Project 1 - Leo Lam

Code

Inauguration speech is a ceremony marking the commencement of a new term of President of America. Often, in this speech, President addresses his policies and future presidential plans. In this study, our goal is to investigate the unfound pattern in these speeches. We focus on the differences in between Republican Presidents and Democratic Presidents as well as Presidents who received graduate school education and Presidents who did not. We want to see if different educational and political backgrounds will have impacts on the topic, word length, sentence length, and emotion of their inauguration speeches.

Before we start analyzing the President inauguration speeches, we have to first install all packages we are going to use in this report.

Hide

Hide

```
packages.used=c("rvest", "tibble", "qdap",
                "sentimentr", "gplots", "dplyr",
                "tm", "syuzhet", "factoextra",
                "beeswarm", "scales", "RColorBrewer",
                "RANN", "tm", "topicmodels",'rJava')
# 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)
}
# load packages
library(ngram)
library(ggplot2)
library("rvest")
library("tibble")
library(qdap)
library(rJava)
library("sentimentr")
library("gplots")
library("dplyr")
library("tm")
library("syuzhet")
library("factoextra")
library("beeswarm")
library("scales")
library("RColorBrewer")
library("RANN")
library("topicmodels")
library(wordcloud)
library(tidytext)
source("~/Desktop/Columbia/Semester 2/ADS/Project1/wk2-TextMining/lib/plotstacked.R")
source("~/Desktop/Columbia/Semester 2/ADS/Project1/wk2-TextMining/lib/speechFuncs.R")
```

Step 1 - Importing speeches

Then, we are going to import our inauguration speeches from http://www.presidency.ucsb.edu/inaugurals.php (http://www.presidency.ucsb.edu/inaugurals.php). We also format the date and remove irrevalent information, which is the last row of our inaug file.

Now, we import a csv file that contain information of the president, including president's name, term, party, date, and number of words containing in the inauguration speech.

Hide

Hide

```
inaug.list=read.csv("~/Desktop/Columbia/Semester 2/ADS/Project1/wk2-TextMining/data/inauglist.csv", stringsAsFactors = FALSE)
```

For the ease of our further investigation, we combine file inaug and inaug.list together into one single file.

Hide

Hide

```
inaug.list$type=c(rep("inaug", nrow(inaug.list)))
speech.list=cbind(inaug.list, inaug)
```

And then, we add the inauguration speech of each president into a new column called fulltext.

Hide

Hide

```
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$fulltext[i]=text
  # Create the file name
  filename <- paste0("~/Desktop/Columbia/Semester 2/ADS/Project1/wk2-TextMining/data/fulltext/",
                     speech.list$type[i],
                     speech.list$File[i], "-",
                     speech.list$Term[i], ".txt")
  sink(file = filename) %>% # open file to write
  cat(text)  # write the file
  sink() # close the file
}
```

Step 2 - Data Preprocessing

Now, we begin to analyze the speech of each president. We first separate each sentence of an inauguration speech and obtain the sentiment of each sentence using get_nrc_sentiment() command, which will return sentiment of anger, anticipation, disgust, fear, joy, sadness, surprise, trust, negative, and positive. We consider “?”, “.”, “!”, “|”, and “;” as the end of each sentence. After we finish the above steps, we remove all the rows with NA value.

Hide

Hide

```
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)
    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))
```

Step 3 - Data Analytics

Let the fun part begins! In this study, we will analyze the differences in between the inauguration speeches of Democratic Presidents and Republican Presidents as well as the presidents who attended graudate school and the presidents who did not attend graudate school. We try to answer the question of what will be the differences in emotion, word length, sentence length, and topics.

First, let us select the Democratic and Republician Presidents.

Hide

Hide

```
democratic = speech.list$File[which(speech.list$Party=='Democratic')]
democratic.speech.ind = which(speech.list$Party=='Democratic')
democratic.sentence.ind = which(sentence.list$Party=='Democratic')
republican = speech.list$File[which(speech.list$Party=='Republican')]
republican.speech.ind = which(speech.list$Party=='Republican')
republican.sentence.ind = which(sentence.list$Party=='Republican')
```

Step 3.1 - Number of words in each sentence

To obtain the number of words in each sentence, we write a function called sentence.length, which will return a vector that contains the length of each sentence of a speech.

Hide

Hide

```
sentence.length <- function(index, df){
  speech.vec = df$fulltext
  a = strsplit(speech.vec[index], '[?!;.|]')
  res.vec = c()
  for(j in 1:length(a[[1]])){
    res.vec = cbind(res.vec, wordcount(a[[1]][[j]]))
  }
  return(res.vec)
}
```

Then, we apply the sentence.length function into both Democratic and Republican inauguration speeches. Since the first democratic president is Andrew Jackson as his predecessors are all democratic-republican party or Fedralist, which both parties do not exist anymore, we will only consider the presidents after President Jackson. We do not consider Whig in this study as it also doesn't exist today.

Hide

Hide

```
mean.length.democratic = 0
for(i in democratic.speech.ind){
  mean.length.democratic = mean.length.democratic + mean(sentence.length(i, speech.li
st))
}
mean.length.democratic = mean.length.democratic/length(democratic.speech.ind)
mean.length.republican = 0
for(i in republican.speech.ind){
  mean.length.republican = mean.length.republican + mean(sentence.length(i, speech.li
st))
}
mean.length.republican = mean.length.republican/length(republican.speech.ind)
cat('Democratic average number of words in each sentence',(mean.length.democratic),'\
n')
```

```
Democratic average number of words in each sentence 23.79952
```

Hide

Hide

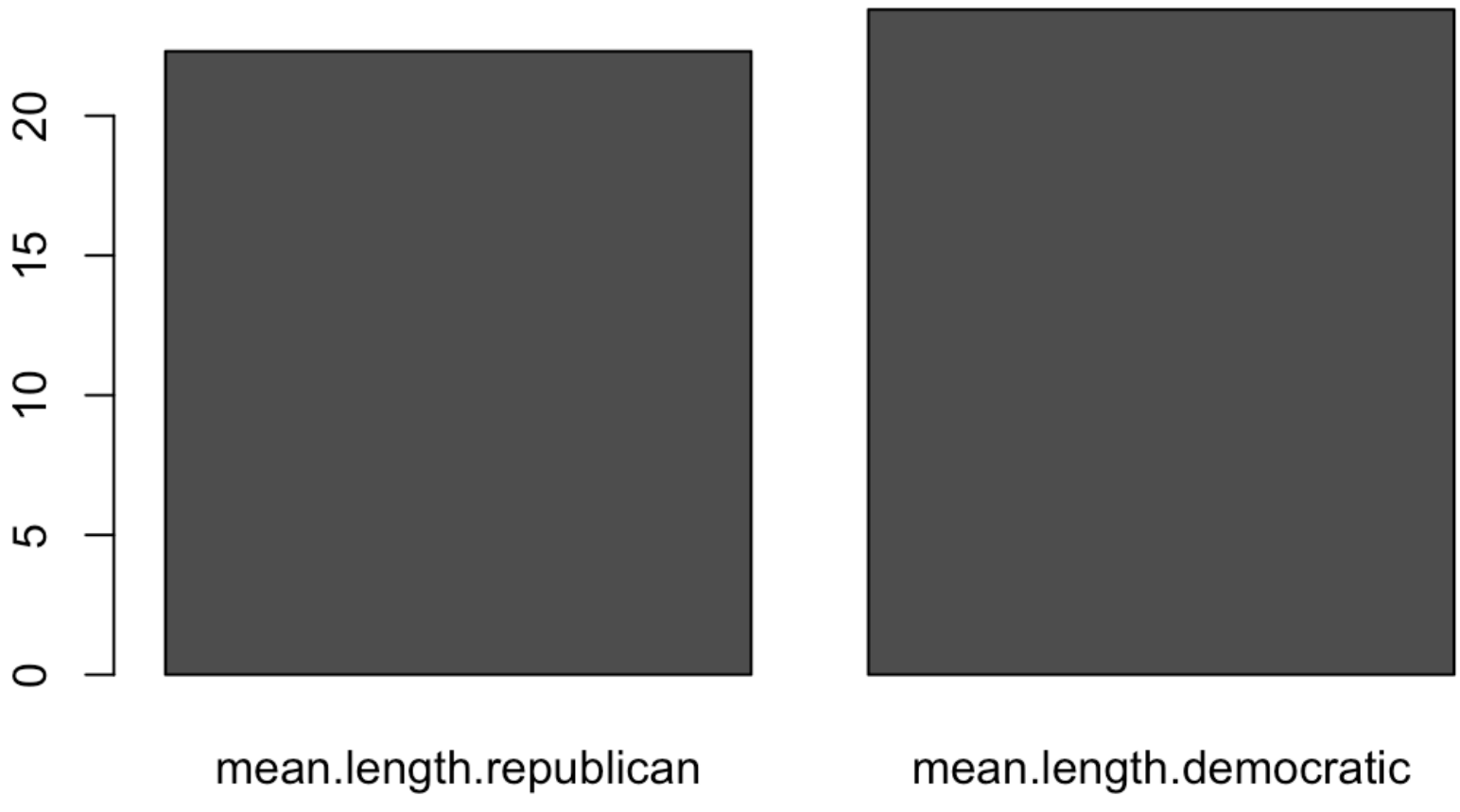
```
cat('Republican average number of words in each sentence',(mean.length.republican))
```

```
Republican average number of words in each sentence 22.30762
```

Hide

Hide

```
df.length.repvsdem = cbind(mean.length.republican,mean.length.democratic)
barplot(df.length.repvsdem)
```



While republican average number of words is lower than democratic(democratic = 23.79952 vs Republican = 22.30762), there isn't a major difference in number of words in each sentence between presidents of both parties. Hence, we begin to suspect would going to graduate school has a positive impact of the average number of words in each sentence. Our hypothesis is going to graduate school may increase the probability of the use of longer sentences.

Using the information from Wikipedia, we obtain the list of presidents in both parties who did not attend graduate school. Again, we only consider presidents after Andrew Jackson. (List of Presidents: https://en.wikipedia.org/wiki/List_of_Presidents_of_the_United_States (https://en.wikipedia.org/wiki/List_of_Presidents_of_the_United_States)) (Education of President: https://en.wikipedia.org/wiki/List_of_Presidents_of_the_United_States_by_education (https://en.wikipedia.org/wiki/List_of_Presidents_of_the_United_States_by_education))

Hide

Hide

```
# republican president who did not attend graduate school
no.grad.rep = c('AbrahamLincoln','UlyssesSGrant','JamesGarfield','BenjaminHarrison',
'WilliamMcKinley','TheodoreRoosevelt','WarrenGHarding','CalvinCoolidge', 'HerbertHoov
er','RonaldReagan','GeorgeBush','DonaldJTrump')
no.grad.rep.speech.ind = c()
for(i in no.grad.rep){
  no.grad.rep.speech.ind = c(no.grad.rep.speech.ind, which(i == speech.list$File))
}
no.grad.rep.sentence.ind = c()
for(i in no.grad.rep){
  no.grad.rep.sentence.ind = c(no.grad.rep.sentence.ind, which(i == sentence.list$Fil
e))
}
# democratic president who did not attend gradaute school
no.grad.dem = c('AndrewJackson','MartinvanBuren','JamesKPolk','JamesBuchanan',
                'GroverCleveland-I','GroverCleveland-II','HarrySTruman',
                'JohnFKennedy')
no.grad.dem.speech.ind = c()
for(i in no.grad.dem){
  no.grad.dem.speech.ind = c(no.grad.dem.speech.ind, which(i == speech.list$File))
}
no.grad.dem.sentence.ind = c()
for(i in no.grad.dem){
  no.grad.dem.sentence.ind = c(no.grad.dem.sentence.ind, which(i == sentence.list$Fil
e))
}
cat('Out of total', length(republican),'republican presidents,','number of republican
presidents who did not attend graduate school:', length(no.grad.rep),'\n')
```

```
Out of total 24 republican presidents, number of republican presidents who did not at
tend graduate school: 12
```

Hide

Hide

```
cat('Out of total', length(democratic),'democratic presidents,','number of democratic
presidents who did not attend graduate school:', length(no.grad.dem),'\n')
```

```
Out of total 22 democratic presidents, number of democratic presidents who did not at
tend graduate school: 8
```

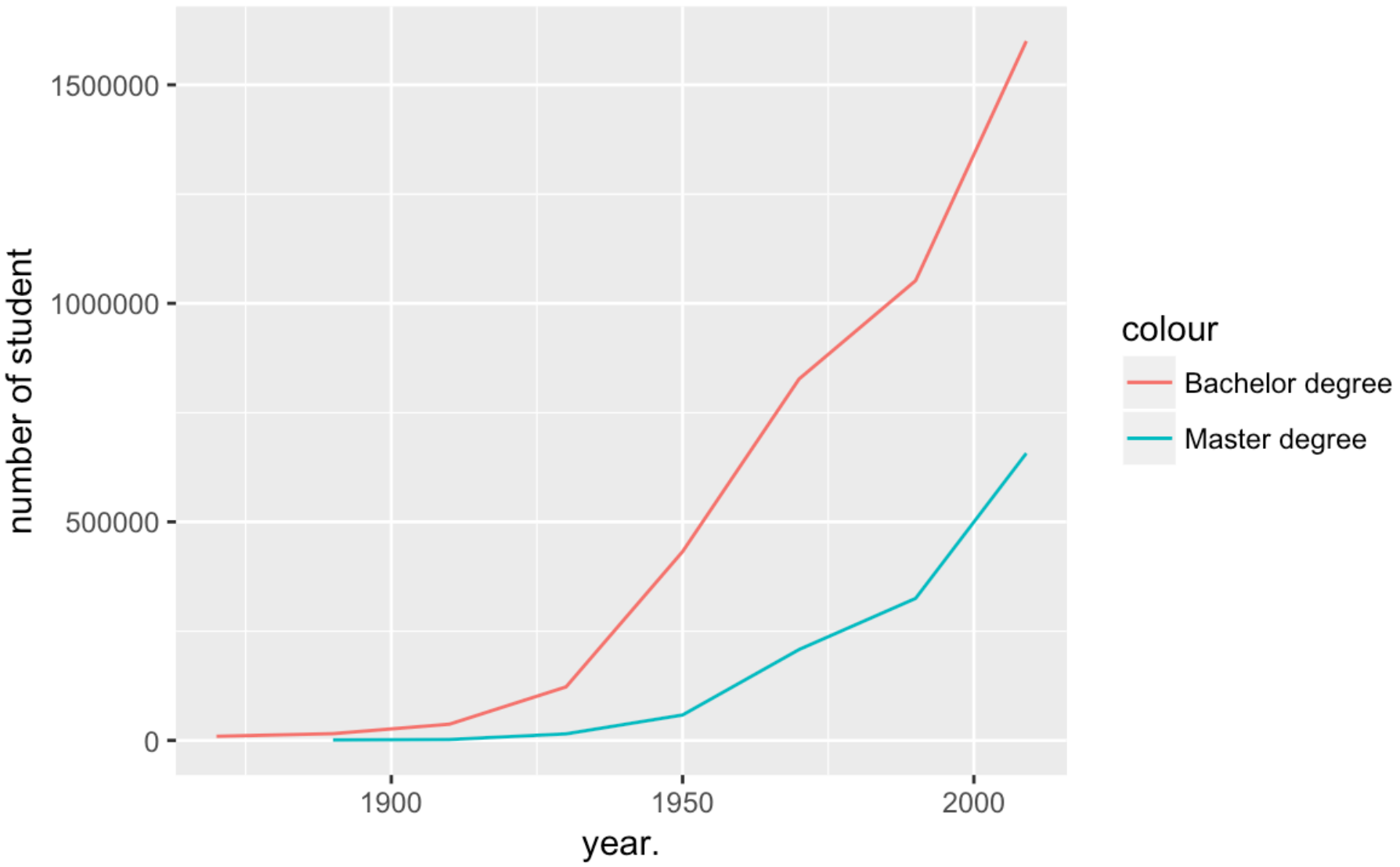
From above, we notice there are more republican presidents who did not attend graduate school. This may explain why republican has an overall lower average number of words in each sentence, which may support our hypothesis of there exists a positive correlation between going to graduate school and the use of longer sentence.

To get a more accurate explanation, let's look at the number of students who attend graduate school in different period of time. We found a dataset from Wikipedia, which provides the number of students who attended undergraduate and graduate school since 1890.
(https://en.wikipedia.org/wiki/History_of_higher_education_in_the_United_States
(https://en.wikipedia.org/wiki/History_of_higher_education_in_the_United_States))

Hide

Hide

```
year. <- c(1870,1890,1910,1930,1950,1970,1990, 2009)
BA. <- c(9400,15500,37200,122500,432000,827000,1052000,1600000)
MA. <- c(NA,1000,2100,15000,58200,208000,325000,657000)
df.degree <- data.frame(year., BA., MA.)
ggplot(df.degree)+
  geom_line(aes(x=year., y=BA., col = 'Bachelor degree'))+
  geom_line(aes(x=year., y=MA., col = 'Master degree'))+
  ylab('number of student')
```



As we can see from the graph, there were fewer than 1000 students attending graduate school before 1890. Therefore, we hypothesize that it was much harder to attend graudate school before 1890, which means the presidents who attended graudate school before 1890 might have a better educational background and ability. The hypothesis here is attending graduate school before 1890 has a stronger impact on the tendency of using longer sentences than attending graduate school after 1890.

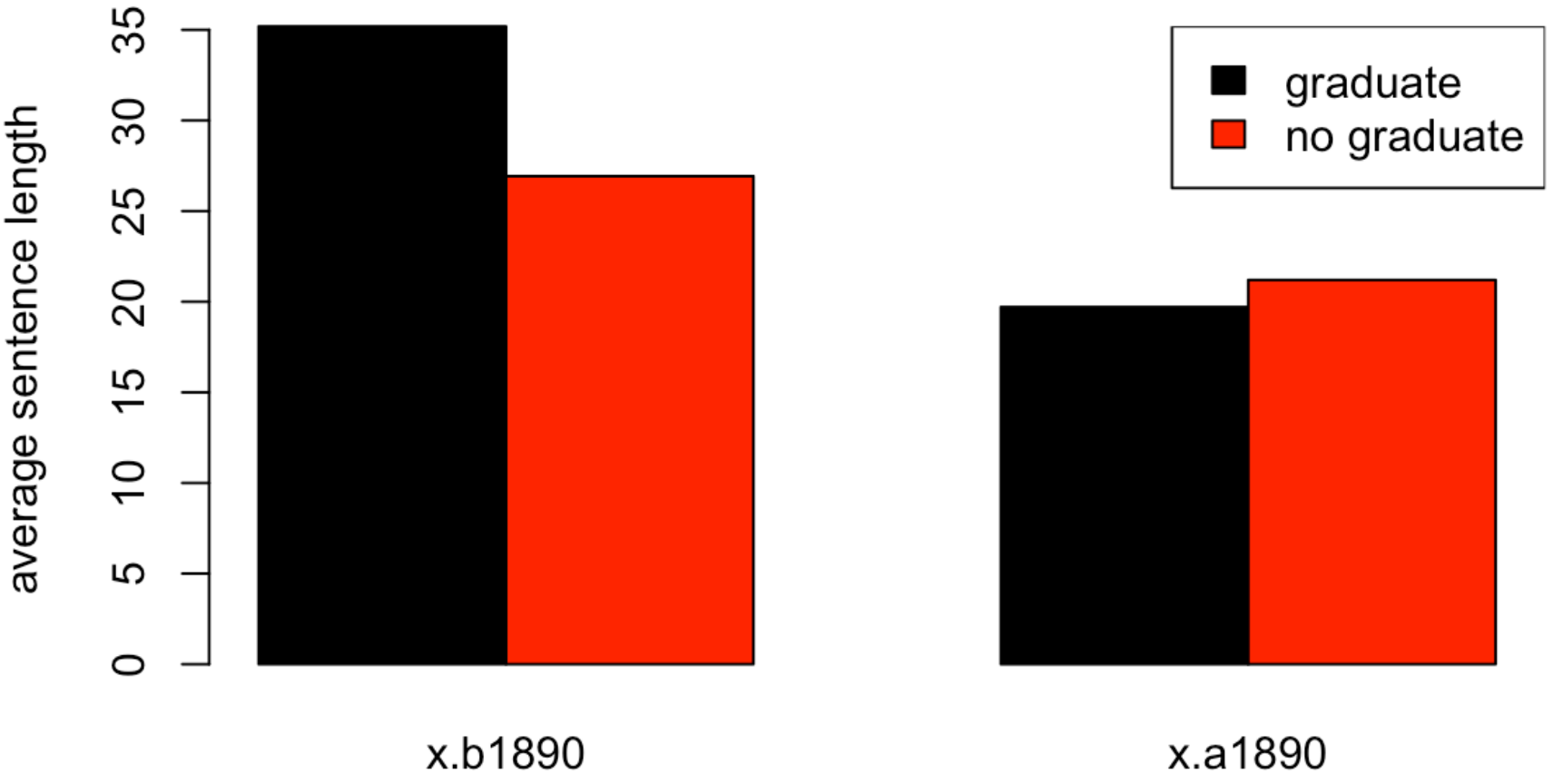
Hide

Hide

```
mean.sentence.length.no.grad.a1890 = 0
for(i in no.grad.a1890.ind){
  mean.sentence.length.no.grad.a1890 = mean.sentence.length.no.grad.a1890 + mean(sent
ence.length(i, speech.list))
}
mean.sentence.length.no.grad.a1890 = mean.sentence.length.no.grad.a1890/length(no.gra
d.a1890.ind)

mean.sentence.length.grad.a1890 = 0
for(i in grad.a1890.ind){
  mean.sentence.length.grad.a1890 = mean.sentence.length.grad.a1890 + mean(sentence.l
ength(i, speech.list))
}
mean.sentence.length.grad.a1890 = mean.sentence.length.grad.a1890/length(grad.a1890.i
nd)
mean.sentence.length.no.grad.b1890 = 0
for(i in no.grad.b1890.ind){
  mean.sentence.length.no.grad.b1890 = mean.sentence.length.no.grad.b1890 + mean(sent
ence.length(i, speech.list))
}
mean.sentence.length.no.grad.b1890 = mean.sentence.length.no.grad.b1890/length(no.gra
d.a1890.ind)

mean.sentence.length.grad.b1890 = 0
for(i in grad.b1890.ind){
  mean.sentence.length.grad.b1890 = mean.sentence.length.grad.b1890 + mean(sentence.l
ength(i, speech.list))
}
mean.sentence.length.grad.b1890 = mean.sentence.length.grad.b1890/length(grad.b1890.i
nd)
x.b1890 = c(mean.sentence.length.grad.b1890,mean.sentence.length.no.grad.b1890)
x.a1890 = c(mean.sentence.length.grad.a1890,mean.sentence.length.no.grad.a1890)
df.ab1890 = (cbind(x.b1890,x.a1890))
rownames(df.ab1890) = c('graduate school', 'no graduate school')
barplot(df.ab1890, beside=T, col=1:2, ylab='average sentence length')
legend('topright',legend=c('graduate', 'no graduate'),col=1:2, fill = 1:2)
```



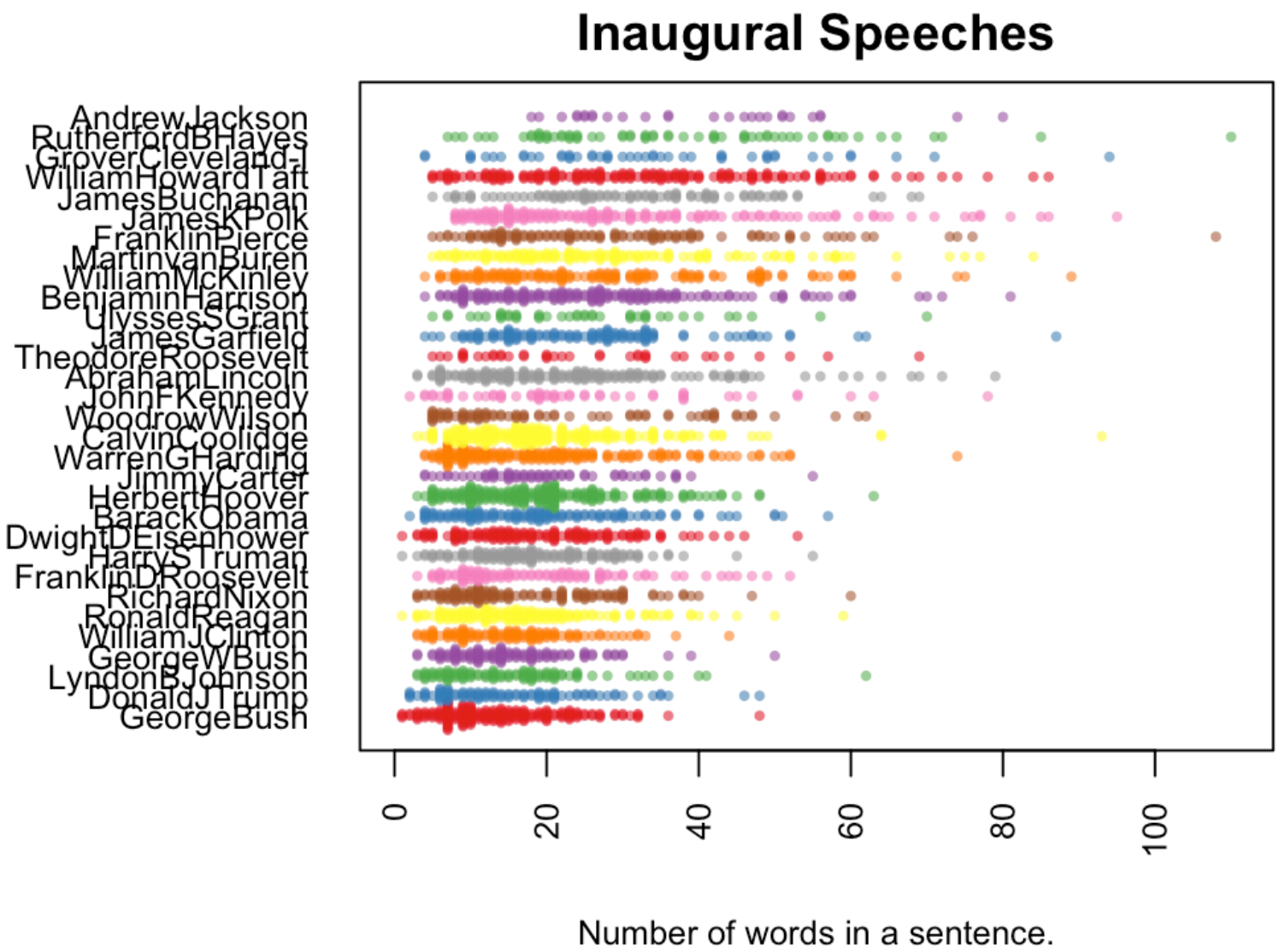
The left shows the average sentence length before 1890 while the right shows the average sentence length after 1890. The result agrees with our hypothesis of going to graduate school before 1890 may have a bigger impact on the tendency of the use of longer sentences. This result also shows going to graduate school after 1890 doesn’t have a significant impact on the average number of words in each sentence, which may indicate the difficulty of going to graduate decreased as there are significant higher amount of graduate student after 1890. This indicates the quality of graduate student might as well decrease causing less impact on the ability of using longer sentences. But we will not further investigate the quality of graduate school education as we would focus more on president inauguration speech in this study.

Now, we will take a look of the number of words in each sentence for each president using beeswarm().

Hide

Hide

```
sentence.list.sel=sentence.list%>%filter(type=="inaug", File%in%c(republican, democra
tic), Term==1)
sentence.list.sel$File=factor(sentence.list.sel$File)
sentence.list.sel$FileOrdered=reorder(sentence.list.sel$File,
                                      sentence.list.sel$word.count,
                                      mean,
                                      order=T)
par(mar=c(4, 11, 2, 2))
beeswarm(word.count~FileOrdered,
         data=sentence.list.sel,
         horizontal = TRUE,
         pch=16, col=alpha(brewer.pal(9, "Set1"), 0.6),
         cex=0.55, cex.axis=0.8, cex.lab=0.8,
         spacing=5/nlevels(sentence.list.sel$FileOrdered),
         las=2, ylab="", xlab="Number of words in a sentence.",
         main="Inaugural Speeches")
```



This graph above further agrees with our hypothesis of presidents who attended graduate school before 1890 tend to use longer sentences as most of the presidents who use longer sentence are the presidents who attend graduate school before 1890 (presidents near the top of the list). The graph aboves also shows us the number of words in each sentence of each president.

Now, let's take a look of the inauguration speech content of the president who used the shortest sentences, George Bush.

Hide

Hide

```
corpus1 = VCorpus(VectorSource(speech.list$fulltext[speech.list$File=='GeorgeBush']))
corpus1<-tm_map(corpus1, stripWhitespace)
corpus1<-tm_map(corpus1, content_transformer(tolower))
corpus1<-tm_map(corpus1, removeWords, stopwords("english"))
corpus1<-tm_map(corpus1, removeWords, character(0))
corpus1<-tm_map(corpus1, removePunctuation)
tdm.all<-TermDocumentMatrix(corpus1)
tdm.tidy=tidy(tdm.all)
tdm.overall=summarise(group_by(tdm.tidy, term), sum(count))
wordcloud(tdm.overall$term, tdm.overall$`sum(count)`,
          scale=c(5,0.5),
          max.words=100,
          min.freq=1,
          random.order=FALSE,
          rot.per=0.3,
          use.r.layout=T,
          random.color=FALSE,
          colors=brewer.pal(9,"Blues"))
```

As we can see from the wordcloud above, President Bush mentioned words like nation, freedom, friends, etc, whcih are topics related to american dream and faith. President Bush used ‘will’ a lot in his speech, which is very reasonable as presidents tend to express their future plan and policy in their inauguration speech.

Next, let’s look at Donald Trump’s wordcloud, who use the second shortest sentences.

Hide

Hide

```
corpus1 = VCorpus(VectorSource(speech.list$fulltext[speech.list$File=='DonaldJTrump']
))
corpus1<-tm_map(corpus1, stripWhitespace)
corpus1<-tm_map(corpus1, content_transformer(tolower))
corpus1<-tm_map(corpus1, removeWords, stopwords("english"))
corpus1<-tm_map(corpus1, removeWords, character(0))
corpus1<-tm_map(corpus1, removePunctuation)
tdm.all<-TermDocumentMatrix(corpus1)
tdm.tidy=tidy(tdm.all)
tdm.overall=summarise(group_by(tdm.tidy, term), sum(count))
wordcloud(tdm.overall$term, tdm.overall$`sum(count)`,
          scale=c(5,0.5),
          max.words=100,
          min.freq=1,
          random.order=FALSE,
          rot.per=0.3,
          use.r.layout=T,
          random.color=FALSE,
          colors=brewer.pal(9,"Blues"))
```

President Trump seems to focus on topic about ‘American’ and ‘Dream’. By comparing the wordcloud of both presidents, President Bush’s inauguration seems to cover more comprehensive words as President Trump seems to focus on relatively limited word choices. Same as President Bush, President Trump also used a lot of ‘will’.

Now, let’s have a closer look of the content of their speeches. We look at the shorter sentences from inauguration speeches of both George Bush and Donald Trump.

Hide

Hide

```
sentence.list%>%
  filter(File=="GeorgeBush",
         type=="inaug",
         word.count<=5)%>%
  select(sentences)%>%sample_n(10)
```

	sentences <fctr>
1	Mr.
18	God bless you.
7	a loving parent;
10	We need harmony;
4	Amen.
11	That war cleaves us still.
9	we've had dissension.
15	Good will begets good will.
5	in important things, diversity;
14	The American people await action.

1-10 of 10 rows

Hide

Hide

```
sentence.list%>%
  filter(File=="DonaldJTrump",
         type=="inaug",
         word.count<=5)%>%
  select(sentences)%>%sample_n(10)
```

	sentences

	<fctr>
3	This is your day.
4	This is your celebration.
2	Thank you.
7	We will not fail.
10	God bless America.
1	They have been magnificent.
9	Thank you.
5	But that is the past.
8	Thank you.
6	America first.

1-10 of 10 rows

The left table shows the sentences with 5 words or fewer from President Bush and the right table shows that of President Trump. Both President Bush and President Trump used a lot of thank you in his shorter sentences. President Trump mentioned his core policy, American first, a lot in his speech. As we can see from above, although both presidents used a lot of short sentences, their contents are significant different. The short sentences of President Bush are about “war”, “parent”, “compromise”, and “dissension” while the short sentences of President Trump are relatively meanless, such as ‘not fail’, ‘celebration’, ‘day’, and ‘bless’.

Now, let’s investigate the presidents who use the shortest sentences in both parties.

Hide

Hide

```
length.all <- c()
for(i in 1:58){
  length.all = c(length.all, mean(sentence.length(i, speech.list)))
}
length.dem <- c()
for(i in democratic.speech.ind){
  length.dem =  c(length.dem, mean(sentence.length(i, speech.list)))
}
length.rep = c()
for(i in republican.speech.ind){
  length.rep =  c(length.rep, mean(sentence.length(i, speech.list)))
}
#ggplot()+
#  geom_line(aes(republican.speech.ind, length.rep, color = 'republican'))+
#  geom_line(aes(democratic.speech.ind, length.dem, color = 'democratic'))
republican[which.min(length.rep)] #GeorgeBush shortest sentence, didn't go to grad sc
hool, 64 years old
```

```
[1] "GeorgeBush"
```

Hide

Hide

```
intersect(no.grad, republican[which.min(length.rep)])
```

```
[1] "GeorgeBush"
```

Hide

Hide

```
z = length.rep
z[which.min(length.rep)] = NA
republican[which.min(z)] #DonaldJTrump second shortest sentence, didn't go to grad sc
hool, 70 years old
```

```
[1] "DonaldJTrump"
```

Hide

Hide

```
intersect(no.grad, republican[which.min(z)])
```

```
[1] "DonaldJTrump"
```

Hide

Hide

```
democratic[which.min(length.dem)] #LyndonBJohnson, 64 years old
```

```
[1] "LyndonBJohnson"
```

Hide

Hide

```
intersect(no.grad, democratic[which.min(length.dem)]) #went to grad school
```

```
character(0)
```

Hide

Hide

```
z1 = length.dem
z1[which.min(length.dem)] = NA
democratic[which.min(z1)] #WilliamJClinton, 46 years old
```

```
[1] "WilliamJClinton"
```

Hide

Hide

```
intersect(no.grad, democratic[which.min(z1)] )
```

```
character(0)
```

From this part, we see both parties have a decreasing trend of number of words in each sentence. Interestingly, we found both of two republican presidents who use the shortest sentences did not attend graduate school while both of the democratic presidents who use the shortest sentences attended graduate school.

We then suspect if there is another factor affecting the length of sentences. We searched the age of the four presidents, and found president Trump was 70 years old and president Bush was 64 years old while president Clinton was 46 years old and president Johnson was 64 years old when they gave the inauguration speech. We suspect that will the length of sentence correlate to the age of the presidents. Here, we used the data from Wikipedia. (https://en.wikipedia.org/wiki/List_of_presidents_of_the_United_States_by_age (https://en.wikipedia.org/wiki/List_of_presidents_of_the_United_States_by_age))

Hide

Hide

```
# Presidents that use shorter sentences
age.name = c('GeorgeBush','DonaldJTrump','LyndonBJohnson','GeorgeWBush','WilliamJClin
ton','RonaldReagan','RichardNixon','FranklinDRoosevelt','HarrySTruman','DwightDEisenh
ower','BarackObama','HerbertHoover','JimmyCarter','WarrenGHarding')
# Age of the presidents that use shorter sentences
age = c(64, 70, 64, 54, 46, 69, 56, 51, 60, 62, 47, 54, 52, 55)
cat('median age of president with shorter sentence:', median(age))
```

```
median age of president with shorter sentence: 55.5
```

According to Wikipedia, the median age of presidents when they entered the White House is 55.6 years old, which is close to the median age of the above data we found. Hence, there is not a correlation between age and length of sentences.

Now, let's investigate if there is a difference in number of words in each sentence between first and second term.

Hide

Hide

Hide

```
first.term = speech.list$File[which(speech.list$Term==1)]
first.term.ind = which(speech.list$Term==1)
second.term = speech.list$File[which(speech.list$Term==2)]
second.term.ind = which(speech.list$Term==2)
mean.first.term.sentence.length = 0
for(i in first.term.ind){
  mean.first.term.sentence.length = mean.first.term.sentence.length + mean(sentence.l
ength(i, speech.list))
}
mean.first.term.sentence.length = mean.first.term.sentence.length/length(first.term.i
nd)
mean.second.term.sentence.length = 0
for(i in second.term.ind){
  mean.second.term.sentence.length = mean.second.term.sentence.length + mean(sentence
.length(i, speech.list))
}
mean.second.term.sentence.length = mean.second.term.sentence.length/length(second.ter
m.ind)
cat('average number of words in each sentence for first term', mean.first.term.senten
ce.length,'\n')
```

```
average number of words in each sentence for first term 25.64794
```

Hide

Hide

```
cat('average number of words in each sentence for second term', mean.second.term.sent
ence.length,'\n')
```

```
average number of words in each sentence for second term 24.87544
```

Therefore, there is not a significant difference in between inauguration speech of first and second term.

Summary: In part 3.1, we found the average sentence length for republican presidents are lower than democratic presidents, and this may due to the fact that there are fewer republican presidnets attended graduate school as compared to democratic presidents. We also found going to graduate school before 1890 has a stronger effect on length of sentences.

Step 3.2 Data Analytics - Length of each word

After analyzing sentence length of president inauguration speeches, let's investigate the length of the word that the presidents use. Here, we predict attending graduate school may be correlated with the length of the word that a president use such that a president who attended graduate school may use a longer word.

We define a function to count the number of character in each word. Since we are interested in the average length of big words that presidents use, we do not consider words with less than five letters.

Hide

Hide

```
word.length <- function(index, df){
  speech.vec = df$fulltext
  a = strsplit(speech.vec[index], '\\s')
  res.vec = c()
  for(j in 1:length(a[[1]])){
    res.vec = c(res.vec, nchar(a[[1]][[j]]))
  }
  res.vec = (res.vec[res.vec > 5])
  return(res.vec)
}
```

Again, same as before, let's see if there is difference in the average word length between republican and democratic presidents.

Hide

Hide

```
# word length of republican and democratic
word.length.dem <- c()
for(i in democratic.speech.ind){
  word.length.dem =  c(word.length.dem, mean(word.length(i, speech.list)))
}
word.length.rep = c()
for(i in republican.speech.ind){
  word.length.rep =  c(word.length.rep, mean(word.length(i, speech.list)))
}
cat('average word length of republican presidents:', mean(word.length.rep), '\n')
```

```
average word length of republican presidents: 8.166799
```

Hide

Hide

```
cat('average word length of democratic presidents:', mean(word.length.dem), '\n')
```

```
average word length of democratic presidents: 8.169233
```

The average word length of both parties are very similar. Then, next look at whether going to graduate school has an effect on average word length.

Hide

Hide

```
no.grad.word.length <- 0
for(i in no.grad.a1890.ind){
  no.grad.word.length = no.grad.word.length + mean(word.length(i,speech.list))
}
no.grad.word.length = (no.grad.word.length)/length(no.grad.a1890.ind)
grad.word.length <- 0
for(i in grad.a1890.ind){
  grad.word.length = grad.word.length + mean(word.length(i,speech.list))
}
grad.word.length = (grad.word.length)/length(grad.a1890.ind)
cat('average word length of presidents who attended graduate school:',grad.word.lengt
h,'\n')
```

```
average word length of presidents who attended graduate school: 8.006391
```

Hide

Hide

```
cat('average word length of presidents who did not attend graduate school:',no.grad.w
ord.length,'\n')
```

```
average word length of presidents who did not attend graduate school: 8.203511
```

Again, there is not a significant difference.

Summary: In part 3.2, we conclude there are not a significant difference in average word length in between the two parties as well as in between presidents who attended graduate school and presidents who did not attend graduate school.

Step 3.3 Data Analytics - Sentiment Analysis

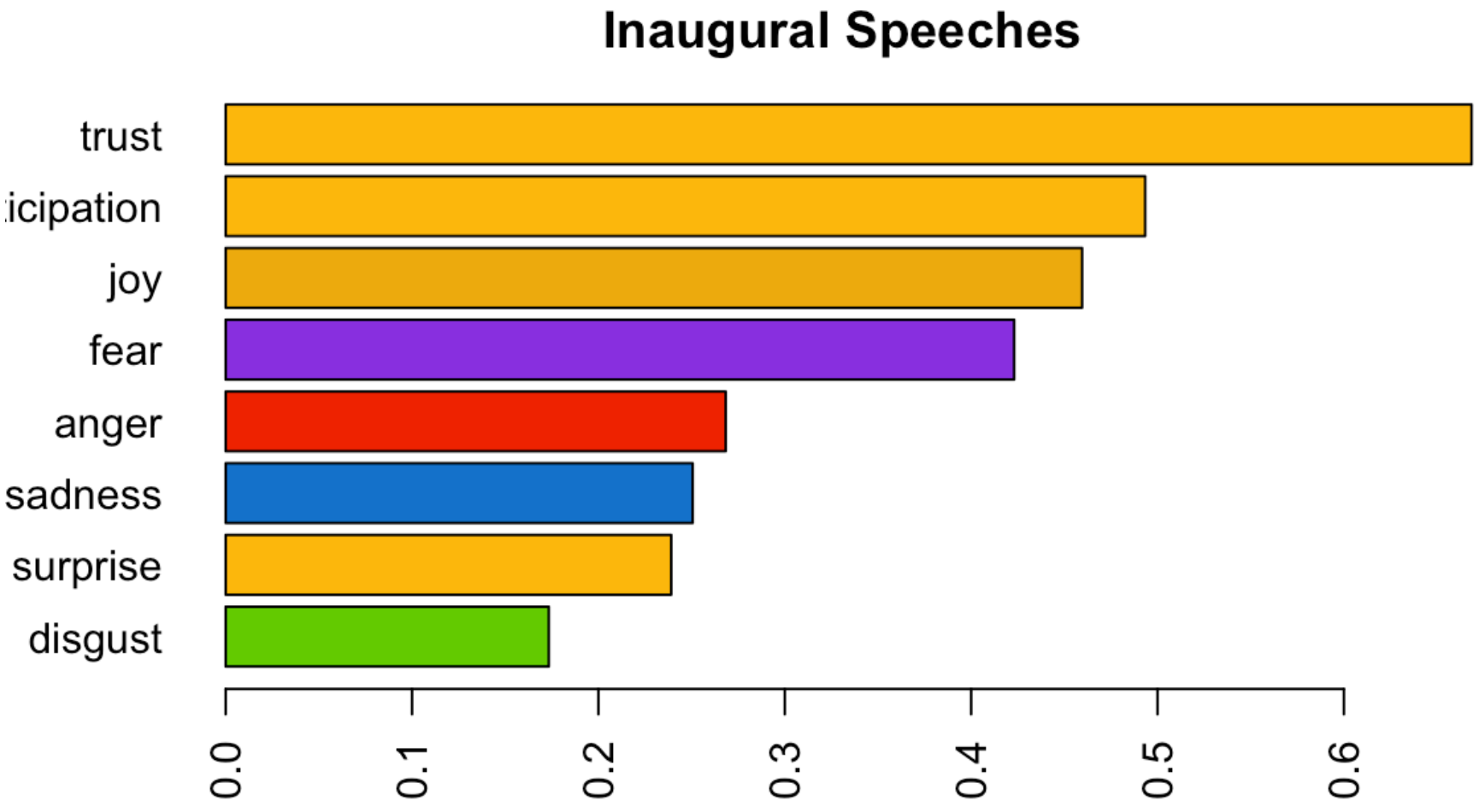
After analyzing both word length and sentence length, let's dig deeper and study the emotion of the inauguration speeches.

First, let's take a look of the overall emotion of all the presidential inauguration speeches.

Hide

Hide

```
emo.means=colMeans(select(sentence.list, anger:trust)>0.01)
col.use=c("red2", "darkgoldenrod1",
            "chartreuse3", "blueviolet",
            "darkgoldenrod2", "dodgerblue3",
            "darkgoldenrod1", "darkgoldenrod1")
barplot(emo.means[order(emo.means)], las=2, col=col.use[order(emo.means)], horiz=T, m
ain="Inaugural Speeches")
```

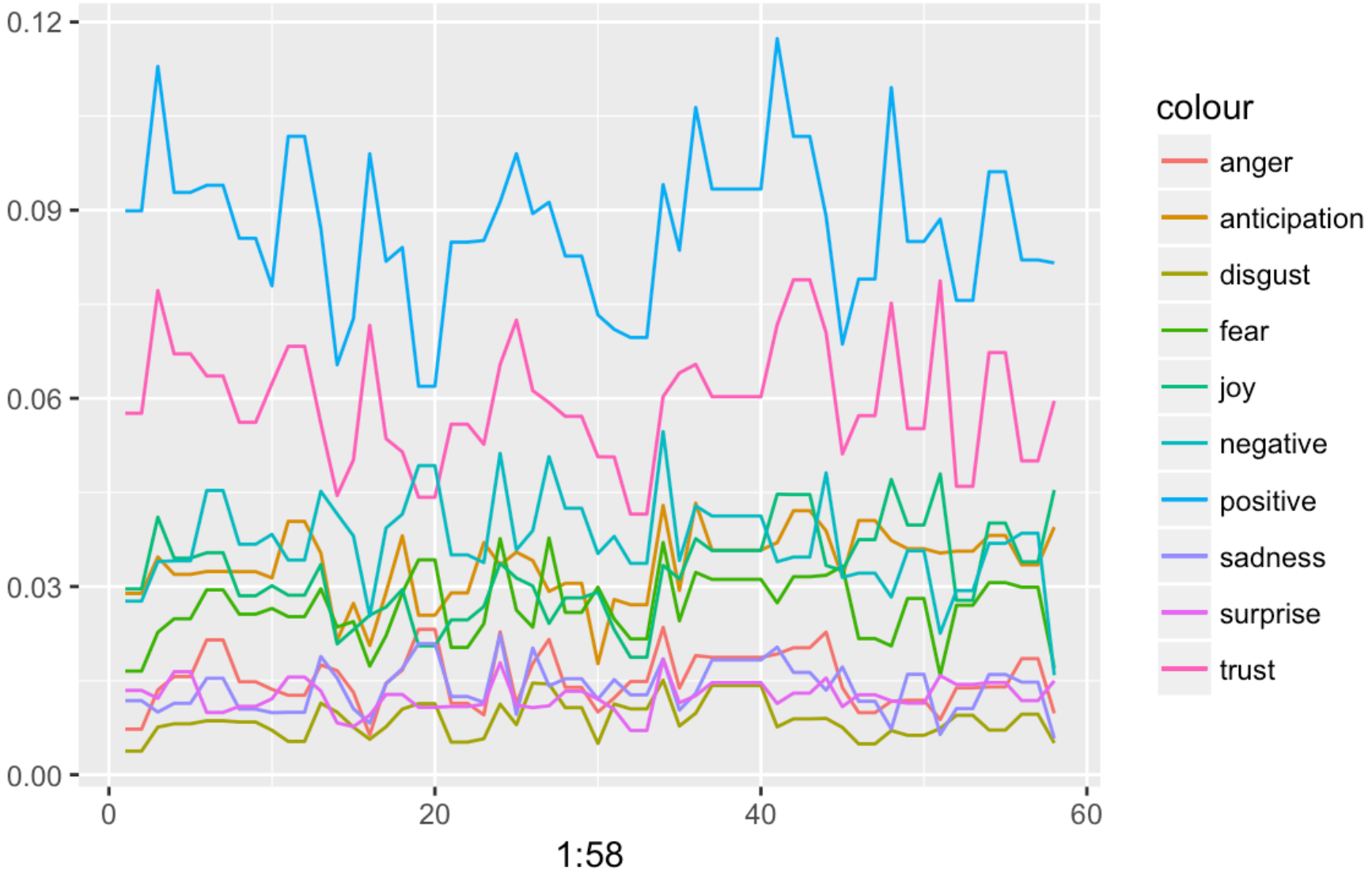


From the above graph, we see most of the presidents like to use words that are related to trust, anticipation, and joy in their inauguration speech. Then, let's take a look to see if the proportion of each emotion changes over the history.

Hide

Hide

```
anger = c(); anticipation =c(); disgust=c(); fear =c(); joy=c(); sadness=c(); surpris
e=c();trust=c();negative=c();positive=c()
for(i in 1:58){
  ran.ind = which(sentence.list$File == speech.list$File[i])
  anger = c(anger, apply(sentence.list[ran.ind,13:22],2, mean)[1])
  anticipation = c(anticipation, apply(sentence.list[ran.ind,13:22],2, mean)[2])
  disgust = c(disgust, apply(sentence.list[ran.ind,13:22],2, mean)[3])
  fear = c(fear, apply(sentence.list[ran.ind,13:22],2, mean)[4])
  joy = c(joy, apply(sentence.list[ran.ind,13:22],2, mean)[5])
  sadness = c(sadness, apply(sentence.list[ran.ind,13:22],2, mean)[6])
  surprise = c(surprise, apply(sentence.list[ran.ind,13:22],2, mean)[7])
  trust = c(trust, apply(sentence.list[ran.ind,13:22],2, mean)[8])
  negative = c(negative, apply(sentence.list[ran.ind,13:22],2, mean)[9])
  positive = c(positive, apply(sentence.list[ran.ind,13:22],2, mean)[10])
}
ggplot()+
  geom_line(aes(x=1:58,y=anger,color='anger'))+
  geom_line(aes(x=1:58,y=anticipation,color='anticipation'))+
  geom_line(aes(x=1:58,y=disgust,color='disgust'))+
  geom_line(aes(x=1:58,y=fear,color='fear'))+
  geom_line(aes(x=1:58,y=joy,color='joy'))+
  geom_line(aes(x=1:58,y=sadness,color='sadness'))+
  geom_line(aes(x=1:58,y=surprise,color='surprise'))+
  geom_line(aes(x=1:58,y=trust,color='trust'))+
  geom_line(aes(x=1:58,y=negative,color='negative'))+
  geom_line(aes(x=1:58,y=positive,color='positive'))+
  ylab('')
```



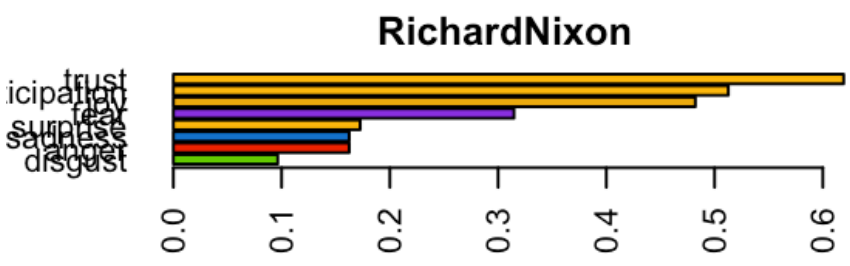
The overall trending of each emotion remain relatively stable except we see a drop in ‘positive’, ‘anticipate’, and ‘trust’ words and a raise in ‘negative’, ‘fear’, ‘anger’, and ‘sadness’ words, which happened at when the inauguration speech of Abraham Lincoln was made. Presidents Lincoln led the US through Civil War, which is often described as the bloodiest war in the US. This may explain the increased in negative words within this period of time. Also, we see a raise in ‘positive’, ‘joy’, and ‘trust’ words in the inauguration speech of Harry S. Truman. President Truman led US toward the victory of World War II, which may explain the raise of the positive emotion words in his inauguration speech.

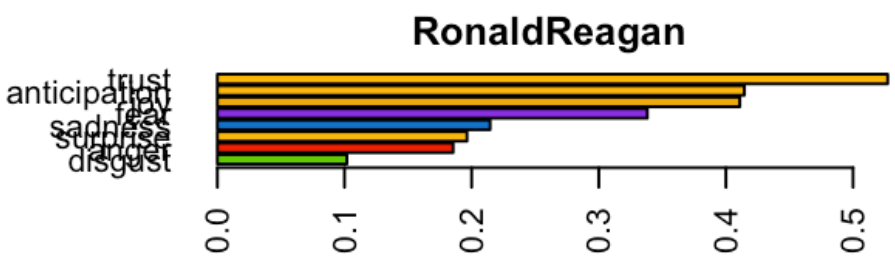
Next, we can investigate the difference in emotion expressed by presidents in both parties. We randomly select a few presidents from each party.

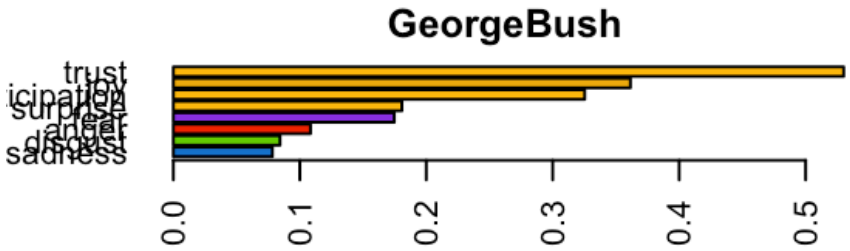
Hide

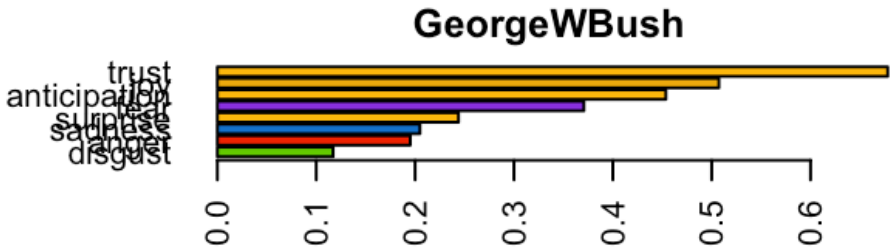
Hide

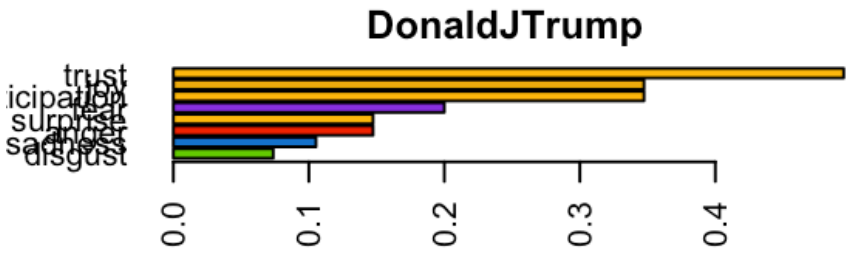
```
# Now, lets look at the emotion of some rep presidents
random.rep=c('RichardNixon','RonaldReagan','GeorgeBush','GeorgeWBush','DonaldJTrump')
par(mfrow = c(3, 2))
for(i in random.rep){
  ran.ind = which(sentence.list$File == i)
  emo.means=colMeans(select(sentence.list[ran.ind,], anger:trust)>0.01)
col.use=c("red2", "darkgoldenrod1",
            "chartreuse3", "blueviolet",
            "darkgoldenrod2", "dodgerblue3",
            "darkgoldenrod1", "darkgoldenrod1")
  barplot(emo.means[order(emo.means)], las=2, col=col.use[order(emo.means)], horiz=T,
main=i)
}
# lets look at the emotion of some dem presidents
random.dem = c('FranklinDRoosevelt','JohnFKennedy','LyndonBJohnson','WilliamJClinton'
, 'BarackObama')
par(mfrow = c(3,2))
```







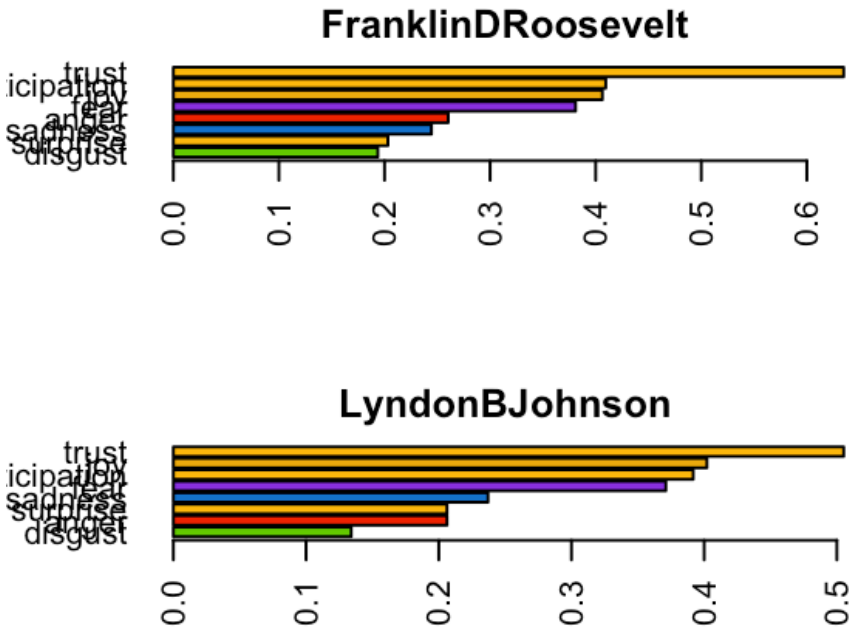


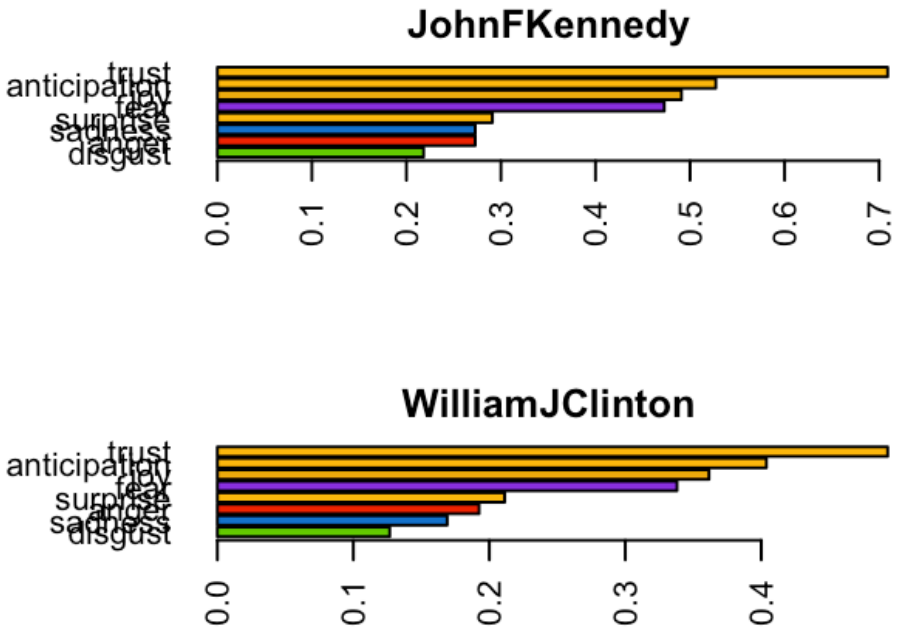


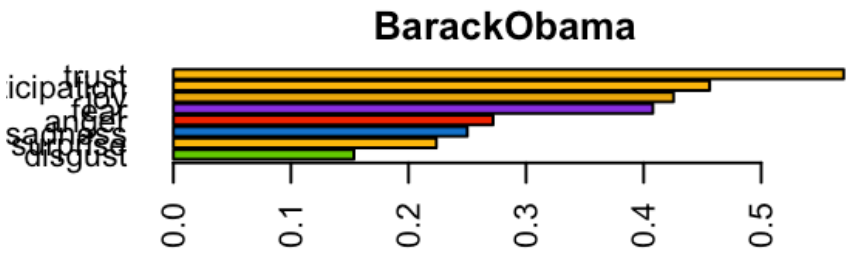
Hide

Hide

```
for(i in random.dem){
  ran.ind = which(sentence.list$File == i)
  emo.means=colMeans(select(sentence.list[ran.ind,], anger:trust)>0.01)
col.use=c("red2", "darkgoldenrod1",
            "chartreuse3", "blueviolet",
            "darkgoldenrod2", "dodgerblue3",
            "darkgoldenrod1", "darkgoldenrod1")
  barplot(emo.means[order(emo.means)], las=2, col=col.use[order(emo.means)], horiz=T,
main=i)
}
```







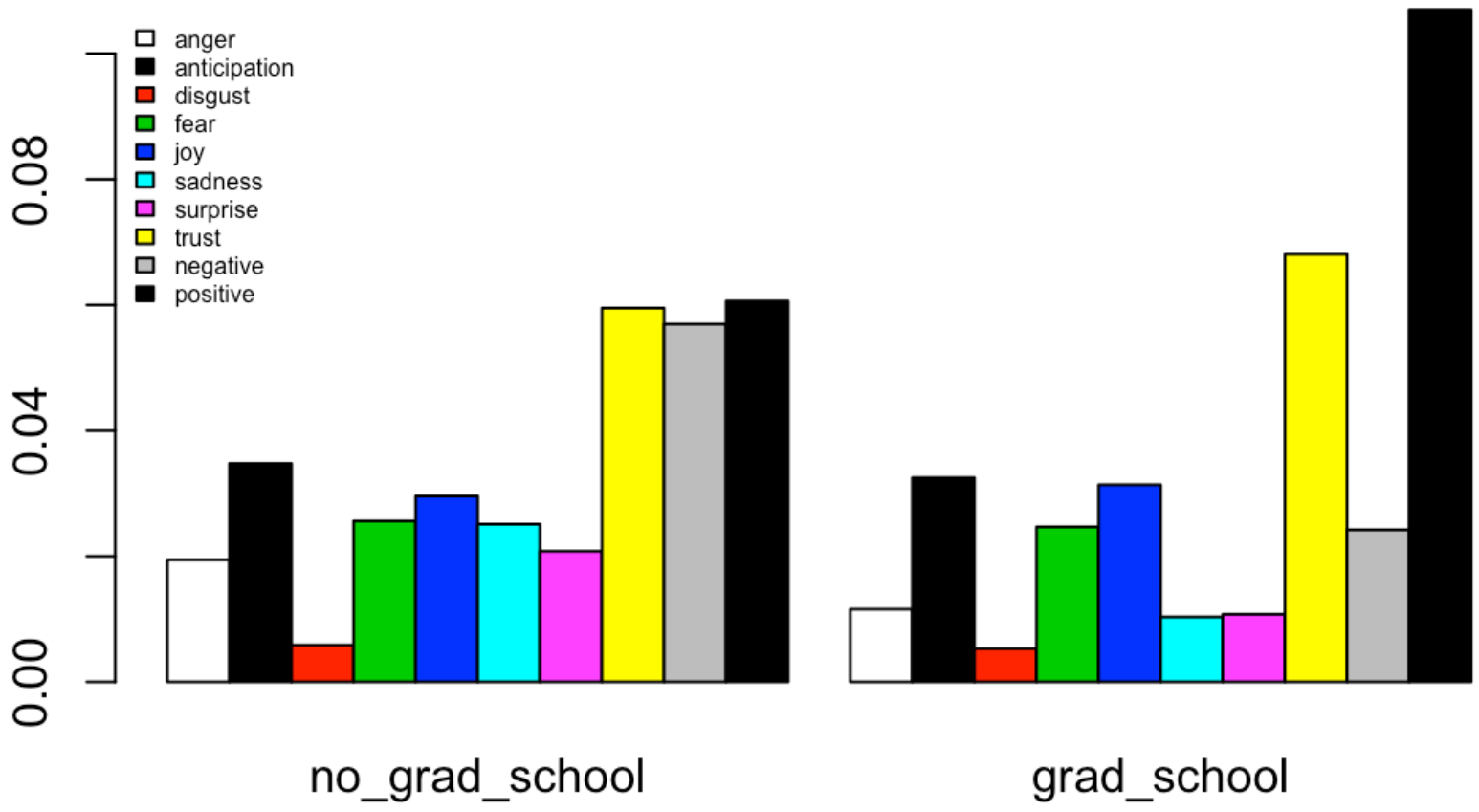
Interestly, we found the emotion in the inauguration speeches of democratic presidents are very consistent while the emotion of republican presidents' inauguration speeches varies among different presidents.

Next, we investigate the emotion of inauguration speeches between president who attended graduate school and president who did not attend graduate school.

Hide

Hide

```
no_grad_school = apply(sentence.list[no.grad.a1890.ind,13:22],2, mean)
grad_school = apply(sentence.list[grad.a1890.ind,13:22],2, mean)
barplot(as.matrix(data.frame(no_grad_school,grad_school)), beside=T, col = 0:9)
legend("topleft", c("anger","anticipation","disgust","fear","joy",'sadness','surprise
','trust','negative','positive'), cex=0.5, bty="n", fill=0:9)
```



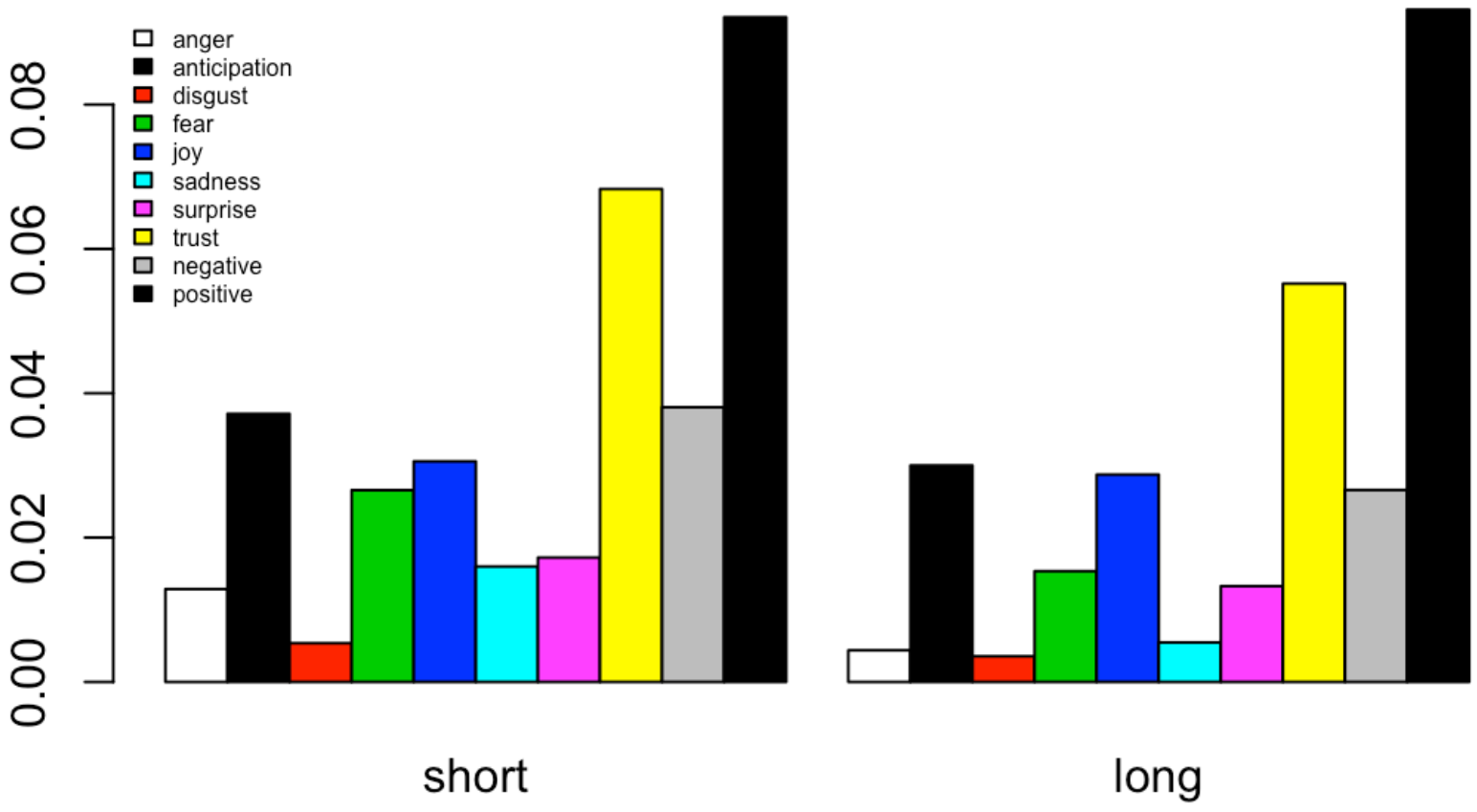
We found presidents who attended graduate school use significantly more positive words and significantly fewer negative words. Presidents who attended graduate school tend to use more ‘trust’ related words and fewer ‘sadness’ and ‘surprise’ related words. This may raise the possibilites hypothesis of presidents who attend graduate tend to give a more positive inauguration speech. However, further investigation and data is needed for additional evidence of this prespective.

Now, we investigate the emotion of the sentences in different length.

Hide

Hide

```
# short sentence vs long sentence
short.ind = which(length.all[11:58] <= mean(length.all))
long.ind = which(length.all[11:58] > mean(length.all))
short = apply(sentence.list[short.ind,13:22],2, mean)
long = apply(sentence.list[long.ind,13:22],2, mean)
barplot(as.matrix(data.frame(short,long)), beside=T, col = 0:9)
legend("topleft", c("anger","anticipation","disgust","fear","joy",'sadness','surprise
','trust','negative','positive'), cex=0.5, bty="n", fill=0:9)
```



Here, we found longer sentences tend to have fewer “negative”, “sadness”, and “fear” related words.

In summary, in 3.3, we analyzed emotion of the inauguration speeches. We found democratic presidents’ inauguration speeches have a more consistent proportion of each emotion while republican presidents’ inauguration speeches do not have a consistent pattern. Also, presidents who went to graduate school tend to give a more positive inauguration speech, and presidents who use longer sentences tend to use fewer negative words.

Part 3.4 Data Analytics: Topic Modeling

Now, we focus on unsupervise study of the speeches. We apply topic modeling into all the inauguration speeches to find the common topics in among the speeches. We then categorize each speech into a topic that it is mostly likely to be.

To do so, we first need to apply natural language processing to our data to remove some unnecessary content, such as number, extra spaces, and meaningless words. We also stem the document, which only the stem part of the word will be remained. For example, ‘supplied’ and ‘supplies’ both become ‘suppli’.

Then, we create a big of words, which is a matrix show in a dummy variable format.

Hide

Hide

```
dtm <- DocumentTermMatrix(docs)
#convert rownames to filenames#convert rownames to filenames
rownames(dtm) <- paste(corpus.list$type, corpus.list$File,
                       corpus.list$Term, corpus.list$sent.id, sep="_")
rowTotals <- apply(dtm , 1, sum) #Find the sum of words in each Document
dtm  <- dtm[rowTotals> 0, ]
corpus.list=corpus.list[rowTotals>0, ]
```

Now, we apply LDA topic modeling into our bag of words.

Hide

Hide

```
#Set parameters for Gibbs sampling
burnin <- 4000 # removing the first 4000 samples
iter <- 2000
thin <- 500 #choose 500th in each round, drop the 499 others
#500 is using every 500th guess because each guess is based on its previous guess. so
n to n+1 won't have much difference but n and n+500th guess would be much difference
and intelligent
seed <-list(2003,5,63,100001,765)
nstart <- 5
best <- TRUE
#Number of topics
k <- 10
#Run LDA using Gibbs sampling
ldaOut <-LDA(dtm, k, method="Gibbs", control=list(nstart=nstart,
                                                  seed = seed, best=best,
                                                  burnin = burnin, iter = iter,
                                                  thin=thin))
#write out results
#docs to topics
ldaOut.topics <- as.matrix(topics(ldaOut))
table(c(1:k, ldaOut.topics))
```

```
  1   2   3   4   5   6   7   8   9  10
565 977 629 555 585 457 435 419 483 436
```

Hide

Hide

```
write.csv(ldaOut.topics,file=paste("~/Desktop/Columbia/Semester 2/ADS/Project1/wk2-Te
xtMining/out/LDAGibbs",k,"DocsToTopics.csv"))
#top 20 terms in each topic
ldaOut.terms <- as.matrix(terms(ldaOut,20))
write.csv(ldaOut.terms,file=paste("~/Desktop/Columbia/Semester 2/ADS/Project1/wk2-Tex
tMining/out/LDAGibbs",k,"TopicsToTerms.csv"))
#probabilities associated with each topic assignment
topicProbabilities <- as.data.frame(ldaOut@gamma)
topicProbabilities$max = apply(topicProbabilities, 1, max)
write.csv(topicProbabilities,file=paste("~/Desktop/Columbia/Semester 2/ADS/Project1/w
k2-TextMining/out/LDAGibbs",k,"TopicProbabilities.csv"))
# use beta distribution
terms.beta=ldaOut@beta
terms.beta=scale(terms.beta)
topics.terms=NULL
for(i in 1:k){
  topics.terms=rbind(topics.terms, ldaOut@terms[order(terms.beta[i,], decreasing = TR
UE)[1:7]])
}
```

After we obtain the list of topic, we look at the word each topic contains and give a title to each topic. Here, we set the title name into a vector, topics.hash.

Hide

```
topics.hash = c('government','american dream', 'economy', 'faith', 'anticipate', 'peo
ple', 'civil right', 'patriot', 'legistration', 'peace', 'max')
corpus.list$ldatopic=as.vector(ldaOut.topics)
#corpus.list$ldahash=topics.hash[ldaOut]
colnames(topicProbabilities)=topics.hash
corpus.list.df=cbind(corpus.list, topicProbabilities)
```

Now, we plot our data into a heatmap to see if some presidents share very similar topics. In a heat map, the same color means represent the same probability. The redder the color is means the higher the probability.

Hide

Hide

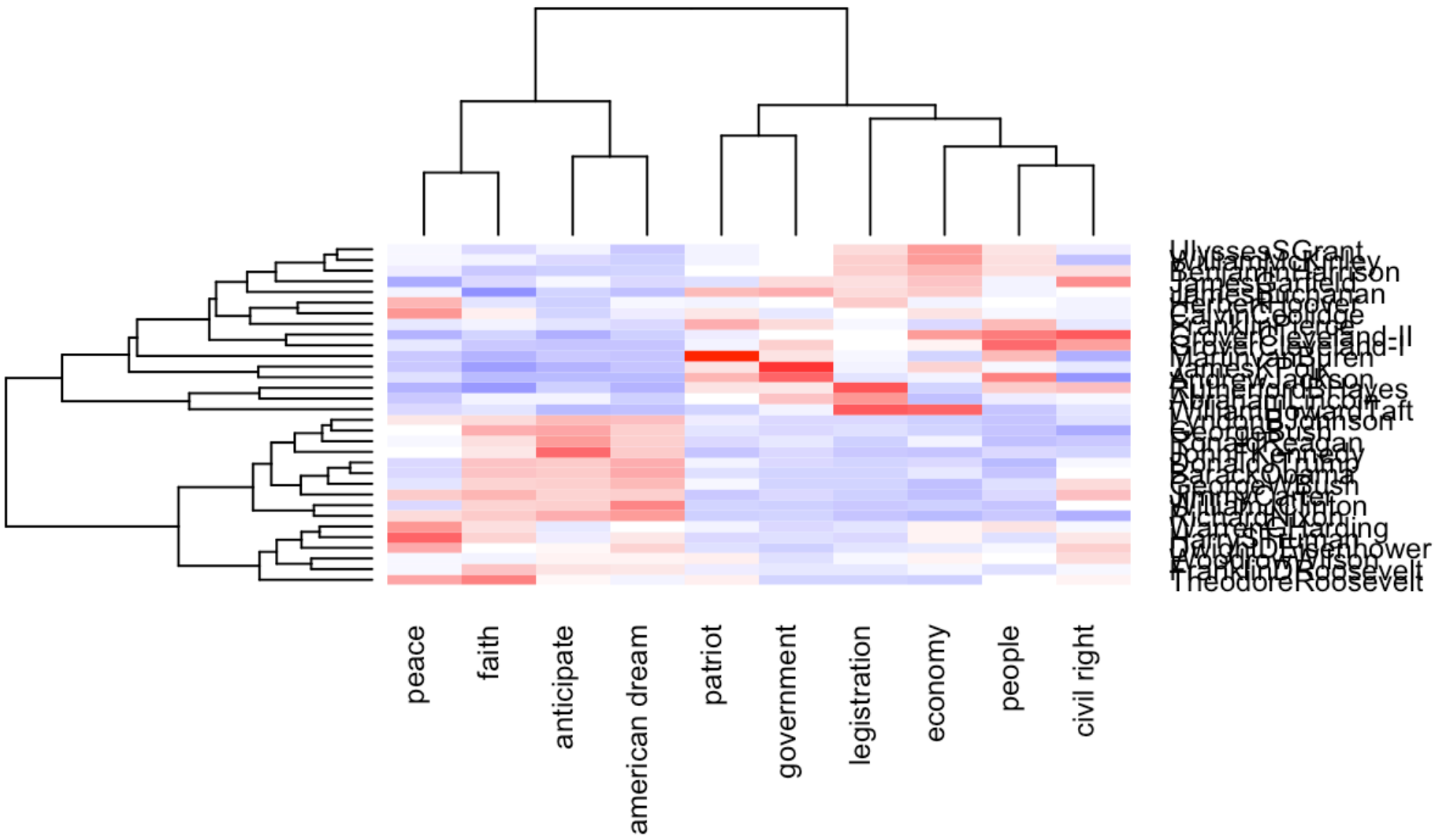
```
par(mar=c(1,1,1,1))
topic.summary=tbl_df(corpus.list.df)%>%
              filter(type%in%c('inaug'), File%in%c(democratic, republican))%>%
              select(File, government:peace)%>%
              group_by(File)%>%
              summarise_each(funs(mean))
```

```
`summarise_each()` is deprecated.
Use `summarise_all()`, `summarise_at()` or `summarise_if()` instead.
To map `funs` over all variables, use `summarise_all()`
```

Hide

Hide

```
topic.summary=as.data.frame(topic.summary)
rownames(topic.summary)=topic.summary[,1]
heatmap.2(as.matrix(topic.summary[,topic.plot+1]),
          scale = "column", key=F,
          col = bluered(100),
          cexRow = 0.9, cexCol = 0.9, margins = c(8, 8),
          trace = "none", density.info = "none")
```



From the heatmap, we saw Dwight D Eisenhower(Korean War), Harry S Truman(WW II), Warren G Harding (right after WW I), and Calbin Coolidge(post WW I) mentioned the topic of ‘peace’ in their inauguration speeches. As I analyzed more about each of their presidency, I found President Eisenhower was in the office during Korean War, President Truman was elected near the very end of World War II, President Harding was in the office right after World War I, and President Coolidge was in the office during the post World War I period. This may explain why they focus to ‘peace’ related topic.

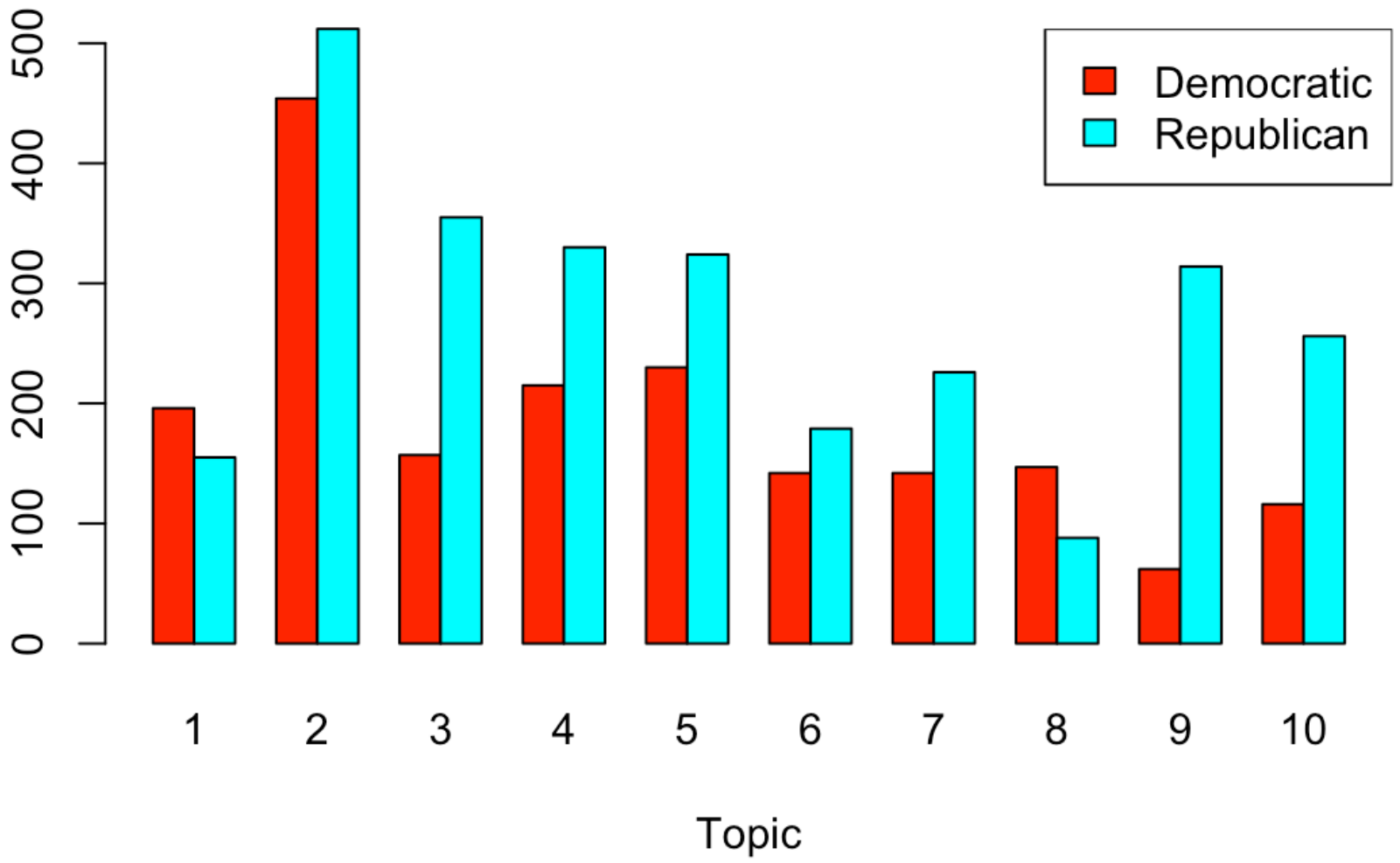
William Howard Taft, Grover Cleveland II, Ulysses S Grant, and William Mckinley mentioned the topic of ‘economy’. As we researched more about their presidential policy and history, we found President Taft was involved in dollary diplomacy, President Cleveland was overwhelmed by nation’s economic disasters-depression, President Grant was in the office during the Gilded Age, a massive industrial growth, and President Mckinley led American through the rapid economic growth. These may explain the expression of ‘economy’ related topic in their inauguration speeches.

Now, let’s investigate the topic difference in between republican and democratic.

Hide

Hide

```
dem.rep.table = table(corpus.list.df[,c('Party','ldatopic')])[c(1,4),]
barplot(dem.rep.table, beside = T, col=c(2,5), xlab = 'Topic')
legend('topright', legend = c('Democratic','Republican'), fill=c(2,5))
```



As we can see from above, republican presidents tend to cover more about topic 3, 9, and 10, which are ‘economy’, ‘legistration’, and ‘peace’ while democratic presidents tend to cover more about topic 1 and 8, which are ‘government’ and ‘patriot’.

Now, let’s investigate on the topic that presidents who attended graduate school tend to cover.

Hide

Hide

```
z.no.grad = corpus.list.df[c(no.grad.dem.sentence.ind, no.grad.rep.sentence.ind),c('l
datopic')]
grad.sentence.ind = (1:nrow(corpus.list.df))[-c(no.grad.dem.sentence.ind,no.grad.rep.
sentence.ind)][692:nrow(corpus.list.df)]
z.grad = corpus.list.df[grad.sentence.ind,c('ldatopic')]
z.grad.table = table(z.grad)
z.no.grad.table = table(z.no.grad)
dem.rep.table = as.data.frame(cbind(z.grad.table, z.no.grad.table))
dem.rep.table
```

	z.grad.table <int>	z.no.grad.table <int>
1	156	262
2	646	321
3	110	406
4	284	260
5	354	207
6	109	239
7	183	196
8	122	163
9	161	280
10	149	232

1-10 of 10 rows

The left table shows the number of sentences in each topic of presidents who attend graduate school while the right table shows that of presidents who did not attend graudate school. As we can see from the table above, presidents who attended graduate school tend to cover topic 2 and 5 in their inauguration speech, which are ‘american dream’ and ‘anticipate’ while presidents who did not attend graduate school tend to cover topic 3, 6, and 9, which are ‘economy’, ‘people’, and ‘legistration’. This may suggest presidents who attended graduate tend to focus on the future and core value of America while the presidents who did not attend graduate school tend to focus on more pragmatic topics.

Now, let’s perform a clustering on presidents who attended graduate school and who did not attend graudate school.

Hide

Hide

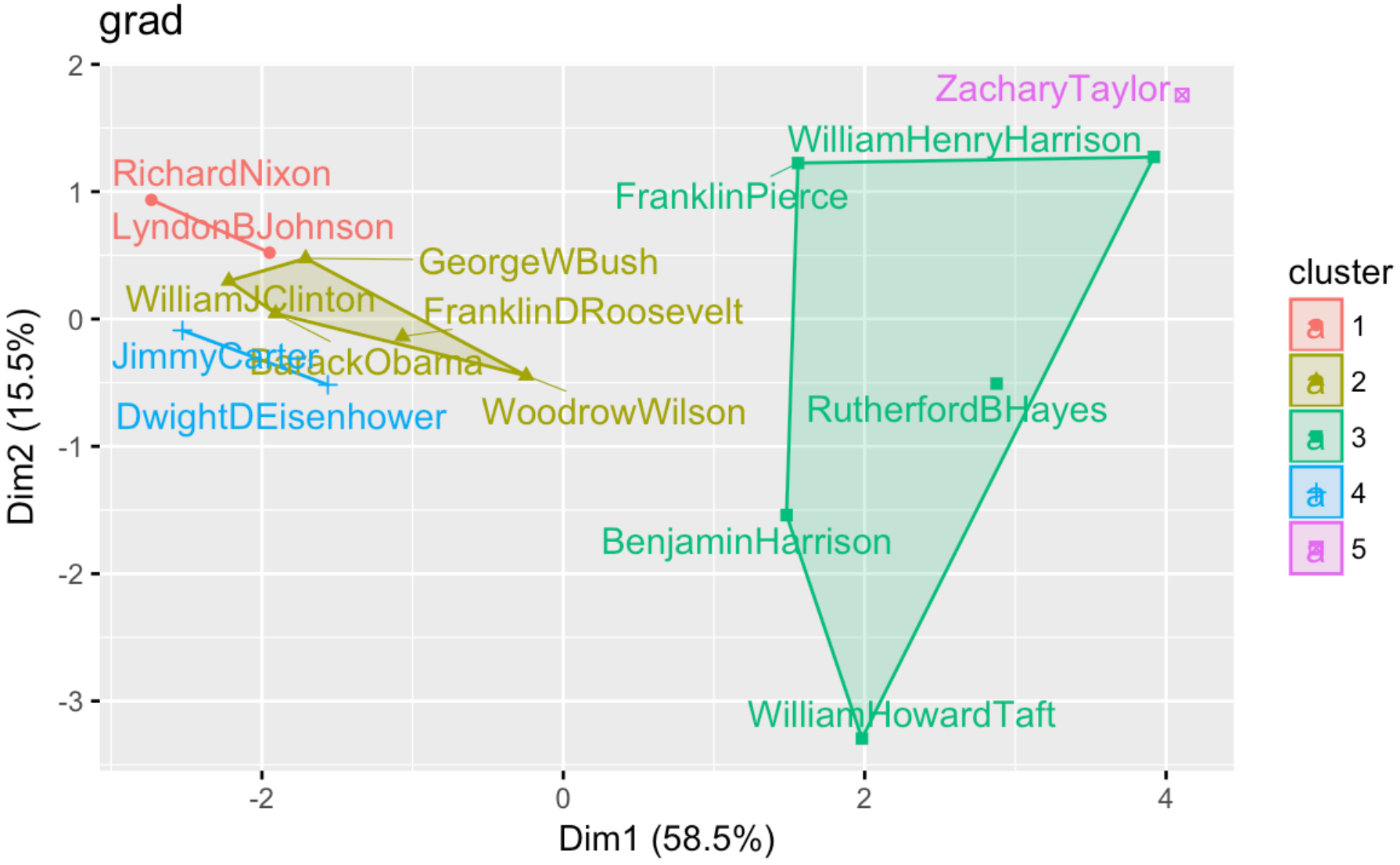
```
presid.summary=tbl_df(corpus.list.df)%>%
  filter(type=="inaug", File%in% speech.list$File[c(grad.a1890.ind,grad.b1890.ind)])%
>%
  select(File, government:peace)%>%
  group_by(File)%>%
  summarise_each(funs(mean))
```

```
`summarise_each()` is deprecated.
Use `summarise_all()`, `summarise_at()` or `summarise_if()` instead.
To map `funs` over all variables, use `summarise_all()`
```

Hide

Hide

```
presid.summary=as.data.frame(presid.summary)
rownames(presid.summary)=as.character((presid.summary[,1]))
km.res=kmeans(scale(presid.summary[,-1]), iter.max=200,
              5)
fviz_cluster(km.res,
             stand=T, repel= TRUE,
             data = presid.summary[,-1],
             show.clust.cent=FALSE, main = 'grad')
```



Hide

Hide

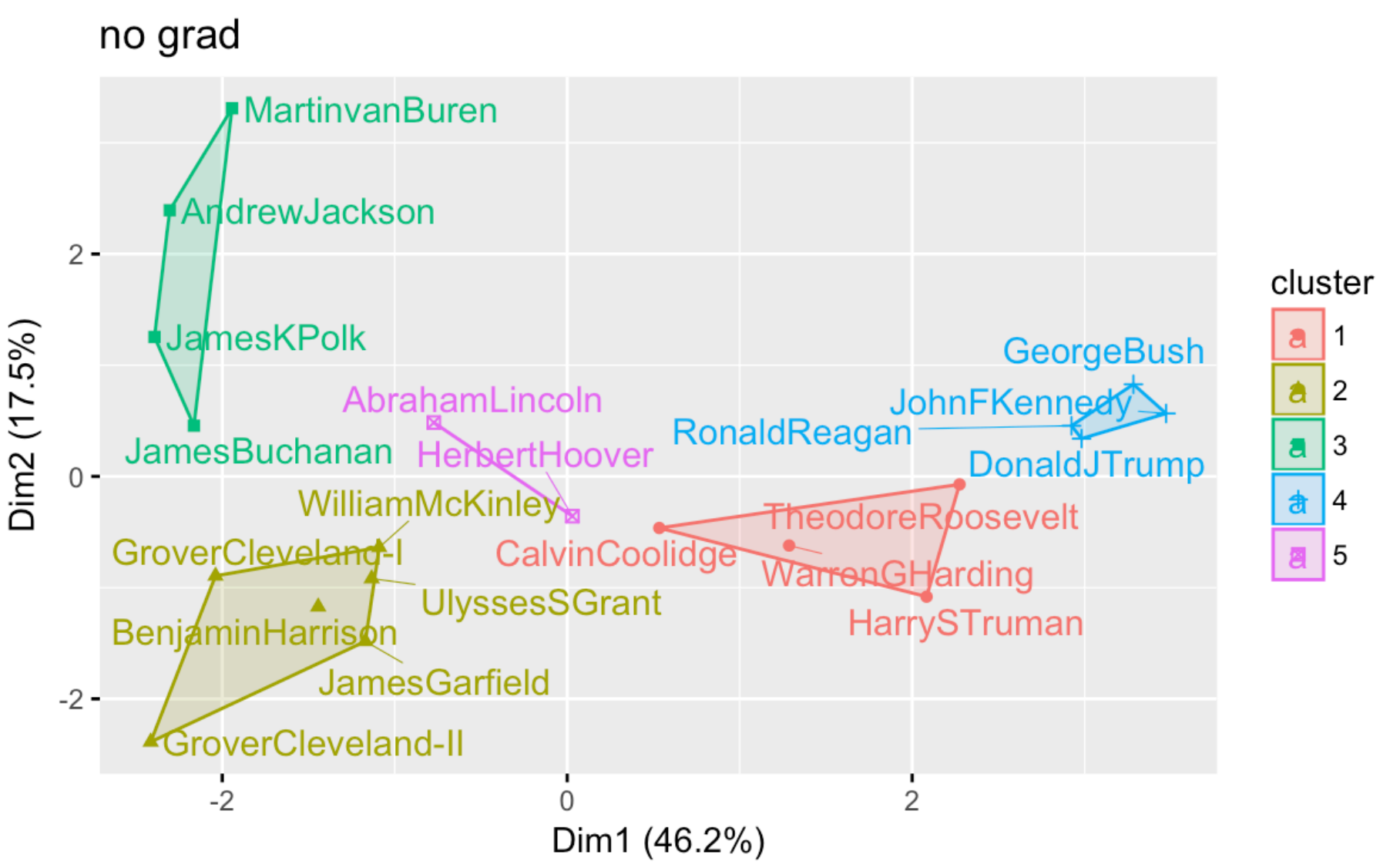
```
presid.summary=tbl_df(corpus.list.df)%>%
  filter(type=="inaug", File%in% speech.list$File[c(no.grad.a1890.ind,no.grad.b1890.i
nd)])%>%
  select(File, government:peace)%>%
  group_by(File)%>%
  summarise_each(funs(mean))
```

```
`summarise_each()` is deprecated.
Use `summarise_all()`, `summarise_at()` or `summarise_if()` instead.
To map `funs` over all variables, use `summarise_all()`
```

Hide

Hide

```
presid.summary=as.data.frame(presid.summary)
rownames(presid.summary)=as.character((presid.summary[,1]))
km.res=kmeans(scale(presid.summary[,-1]), iter.max=200,
              5)
fviz_cluster(km.res,
             stand=T, repel= TRUE,
             data = presid.summary[,-1],
             show.clust.cent=FALSE, main = 'no grad')
```



The graph in the right shows the clustering of presidents who did not attend graduate school and the graph in the left shows that of presidents who attended graduate school. As we observe the cluster plot above, we notice president from the same party tend to form it's own cluster. For example, in 'no grad' cluster, all presidents in cluster 3 are democratic while all presidents in cluster 5 are republican.

Then, we think if the presidents are clustered together, this means they gave similar inauguration speeches, which suggests they might also share similar policies. Therefore, if an American supports one of the presidents in the cluster, he/she may as well supports another president in the same cluster. Hence, we found a dataset from Wikipedia that contains approval rate of all presidents since 1937. (https://en.wikipedia.org/wiki/United_States_presidential_approval_rating (https://en.wikipedia.org/wiki/United_States_presidential_approval_rating))

Hide

Hide

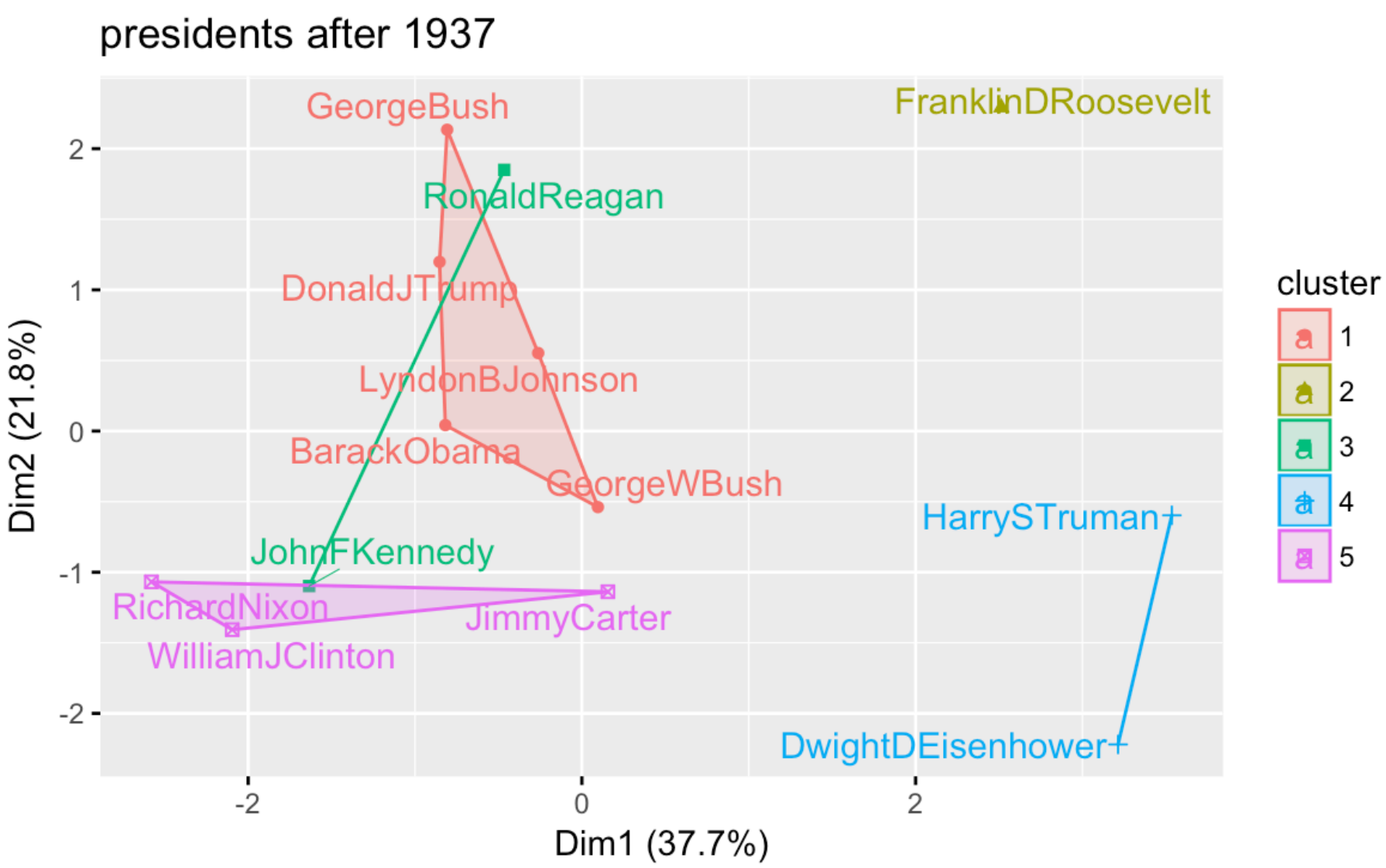
```
presid.summary=tbl_df(corpus.list.df)%>%
  filter(type=="inaug", File%in% speech.list$File[37:58])%>%
  select(File, government:peace)%>%
  group_by(File)%>%
  summarise_each(funs(mean))
```

```
`summarise_each()` is deprecated.
Use `summarise_all()`, `summarise_at()` or `summarise_if()` instead.
To map `funs` over all variables, use `summarise_all()`
```

Hide

Hide

```
presid.summary=as.data.frame(presid.summary)
rownames(presid.summary)=as.character((presid.summary[,1]))
km.res=kmeans(scale(presid.summary[,-1]), iter.max=200,
              5)
fviz_cluster(km.res,
             stand=T, repel= TRUE,
             data = presid.summary[,-1],
             show.clust.cent=FALSE, main = 'presidents after 1937')
```



From the graph, we look at cluster 2, where there are three presidents including Richard Nixon, William Clinton, and Barack Obama. According to the historical approval rate from Wikipedia, Preisdent Nixon has an average approval rate of 49.1, President Clinton has an average approval rate of 55.1, President Obama has an average approval rate of 47.9. Also, if we look at cluster 4, where there are four presidents including George Bush, Franklin Roosevelt, Lyndon Johnson, and Donald Trump. As President Trump has just been into an office for a year, we will account him into this approval rate analysis. President Johnson has an average approval rate of 55.1, President Bush has an average approval rate of 60.9, and President Roosevelt has an average approval rate of 63. These results match our hypothesis of presidents have similar inauguration speech tend to have similar approval rate. But further analysis is required as the collection of approval rate started in 1937 meaning we only have very limited data.

Summary: In part 3.4, we applied topic modeling into the inauguration speeches, we found presidents from differents parties tend to focus on different topics and presidents who attended to graduate school also tend to focus on different topics. Furthermore, we found presidents in the same cluster tend to be in the same party and have the similar approval rate.

Conclusion

In conclusion, in this study, we showed that Republican Presidents tend to use shorter sentences as compared to Democratic Presidents, and this may due to the fact that there are fewer Republican Presidents who attended graduate school. We also showed that Presidents who attended graduate school tend to use more positive words in their inauguration speech while presidents who use longer sentences tend to use fewer negative words. Furthermore, we found that going to graduate school before 1890 may have a stronger

impact in sentence length as graduate school were harder to get in before 1890. Finally, presidents from the same party tend to cover similar topics in their inauguration speech, and presidents who mention similar topics in their inauguration speech may share similar approval rate.

Reference

https://en.wikipedia.org/wiki/United_States_presidential_approval_rating
(https://en.wikipedia.org/wiki/United_States_presidential_approval_rating)
https://en.wikipedia.org/wiki/List_of_presidents_of_the_United_States_by_age
(https://en.wikipedia.org/wiki/List_of_presidents_of_the_United_States_by_age)
https://en.wikipedia.org/wiki/History_of_higher_education_in_the_United_States
(https://en.wikipedia.org/wiki/History_of_higher_education_in_the_United_States)
https://en.wikipedia.org/wiki/List_of_Presidents_of_the_United_States_by_education
(https://en.wikipedia.org/wiki/List_of_Presidents_of_the_United_States_by_education)
https://en.wikipedia.org/wiki/List_of_Presidents_of_the_United_States
(https://en.wikipedia.org/wiki/List_of_Presidents_of_the_United_States)