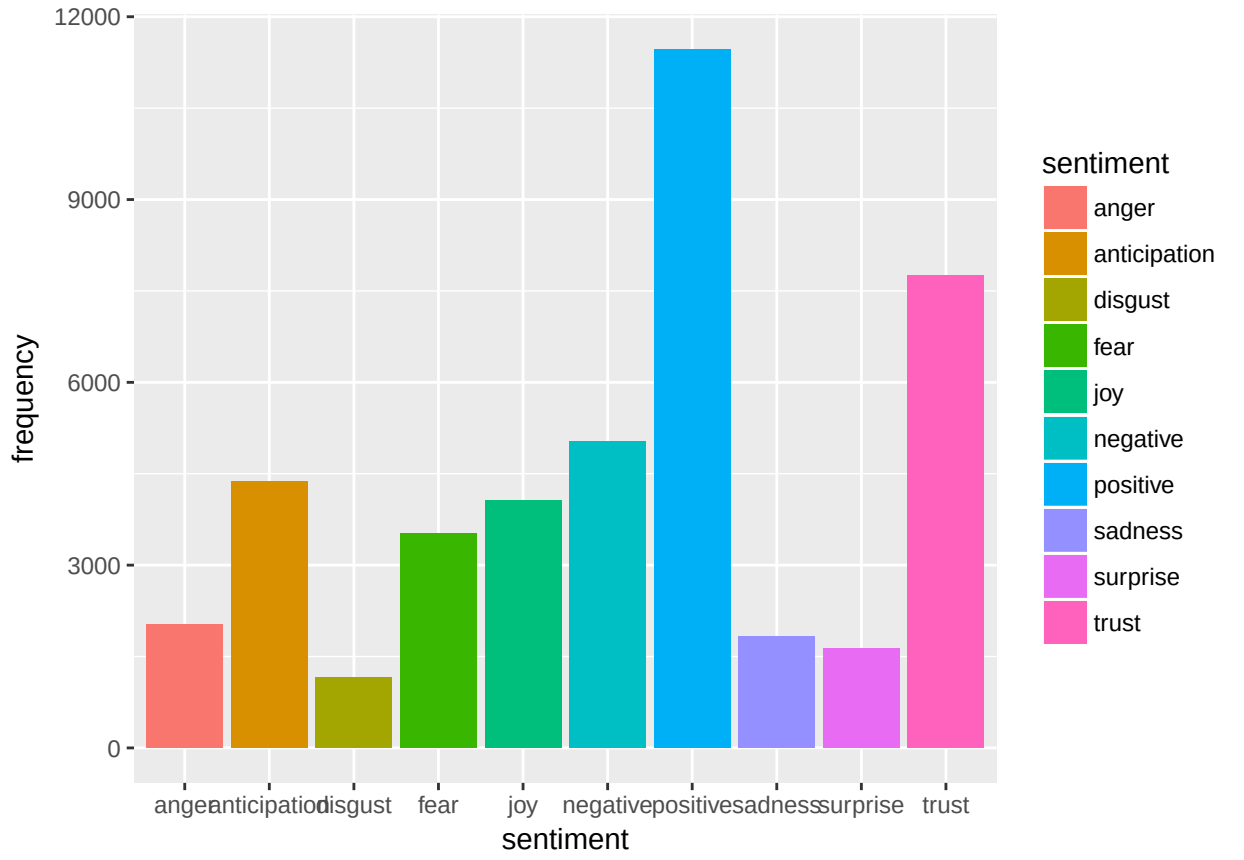
```
## # A tibble: 6 x 2
##   word      sentiment
##   <chr>     <chr>
## 1 abacus    trust
## 2 abandon   fear
## 3 abandon   negative
## 4 abandon   sadness
## 5 abandoned anger
## 6 abandoned fear
```

```
# compute the frequency of different sentiments in each year
sentiments <- inner_join(unnest_tokens(speech.list,word,fulltext), get_sentiments('nrc'), by = "word")
sentiments.count <- count(sentiments,Year,sentiment)
head(sentiments.count,10)
```

```
## # A tibble: 10 x 3
##     Year sentiment        n
##    <dbl> <chr>        <int>
##  1  1789 anger            9
##  2  1789 anticipation    49
##  3  1789 disgust          6
##  4  1789 fear            25
##  5  1789 joy             46
##  6  1789 negative        42
##  7  1789 positive       121
##  8  1789 sadness         13
##  9  1789 surprise        19
## 10  1789 trust           76
```

Firstly, we want to have a general image of the constitution of sentiments in presidents' inauguration speeches.

```
sentiments.count.all <- count(sentiments,sentiment)

ggplot(sentiments.count.all) +
  geom_col(aes(sentiment, n, fill = sentiment)) +
  labs(x = "sentiment",y = "frequency")
```

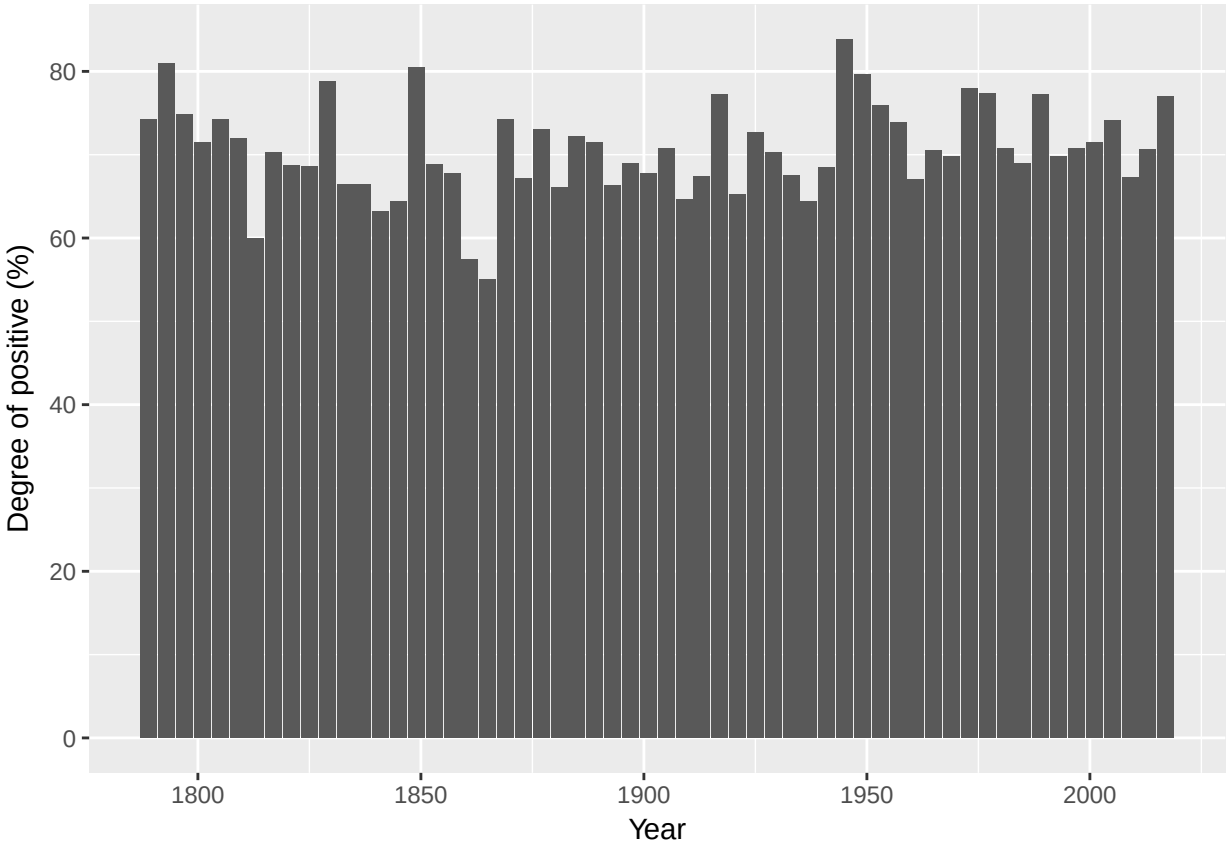


It's shown that besides "negative" and "positive" sentiments, top 3 sentiments are "trust", "anticipation" and "joy".

Next, we are insterested in whether we can find any significant changing pattern for presidents' sentimental constitution in inauguration speeches over time.

We consider the degree of "positive" (DP) sentiment. We measure this by the following way: $DP = \#positive/(\#positive + \#negative)$

```
DP <- daply(sentiments.count,.(Year),compute.DP)

ggplot(data.frame(DP)) +
  geom_bar(aes(x = speech.list$Year,y = 100*DP),stat = "identity") +
  labs(x = "Year",y = "Degree of positive (%)")
```



The plot shows that there is no significant difference among the degree of positive sentiment in all the presidents' speeches.

4. Clustering

From what has been discussed above, we know that there are two quantative indexes that are related to the time, which are the average words per sentene in a speech and the proportion of verbs in a speech. Now, we make clustering analysis based on these two indexes.

```
cluster.df <- data.frame("Year" = speech.list$Year,
                         "President" = speech.list$President,
                         "avg.words" = avg.words.per.sentence,
                         "prop.verb" = prop.verb.over.time)
km.res <- kmeans(cluster.df[,c(-1,-2)],iter.max = 200,4)
```

```
set.seed(123)
fviz_cluster(km.res, stand = TRUE,
             data = cluster.df[,c(3,4)],
             xlab = "Average words per sentence (standardized)",
             ylab = "Proportion of verbs in a speech (stadardized)",
             show.clust.cent=FALSE)
```

The plot shows that in each group, the interval between the corresponding "Years" are not large. Take group 4 as an example, all the years in this group are around 1850s. This indicates that the pattern of inauguration speech is changing as time goes by, and "average words per sentence" and "proportion of verbs in a speech" are two good indexes that could "measure" this changing to some extent.

Part III: Conclusion

1. The decreasing tendency of number of words per sentence in the presidents' inauguration speech is confirmed. This indicates that American Presidents are making the speeches more and more concise. One reason for this phenomenon may due to the fact that shorter sentences could express emotions stronger than longer ones. There is a trend that presidents are trying to motivate people with shorter sentences.
2. Republican always stresses words like "hold", "continue" and "force". It seems that Republican is more willing to stick to a plan. Meanwhile, Democratic prefers words like "change", "form" and "set", indicating that Democratic is more likely to making changes and trying something new.
3. The increasing tendency of proportion of verbs in the presidents' inauguration speech over time is confirmed. This indicates that American Presidents are more and more willing to use verbs in their speeches. This may indicate that presidents focus more on what actions they are going to take and express their ideas more directly without polishing their language too much.
4. There is no significant difference among the degree of positive sentiment in all the presidents' speeches.
5. The pattern of inauguration speech is changing as time goes by, and "average words per sentence" and "proportion of verbs in a speech" are two good indexes that could "measure" this changing to some extent.