Code

Project 1

KE HAN kh2793

Summary

Here, this notbook include 4 parts:

- Data preprocessing
- Sentimental and length analysis
- Wording analysis
    - **A**. Speech style
    - **B**. Wording to economy
- Summary

In the sentimental analysis and the length, I ploted all the information into a same plot. This plot shows the sentiment in the speech, and it generated some ideas for the further review.

Then, it comes to the wording analysis. It is quite similary to word cloud, since they both using the freqency of the words. And two parts are in it, the first is using stop words to analyzing speech style. And the second is using other data, like yearly GDP, GDP growth, financial market data and CPI to crosss analyze the presidental inaugration speech. The inuition is the inaugration speech would reflect the president's wish and goal which would benefit or feeled by Americans. And what's more, this topic interests me because in clustering it would show sentimental anlysis's value.

And in the wording to economy, some of the variable I used are derived from papers and articile I read before, some of the url link is in the reference.

Party/Change of party: one of the article I read analyzed the relationship of party change to stock trends(that part is really interesting, since market trends tell us ahead of time, which party would win). So, I used the SP500 index to indicate the stock market. Also, I know that some student use NYSE or other index, but that is not reflecting overall American. The diviation of small company/ small market might because of how Trump promised to SMB (small-midium size business), so I looked at NYSE but not included here.

(Notice: Some function takes more than 10 minutes to run, recommend using the test instead of the step1&step2 real file)

Step 0: load data & data cleaning & exploratory analysis

The step 0 is to prepare the data, including data collecting and data cleaning. Also, I counted words in the speech, get the somekind panel data. I find that working with numerical data would tell us something. What's more, numerical representitive of the speech would include more tools in analyzing.

And the methodology here, inculding left_join varibales and speech data, descriptive summary of speeches, reshape the data into data.frame which would be easy handle.

Counting the words with table function:

Hide

```
getwd()
```

```
[1] "/Users/HelenHan/Desktop/stat MA/Second Semester/5243 Applied Data Science"
```

Hide

```
#setwd("")
#get the word count for further analysis
#to avoid reproductivity problem, use "file.choose()"
Test<-readLines(file.choose())
```

```
incomplete final line found on '/Users/HelenHan/Desktop/stat MA/Second Semester/5243
Applied Data Science/InauguralSpeeches/inaugBarackObama-2.txt'
```

Hide

```
# to lower
Test<-tolower(Test)
#process, using , . and space to seperate the sentence
Test.words <- strsplit(Test, split = "[ |. |, |  ]")
Test_count<-table(Test.words)
Test_count<-sort(Test_count,decreasing = TRUE)
head(Test_count,10)
```

```
Test.words
      the  and  our   of   we   to that
 190  104   88   75   69   67   65   54
   a  for
  37   27
```

Hide

```
tail(Test_count)
```

```
Test.words
    work;   workers workforce     worst
         1         1         1         1
    young     youth
         1         1
```

Using this trunk of code, select file of speech, this function can give out any top words of any president's speech.

Left join the tables with financial data:

Hide

```
Test <-readLines(file.choose())
Financials<-readLines(file.choose())
Test <- left_join(Test, Financials, by = "year_president")
#some of the errors/unmatch in president speech time was corrected in the finance fil
e, so the following analysis would use step2.csv, which is a processed data set, cove
ring from 1930s, by when there is GDP, CPI and financial market data
```

Combine all speeches into one file, making it easy to compare others with Trump:

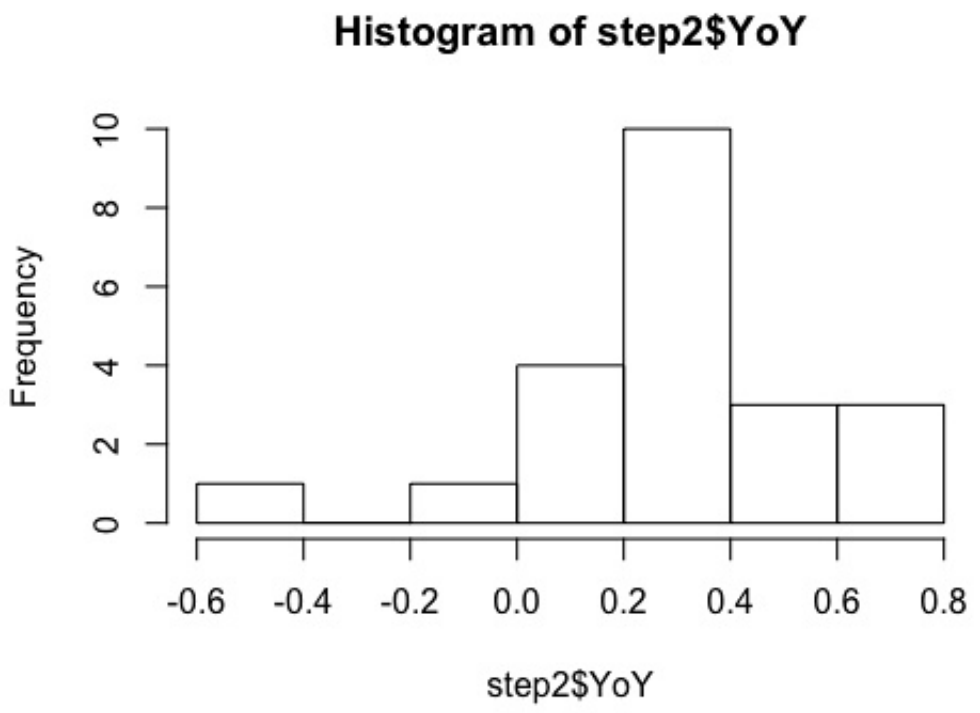
Hide

```
#files <- list.files("./data/InauguralSpeeches/")
#sometimes the previous line would not work, using file.choose would avoid problems
files <- list.files(file.choose())
files <- files[-length(files)]
lng_files <- length(files)

sentences <- data.frame(speech = rep(NA, lng_files),
            year_president = rep(NA, lng_files))
for (i in 1 : lng_files) {
    # import each individual speech
    file_path <- paste0("./data/", files[i])
    temp <- readLines(file_path)
    temp <- paste(temp, collapse = " ")  # concatenate all characters to one string
    sentences$speech[i] <- temp
}
#plot(step2$GDP)
#plot(step2$CPI)
```

A preview of the data we have, after those processing steps.

And some view of the GDP YoY can be seen here.



image

year	presi	speech
2017		Chief Justice Roberts, President Carter, President Clinton, Preside
2013		Thank you. Thank you so much. Vice President Biden, Mr. Chief Justi
2009		My fellow citizens, I stand here today humbled by the task before
2005		Vice President Cheney, Mr. Chief Justice, President Carter, Preside
2001		Thank you, all. Chief Justice Rehnquist, President Carter, Presider
1997		My fellow citizens, at this last Presidential Inauguration of the
1993		My fellow citizens, today we celebrate the mystery of American ren
1989		Mr. Chief Justice, Mr. President, Vice President Quayle, Senator M
1985		Senator Mathias, Chief Justice Burger, Vice President Bush, Speake
1981		Senator Hatfield, Mr. Chief Justice, Mr. President, Vice President
1977		For myself and for our Nation, I want to thank my predecessor for

President	time	Term	Party	Words	content	year	GDP	YoY	Financial	CPI
Donald J.	1/20/17	1	Republican	1455	Chief Jus	2017	18,860.80	0.12996435	2280	238.6575
Barack Oba	1/21/13	2	Democratic	2096	Thank you.	2013	16,691.50	0.15762864	1606	232.965
Barack Oba	1/20/09	1	Democratic	2395	My fellow	2009	14,418.70	0.1011937	1057	214.565833
George W. I	1/20/05	2	Republican	2071	Vice Pres	2005	13,093.70	0.2327195	1191	195.266667
George W. I	1/20/01	1	Republican	1592	Thank you,	2001	10621.8	0.2338735	1224	177.041667
William J.	1/20/97	2	Democratic	2155	My fellow	1997	8608.5	0.25147194	885	160.525
William J.	1/20/93	1	Democratic	1598	My fellow	1993	6878.7	0.21581208	451	144.475
George Bus	1/20/89	1	Republican	2320	Mr. Chief	1989	5657.7	0.30160812	317	123.941667
Ronald Rea	1/21/85	2	Republican	2561	Senator Ma	1985	4346.7	0.35369044	185	107.6
Ronald Rea	1/20/81	1	Republican	2427	Senator Ha	1981	3211	0.53930968	131	90.9333333
Jimmy Cart	1/20/77	1	Democratic	1229	For mysel	1977	2086	0.46027301	100	60.6166667
Richard Ni	1/20/73	2	Republican	1803	I, RICHARI	1973	1428.5	0.40062751	104	44.425
Richard Ni	1/20/69	1	Republican	2128	Senator D	1969	1019.9	0.37138631	93	36.6833333
Lyndon B.	1/20/65	1	Democratic	1507	My fellow	1965	743.7	0.32025564	90	31.5283333
John F. Ke	1/20/61	1	Democratic	1366	Vice Pres	1961	563.3	0.18614445	67	29.9016667
Dwight D. I	1/21/57	2	Republican	1658	Mr. Chairr	1957	474.9	0.21862972	44	28.1133333
Dwight D. I	1/20/53	1	Republican	2459	My friend	1953	389.7	0.71071115	29	26.7683333
Harry S. T	1/20/49	1	Democratic	2273	[Delivere	1949	227.8	-0.0017528	18	23.8091667

Some findings:

- The GDP YoY can help us to devide all the speeches into 4 category.
- The speech data can be analyzed using polarity, which is a direct way to get information.
- Comparision of wording, cross analysis of data from other sources to further analyze

Step 1: analysis of the whole speech & visualization

The goal of this part is to integrate the analysis of sentiment and length. Some of the thought is derived from the Smart Data with R webpage. Also, join the information from presidentinfo, to get the party infomation.

Here, all the steps would help us to get to the speech basic text analysis, including average number of words per sentence, sentiment, party membership and length of speech.

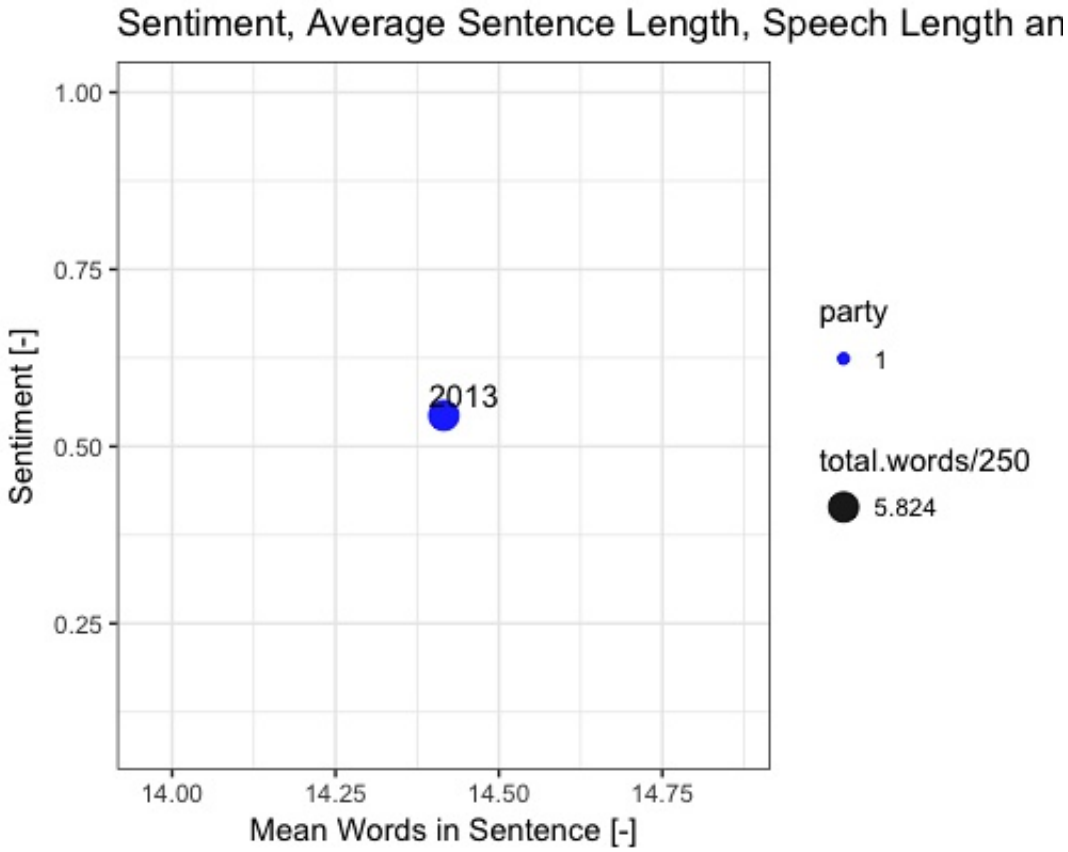
The steps are:

- Library
- Data Import
- Sentence split & left_join of difference data source& sentimental analysis
- Data visulization

Also, the first trunk is the analysis of Trump only. And the next includ all presidents.

Hide

```
library(rJava)
suppressPackageStartupMessages(library(qdap))
suppressPackageStartupMessages(library(dplyr))
suppressPackageStartupMessages(library(ggplot2))
library(ggrepel)
library(rio)
# test part, using small file to run the whole process
#if this one not working, please use all<-readLines(file.choose())
t <- data.frame(speech = rep(NA, 1),
                year_president = rep(NA, 1))
t$speech[1]=Trump
t$year_president[1]=2017
tt <- sentSplit(t, "speech", verbose = F)
pol <- polarity(tt$speech, tt$year_president)
pol_df <- pol$all
pol_df <- pol_df %>% dplyr::filter(!is.na(year_president))
pol_df$year_president <- as.factor(pol_df$year_president)
pol_df$pos.words <- NULL
pol_df$neg.words <- NULL
pol_group <- pol$group
pol_group$party <- as.factor(1)
# Polarity vs. Mean Sentence Length
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 = party,
            size = total.words/250),
            alpha = .9)
g <- g + geom_text_repel(aes(x =total.words / total.sentences,
                y = stan.mean.polarity,
                label = factor(year_president)))
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_bw()
g
```



image

The previous image is the one with Trump only, and it is all about sentiment, average sentence/speech.

This is the whole process, which would help to draw the informative picture.

polarity() can be found here. polarity score (https://www.rdocumentation.org/packages/qdap/versions/2.2.5/topics/polarity)

Hide

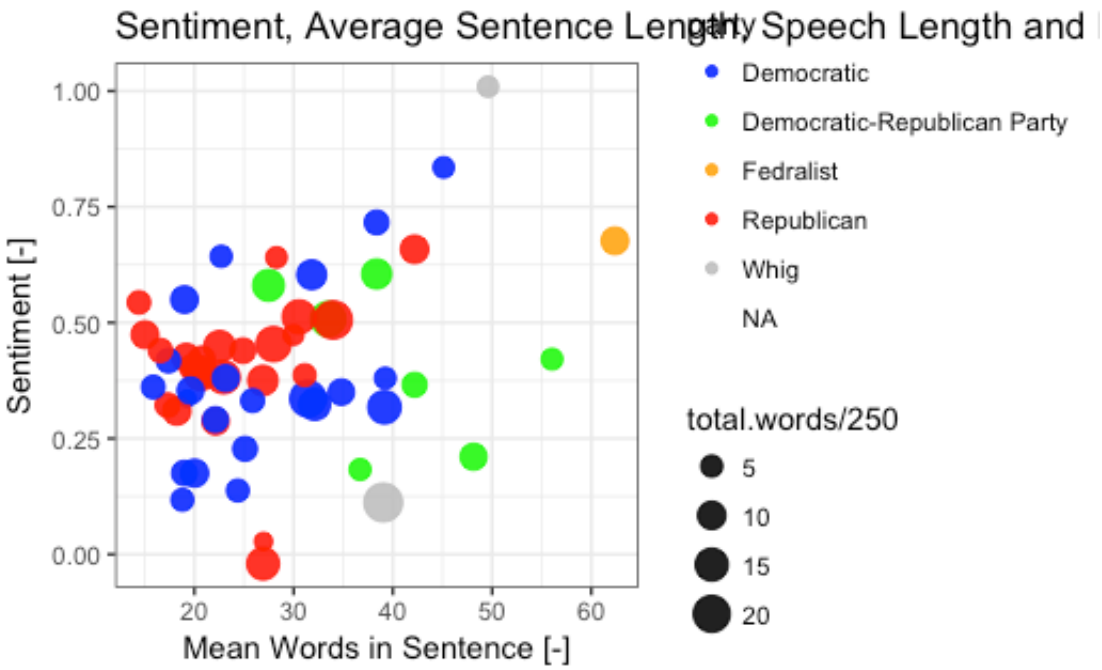
```
#real process, with the whole speech and the whole run
# Import Data
# using the speech data from step0
# sentences
all<-readLines("all.txt")
sentences <- data.frame(speech = rep(NA, 58),
            year_president = rep(NA, 58))
for(i in 1:58){
  sentences$year_president[i]=2017-4*(i-1)
  sentences$speech[i]=all[i]
}
sentences <- sentSplit(sentences, "speech", verbose = F)

#polarity() is a function that provides sentiment analysis
# polarity function would last for a few minutes
pol <- polarity(sentences$speech, sentences$year_president)
pol_df <- pol$all
pol_df <- pol_df %>% dplyr::filter(!is.na(year_president))
pol_df$year_president <- as.factor(pol_df$year_president)
pol_df$pos.words <- NULL
pol_df$neg.words <- NULL
pol_group <- pol$group
```

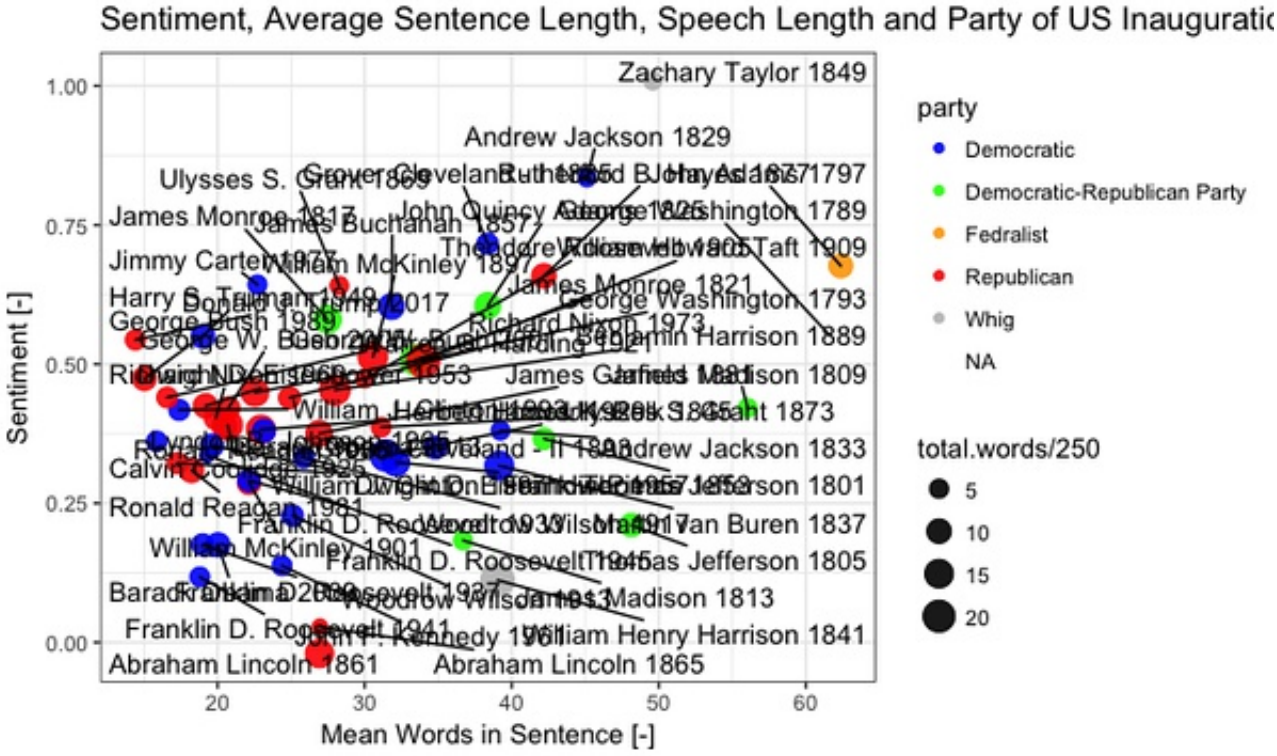
Here, it is quite clear Sentiment numerical result can be used afterwards in trend analysis. Package polarity() is very useful here.

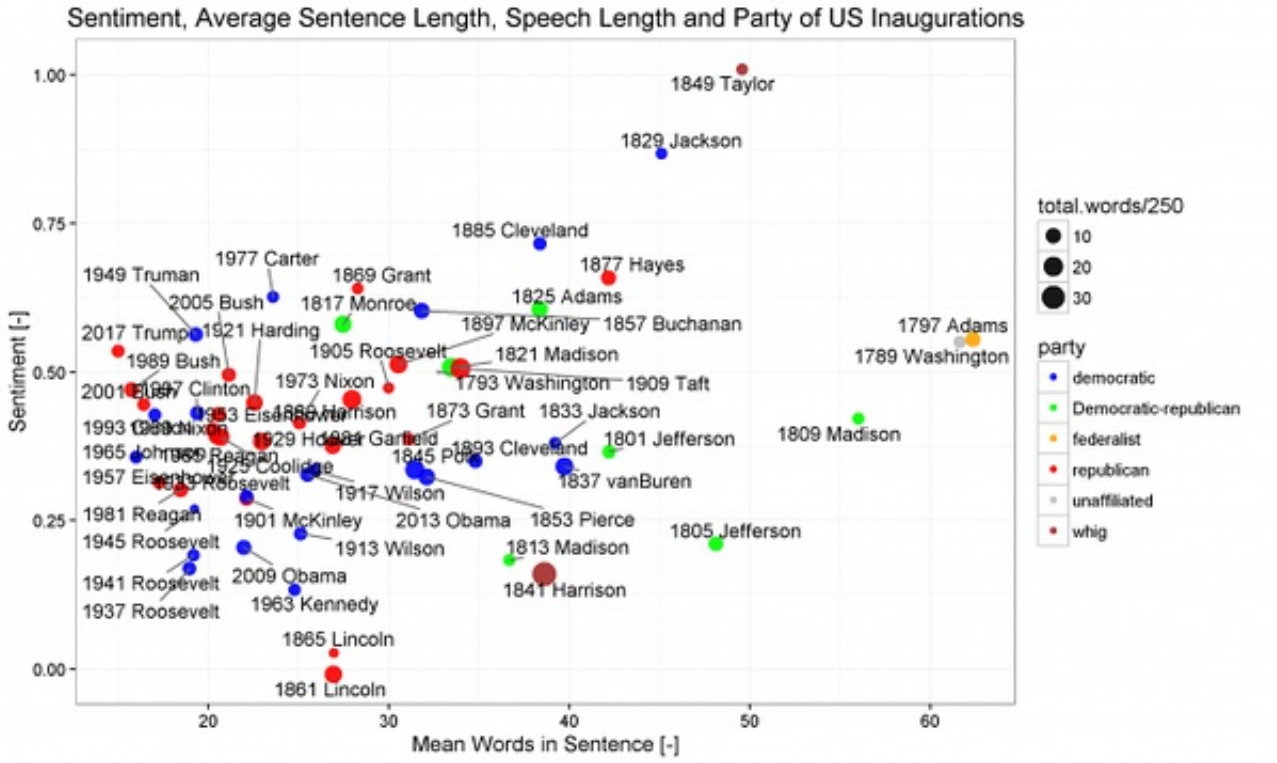
Hide

```
# get party information
#pol_group <- left_join(pol_group, pres_party, by = "year_president")
#pol_group$party <- as.factor(pol_group$party)
# the join here, sometimes not work out, so, can use step0, which is already processe
d
pol_group$year_president <- gsub("_", " ", pol_group$year_president)
# Polarity vs. Mean Sentence Length
pol_group<-read.csv("step0.csv")
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 = 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)))
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_bw()
g
```



Here, the text is a little bit large to fit in. So, the following 2 plot shows the adjusted graphics.





Some findings:

- For different party, the difference is quite obvious. Especially the sentimental ones, since polarity() shows the sentiment (polarity) of text by grouping variable(s). Actucally Maybe clustering would make sense here. Also, this is a proof to one of the articles I read. Democrat vs. Republican

(http://www.diffen.com/difference/Democrat_vs_Republican) And I think this is part of the common point I am looking for, that is althongh time change, president stayed same. Further analysis see part 2.

- Trump uses the shortest sentences of all presidents, and his speech has a higher numerical sentiment compared to both Obama speeches. His speech is very short(but not the shortest one).
- Time trend exists. Adams speech in 1797 used longest sentneces. As the discussion in class, media trend and how people comunicate evolve. I also read some paper about this, one of the reason is that new words are coming up, some words mean more than one concept. See the link for more information. It also analyzed the complexity of the sentence, the structure and the visulization of sentences., showing same pattern. Stylistic analysis (http://cdmd.cnki.com.cn/Article/CDMD-10487-2009036843.htm) Writer use shorter sentences (https://www.reference.com/education/writers-use-short-sentences-740f5aad12fb3b24)
- Lastly, this is a inuitively grapgic to further analysis on interesting topics for business purpose, and thus further test is needed here. Such as sentences getting shorter, Twitter can switch to shorter text. Instagram getting popular, since people are more interested in pictures and shorter sentences, so there is no need for long sentences.

Step 2: analysis of the speech style

Here, I define the speech as the style, which is represented as stopwords. As we start to learn English, we were taught that WOULD WILL SHOULD SHALL COULD CAN all differ. And one of the feeling I have here, is that people here in US are really polite and respective to people around, thus using less pushing words. Finding on how stopword differ is interesting in this sense.

What mainly in this part, is the comparision (using cosine similarity) between the central to all president and how Trump is different from others. And the graph in the attached file show the clear distinction between Trump and others.

Actually, Trump enjoyed using WE and OUR, I think this is a way to gain trust from the others. And it is somehow an exaggeration. Since, there is some many fighting back voice, WE and OUR is a way to get near to others, showing that everyone is on the same boat. Just like Obama, “Together we can!”, closer relationship with everyone. Obama and Trump both use WE and OUR more than other presidents.

Hide

```
# all about Trump Speech style
# if needed, truning sentences into one long string
# HC <- paste(HC, collapse = " ")
# load
Trump<-readLines("Trump.txt")
# to lower
Trump<-tolower(Trump)
#process
Trump.words <- strsplit(Trump, split = "[ |. |, |  ]")
Trump_count<-table(Trump.words)
Trump_count<-sort(Trump_count,decreasing = TRUE)
head(Trump_count,10)
```

```
Trump.words
           and     the      of     our
   188      76      71      48      47
    we    will      to      is america
    47      43      37      20      19
```

Hide

```
tail(Trump_count)
```

```
Trump.words
   wisdom wonderful     words      work
        1         1         1         1
      yes     young
        1         1
```

Hide

```
############ comparision########
Others<-readLines("Others.txt")
Others<- paste(Others, collapse = " ")
Others<-tolower(Others)
Others.words<-strsplit(Others,split = "[ |. |, |  ]")
Others_count<-table(Others.words)
Others_count<-sort(Others_count,decreasing = TRUE)
head(Others_count,10)
```

```
Others.words
        the    of   and    to    in
10883  9733  6850  5129  4351  2700
    a   our  that    we
 2167  2097  1722  1662
```

Hide

```
tail(Others_count)
```

```
Others.words
you--ask  younger youngest yourself
       1        1        1        1
youthful     zone
       1        1
```

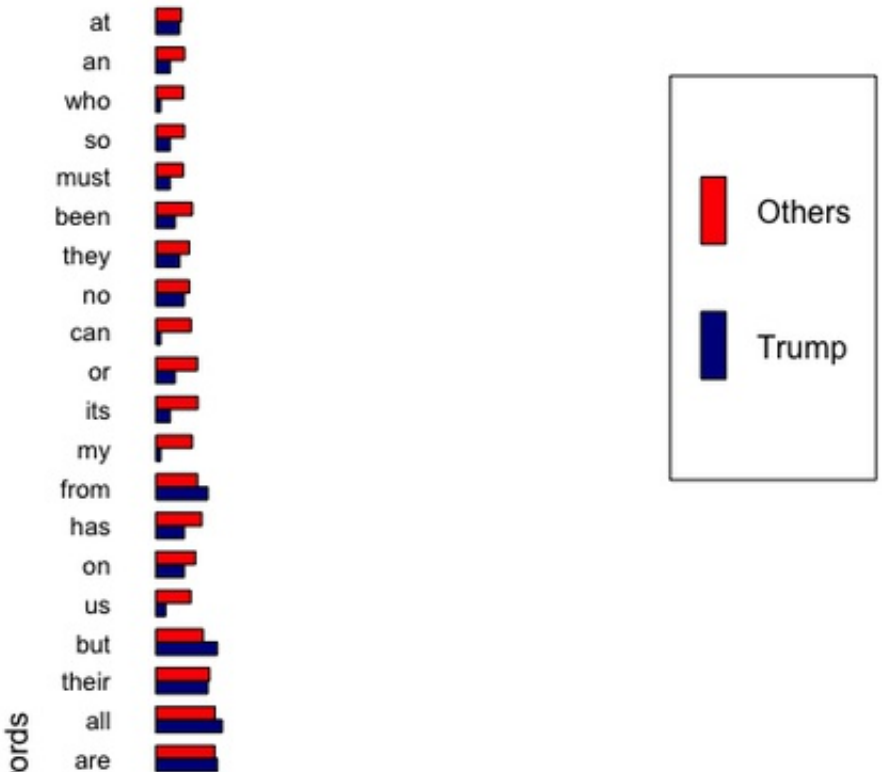
Here, the top words already differ from Trump to others. The number here is the absolute time the word mentioned. But the usage of words, and their rank do show something. Here, he mentioned WILL, AMERICA, ALL, YOU, and the words he did not frequently mentioned are THAT, BE, IT, BY.

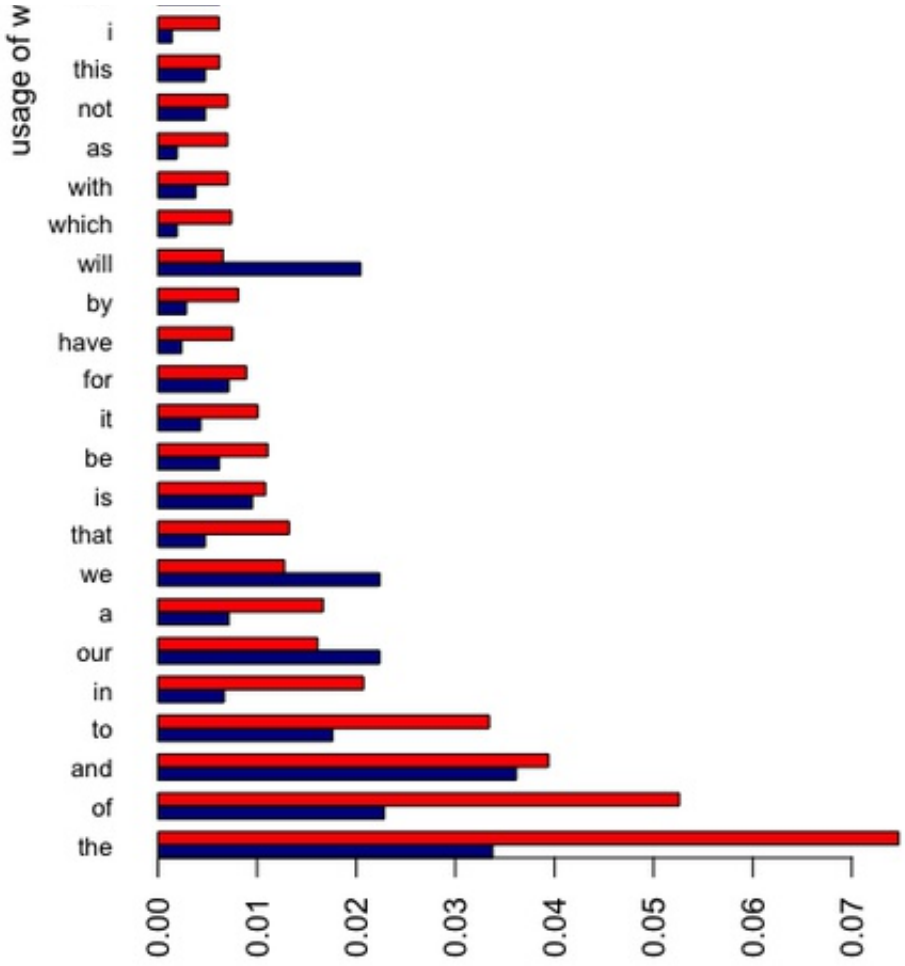
The interesting thing I found is that, Trump is using less relative clause since there is less THAT. One of the reason might be the people who vote for him is different from the one used to be. Also the segmentation of the audience is different. See reference: Reality Check: Who Voted for Donald Trump? (http://www.bbc.com/news/election-us-2016-37922587). Also, Trump is not using as much BE as others. One of the reason might be he is using WILL instead, another way to express future and deliver it to the Americans.

It is similiar to the discussion in class, that media is changing the way people speak. And evidence is that, people are using shorter sentences.

Hide

```
## now, comparing only the stopwords & the GRAPHICS
# stopword<-readline('stopwords_300.txt') After trying the stopwords, I only list out the top 42 stopwords here.
stopwords<-c("the","of","and","to","in","our","a","we","that","is","be","it","for","have","by","will","which","with","as","not","this","i","are","all","their","but","us","on","has","from","my","its","or","can","no","they","been","must","so","who","an","at")
step2stop<-matrix(NA,ncol=2,nrow=length(stopwords))
j=1
for( i in stopwords){
  step2stop[j,1]=Trump_count[i]/2103
  step2stop[j,2]=Others_count[i]/130231
  j=j+1
}
colnames(step2stop) <- c("Trump", "Others")
rownames(step2stop) <- stopwords
barplot(t(step2stop), beside=T,horiz = T, ylab="usage of words", cex.names=0.8, las=2,col=c("darkblue","red"),legend.text = c("Trump", "Others") )
# then plot the as bar chart, it is quite distinct between Trump and all other president in using stopwords
barplot(t(step2stop), beside=T,horiz = T, ylab="usage of words", cex.names=0.8, las=2,col=c("darkblue","red"),legend.text = c("Trump", "Others") )
```





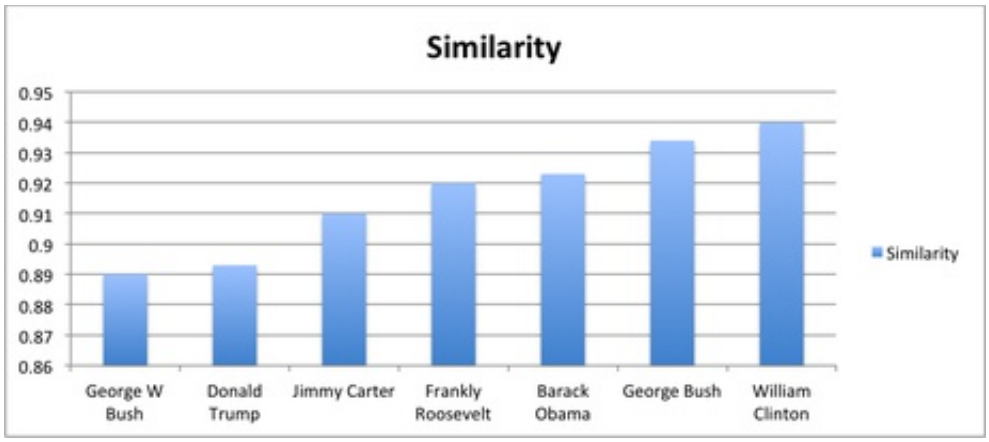
image

Finding is that, first thing everyone would notice is that Trump use extremely less THE/OF/TO/A/THAT/IT. And extremely more WILL/WE. So, street language, more future tense, more close relationship are the thing define Trump.

Here, from the perspective of cosine similarity and comparison to centroid. The following chart showed the top 5 distinction president.

Hide

```
# style comparision to other presidents, here we use the cosine similarity to compar
major difference and similarity of presidents
library(proxy)
dist(m, method="cosine")
# the finding is not that interesting, so I just attached the plot here
```



image

Some finding:

- Very interesting in stop words Trump is using. From the perspective of data science, it is clear to show that Trump's speech turned out to be the second farthest from the centroid, right behind George W. Bush's address of 2001.
- Trump was strong in using OUR and WE – and almost never said "I". I think this is a pattern shown about exaggeration, and this is similar to Obama's 2013 speech
- Trump is the all-time winner in using WILL (and interestingly, a large portion of his speech is in the future tense). Also, if looking at the actural sentences, Trump use WE'VE, the oral language. And no people did this before.
- For the part of comparision between centriod to president, Trump is in top5. Although not Top1, it is different enough.
- Further analysis and underlying business usage would be, president speech is a way to reflect common people's life.

Step 3: analysis of the speech content (words)

Here, the analysis of the speech mostly work with words.

time trend

In the tutorial, teacher analysis the wording and find the time trend behind it. For example, the wording in 1920s maybe more harsh, about war. And nowadays, focused more on economy. And to me, I think this part is very important, since it is most relevant part to people's daily life.

Here, I dig deepper into the area of speech's correlation with economy. Do the data mining work, and do the feature selection with all the words we have.

Hide

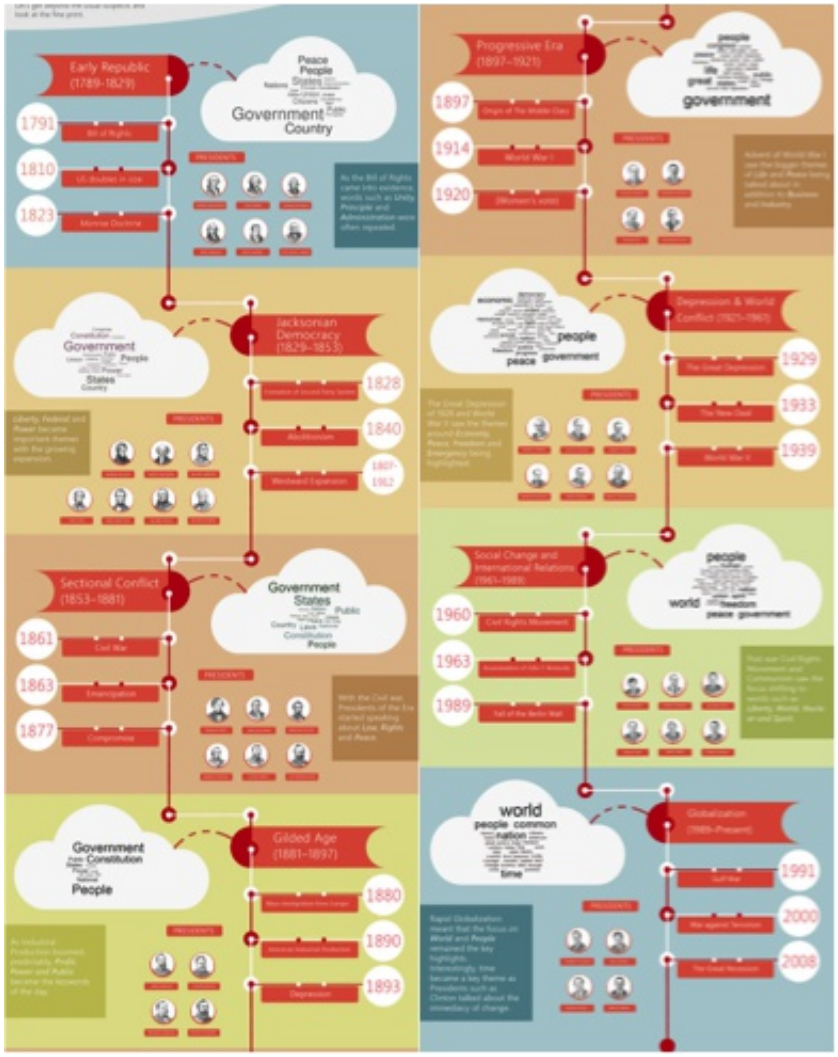
```
#in the data processing part, I use let_join to get the GDP/YoY/Financial Market/CPI
data into one single sheet
# since the data of GDP only from 1929 great depression, so only analyze part of the
data after 1930s
step2<-read.csv("step2.csv")
str(step2)
```

```
'data.frame':   18 obs. of  18 variables:
 $ President         : Factor w/ 12 levels "Barack Obama",..: 2 1 1 5 5 12 12 4 11 11
...
 $ time              : Factor w/ 18 levels "1/20/01","1/20/05",..: 4 16 3 2 1 15 14 1
3 18 12 ...
 $ Term              : int  1 2 1 2 1 2 1 1 2 1 ...
 $ Party             : Factor w/ 2 levels "Democratic","Republican": 2 1 1 2 2 1 1 2
2 2 ...
 $ Words             : int  1455 2096 2395 2071 1592 2155 1598 2320 2561 2427 ...
 $ content           : Factor w/ 18 levels " [Delivered in person at the Capitol] Mr.
Vice President, Mr. Chief Justice, fellow citizens: I accept with humility the honor
"| __truncated__,..: 2 18 8 16 15 7 9 6 14 13 ...
 $ year              : int  2017 2013 2009 2005 2001 1997 1993 1989 1985 1981 ...
 $ GDP               : Factor w/ 18 levels "1019.9","10621.8",..: 7 6 4 3 2 18 16 15
12 10 ...
 $ YoY               : num  0.13 0.158 0.101 0.233 0.234 ...
 $ Financial.market  : int  2280 1606 1057 1191 1224 885 451 317 185 131 ...
 $ CPI               : num  239 233 215 195 177 ...
 $ total.sentences   : int  101 91 119 100 96 110 92 154 125 133 ...
 $ total.words       : int  1456 2107 2383 2073 1591 2157 1600 2313 2559 2425 ...
 $ ave.polarity      : num  0.2031 0.1247 0.0613 0.1602 0.1709 ...
 $ sd.polarity       : num  0.374 0.328 0.35 0.381 0.388 ...
 $ stan.mean.polarity: num  0.544 0.38 0.175 0.421 0.44 ...
 $ party             : Factor w/ 2 levels "Democratic","Republican": 2 1 1 2 2 1 1 2
2 2 ...
 $ president         : Factor w/ 18 levels "Barack Obama 2009",..: 3 2 1 8 7 18 17 6
16 15 ...
```

Hide

```
# lots of the data in the sheet would turned into factor variable
pairs(step2)
```

Interesting methods are the one discussed in class, using categorical and sentimental analysis to get the sense of how president care about economy. And also, the word cloud also show the relationship of speech to time. The following graphics shows the time trend of word cloud.



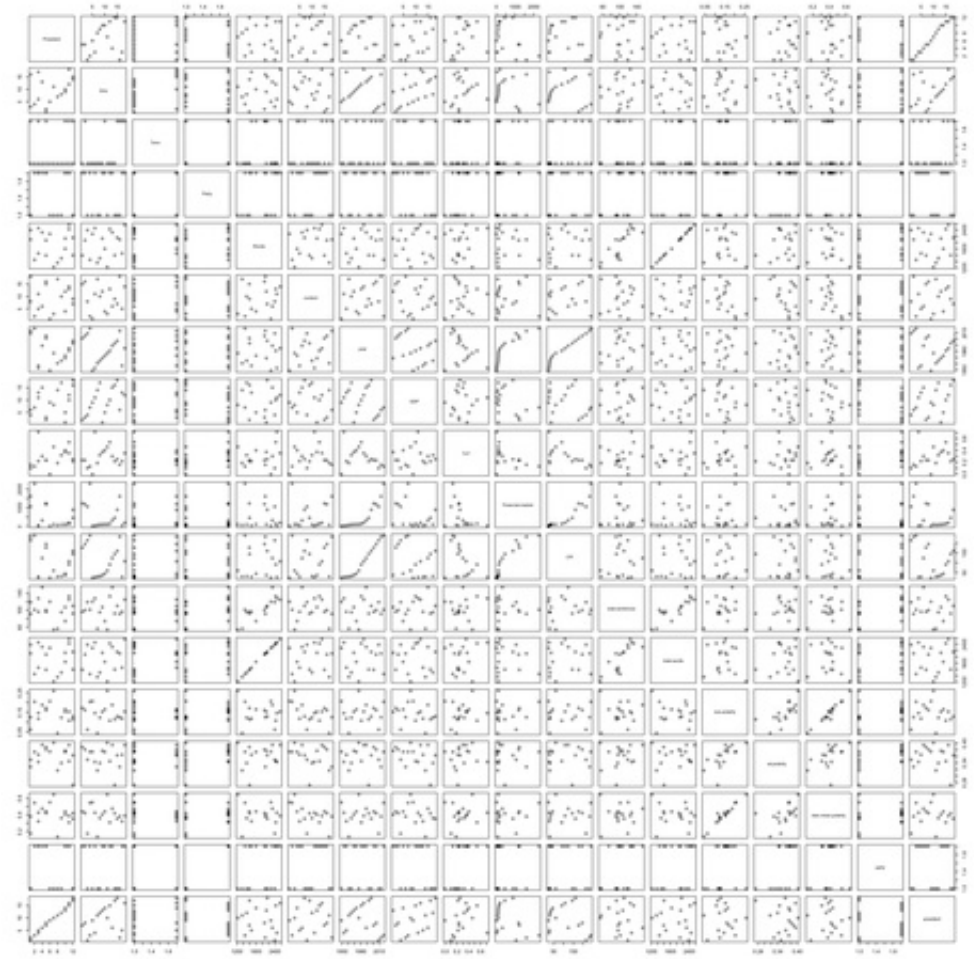
image

Hide

```
#using the preprocess data with financial data(GDP, YoY, Financial market data, dolla
r price)
#due to data constrain, the analysis only include 1940s till now
# run a linear model of these data
pairs(step2)
```

ıre margins too large

Too see clearly, use the image below.



image

But my model, I would use simple linear regression model on the data. I have a concern on this, since there is limited data, some regression issue might be hard to resolve. So, after running some linear model, I find one of the most ineresting topic with those data. That is wether Trump would have a second term! Then I change the data into training and test(only Trump data), then

If possible, maybe tree and SVM can be introduced.

Hide

```
library(leaps)
train<-read.csv("step3.csv")
train<-train[,3:15]
train$GDP<-as.numeric(train$GDP)
regfit.full=regsubsets(train$Term~.,data=train,nvmax=8)
```

```
2  linear dependencies found
```

Hide

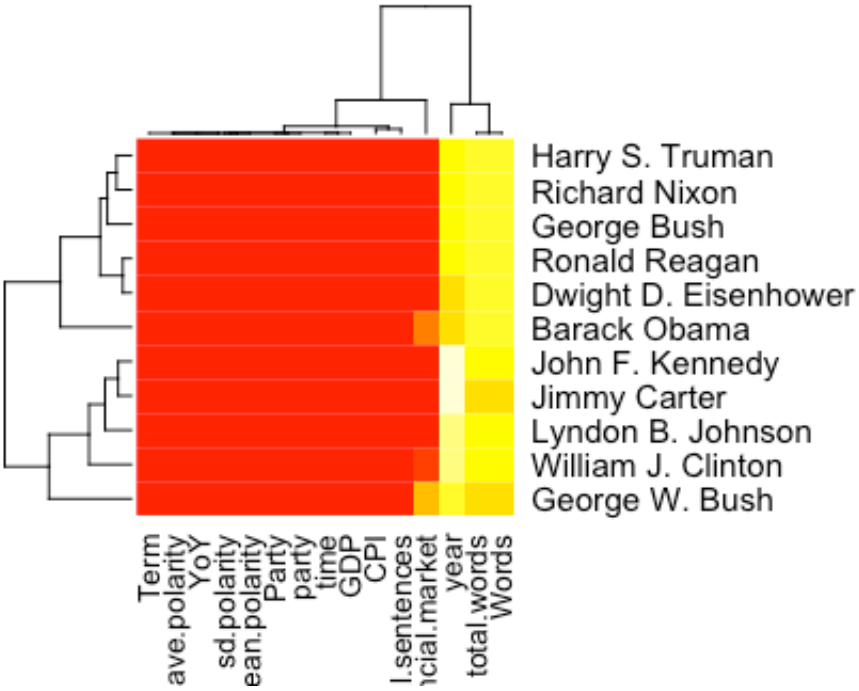
```
summary(regfit.full)
```

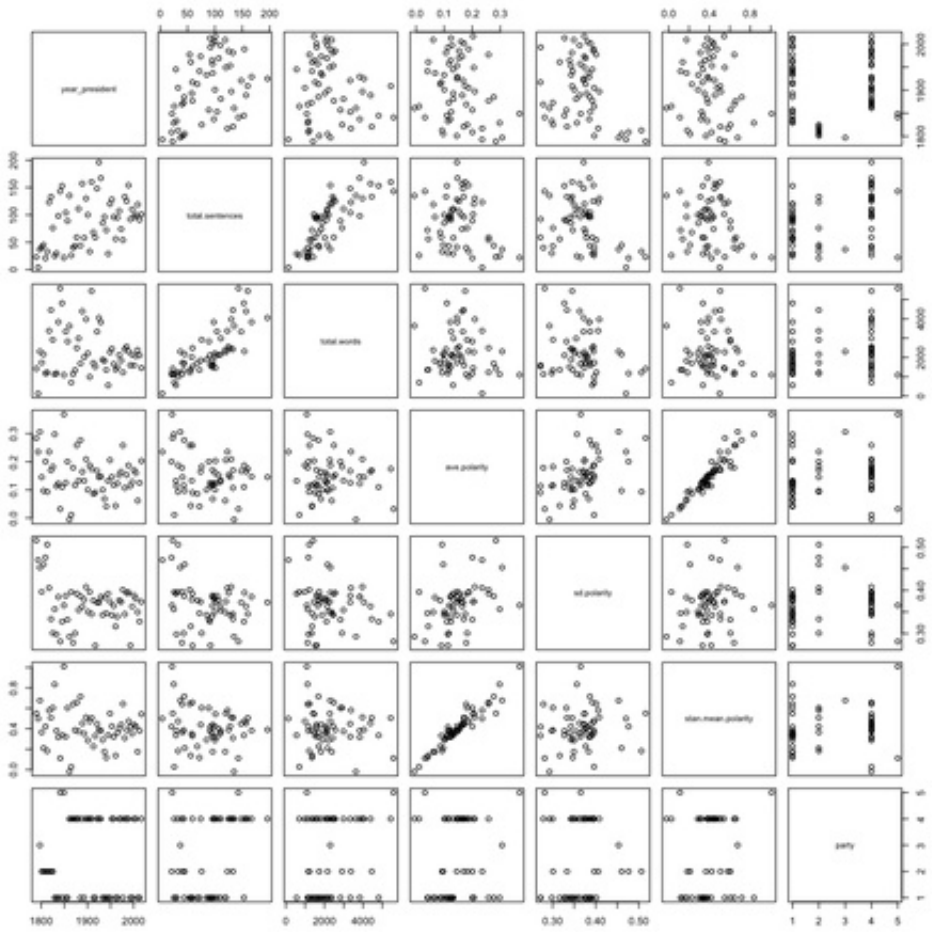
```
Subset selection object
Call: regsubsets.formula(train$Term ~ ., data = train, nvmax = 8)
12 Variables  (and intercept)
                  Forced in Forced out
PartyRepublican        FALSE      FALSE
Words                  FALSE      FALSE
year                   FALSE      FALSE
GDP                    FALSE      FALSE
YoY                    FALSE      FALSE
Financial.market       FALSE      FALSE
CPI                    FALSE      FALSE
total.sentences        FALSE      FALSE
total.words            FALSE      FALSE
ave.polarity           FALSE      FALSE
sd.polarity            FALSE      FALSE
stan.mean.polarity     FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: exhaustive
         PartyRepublican Words year GDP YoY Financial.market CPI
1  ( 1 ) " "             " "   " "  "*" " " " "              " "
2  ( 1 ) " "             " "   " "  "*" " " " "              " "
3  ( 1 ) " "             "*"   " "  " " " " "*"              "*"
4  ( 1 ) " "             "*"   " "  " " "*" "*"              "*"
5  ( 1 ) " "             "*"   " "  " " "*" "*"              "*"
6  ( 1 ) " "             "*"   "*"  "*" " " " "              " "
7  ( 1 ) " "             "*"   "*"  "*" " " " "              "*"
8  ( 1 ) " "             "*"   "*"  "*" " " "*"              " "
         total.sentences total.words ave.polarity sd.polarity
1  ( 1 ) " "             " "         " "          " "
2  ( 1 ) " "             " "         "*"          " "
3  ( 1 ) " "             " "         " "          " "
4  ( 1 ) " "             " "         " "          " "
5  ( 1 ) "*"             " "         " "          " "
6  ( 1 ) "*"             "*"         " "          "*"
7  ( 1 ) "*"             "*"         " "          "*"
8  ( 1 ) "*"             "*"         "*"          "*"
         stan.mean.polarity
1  ( 1 ) " "
2  ( 1 ) " "
3  ( 1 ) " "
4  ( 1 ) " "
5  ( 1 ) " "
6  ( 1 ) " "
7  ( 1 ) " "
8  ( 1 ) " "
```

And the economy data is relevant to this. ALso, the heat map is quite vague, but mean something.

Hide

```
train<-read.csv("step3.csv")
row.names(train) <- train$President
tt<-data.matrix(train[,2:16])
heatmap(tt)
```





image

This graph somehow restreghten the conclusion in part 1.

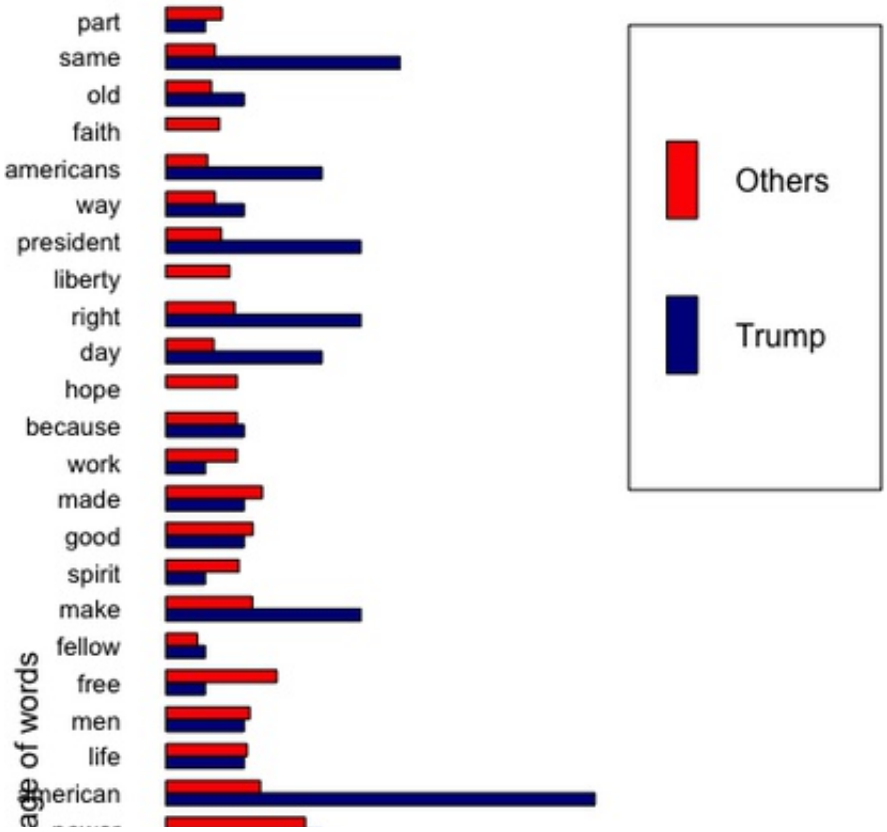
not about time

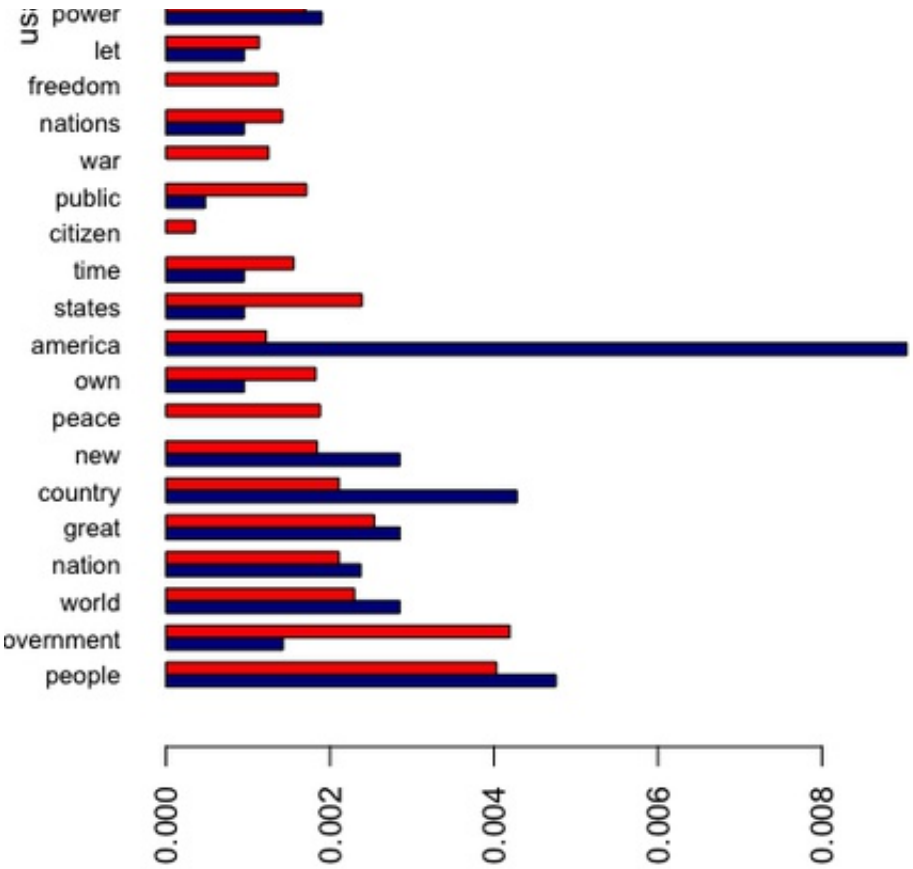
I know, there must be something across time, which is classic american, not changing with time trend.

From the tutorial, the length of sentence, the emotion of paragrph and the clustering of working are all about content. Here, I analyzed the top related non-stop word.

Hide

```
#using the same one from step1(using the word counting method for the non- stoping wo
rds)
content<-c("people","government","world","nation","great","country","new","peace","ow
n","america","states","time","citizen","public","war","nations","freedom","let","powe
r","american","life","men","free","fellow","make","spirit","good","made","work","beca
use","hope","day","right","liberty","president","way","americans","faith","old","same
","part")
contentcontent<-matrix(NA,ncol=2,nrow=length(content))
j=1
for( i in content){
  contentcontent[j,1]=Trump_count[i]/2103
  contentcontent[j,2]=Others_count[i]/130231
  j=j+1
}
colnames(contentcontent) <- c("Trump", "Others")
rownames(contentcontent) <- content
barplot(t(contentcontent), beside=T,horiz = T, ylab="usage of words", cex.names=0.8,
las=2,col=c("darkblue","red"),legend.text = c("Trump", "Others") )
# since the centroid did not show much, so it is not here.
```





image

The graph shows a lot. “Make America Great Again.” In this slot, words MAKE/AMERICA used quite often. Trump is Top1 among all the presidents in saying AMERICA and AMERICAN. On the other hand, he never used the words PEACE, FREEDOM, CONSITITUTION, LIBERTY, DUTY, etc which had been so popular with his predecessors.

And the most interesting words are FACTORIES, BORDERS, and DREAM - right go abreast with Trump’s agenda.“Build a wall…. Build factories to stimulate econoy” see link for more infomation. Trump to Order Mexican Border Wall and Curtail Immigration (https://www.nytimes.com/2017/01/24/us/politics/wall-border-trump.html?_r=0)

Some findings:

- GDP/financial market related to the time trend of the speeches. As time goes by, sentences getting shorter and words in sentences same trend.
- Trump using American more than others. Thus showing exaggeration on the unity of American. Also using same and we, which is a evidence of similarity and unity in USA.
- Lastly, Trump has a clear theme compared to others. Thus lots of words a used in emphsize his goal as a president.

Step 4: summary

The most interesting result I have see is that, Trump do differ from others. And here is the summary of the findings:

- The GDP YoY can help us to devide all the speeches into 4 category. Also, GDP/financial market related to the time trend of the speeches. As time goes by, sentences getting shorter and words in sentences same trend, since they are top in stepwise feature selection.
- For different party, the difference is quite obvious. Especially the sentimental ones, since polarity() shows the sentiment (polarity) of text by grouping variable(s). Actucally Maybe clustering would make sense here. Also, this is a proof to one of the articles I read. And I think this is part of the common point I am looking for, that is althongh time change, president stayed same. Further analysis see part 2.
- Trump uses the shortest sentences of all presidents, and his speech has a higher numerical sentiment compared to both Obama speeches. His speech is very short(but not the shortest one).
- Time trend exists. Adams speech in 1797 used longest sentneces. As the discussion in class, media trend and how people comunicate evolve. I also read some paper about this, one of the reason is that new words are coming up, some words mean more than one concept. See the link for more information. It also analyzed the complexity of the sentence, the structure and the visulization of sentences., showing same pattern.
- Very interesting in stop words Trump is using. From the perspective of data science, it is clear to show that Trump's speech turned out to be the second farthest from the centroid, right behind George W. Bush's address of 2001. For the part of comparision between centriod to president, Trump is in top5. Although not Top1, it is different enough. Trump was strong in using OUR and WE – and almost never said "I". I think this is a pattern shown about exaggeration, and this is similar to Obama's 2013 speech
- Trump is the all-time winner in using WILL (and interestingly, a large portion of his speech is in the future tense). Also, if looking at the actural sentences, Trump use WE'VE, the oral language. And no people did this before. Trump using American more than others. Thus showing exaggeration on the unity of American. Also using same and we, which is a evidence of similarity and unity in USA. Trump has a clear theme compared to others. Thus lots of words a used in emphsize his goal as a president.

Here is potential business interests:

- Further analysis and underlying business usage would be, president speech is a way to reflect common people's life.
- Inuitively grapgic to further analysis on interesting topics for business purpose, and thus further test is needed here. Such as sentences getting shorter, Twitter can switch to shorter text. Instagram getting popular, since people are more interested in pictures and shorter sentences, so there is no need for long sentences.

After analysing the word data and the whole speech with text mining techniques, I think more work or comparsion might be the one between candidata and between candidata/presidental one. Since, it is more currently related. And it might be helpful for us to find out whether Trump is trust worth or not. For example, Trump has not change wording in candidata domination and presidental speech, then might some regulation would come soon, which is more relative to the people.

reference

http://smartdatawithr.com/en/ (http://smartdatawithr.com/en/)
http://avalon.law.yale.edu/subject_menus/inaug.asp (http://avalon.law.yale.edu/subject_menus/inaug.asp)