Project 1

Wenyuan Gu

Step 1:Data Importing

```
#Load packages
library("rvest")
library("tibble")
library("qdap")
library("sentimentr")
library("gplots")
library("dplyr")
library("tm")
library("syuzhet")
library("factoextra")
library("beeswarm")
library("scales")
library("RColorBrewer")
library("RANN")
library("tm")
library("topicmodels")
library("readxl")

source("../lib/plotstacked.R")
source("../lib/speechFuncs.R")

#read speeches
inaug <- read_excel("~/Spring2018-Project1-wenyuangu/data/InaugurationInfo.xlsx")

inaug <- inaug[-41,] #remove line 41 whose speech data is missing
inaug <- inaug[-57,] #remove line 57 because we do not analyse it now

inaug$fulltext <- NA
for (i in seq(nrow(inaug))){
  text <- readLines(paste0("~/Spring2018-Project1-wenyuangu/data/InauguralSpeeches/
inaug",inaug$File[i],"-",inaug$Term[i],".txt"))
  inaug$fulltext[i] <- text
}                   # add full text
```

Step 2:Data Processing

Here, I categorize our data according to party.We can see from the data that there are 6 categories, which are “Fedralist”,“Democratic-Republican Party”,“Democratic”,“Whig”,“Republican” and “NA”.Since some of these categories contain very few presidents,we only analyse two of them,“Democratic” and “Republican”.

```
inaug$Words <- as.numeric(inaug$Words)

unique(inaug$Party)
```

```
## [1] "NA"                            "Fedralist"
## [3] "Democratic-Republican Party" "Democratic"
## [5] "Whig"                          "Republican"
```

```
democ <- inaug[inaug$Party == "Democratic",]
repub <- inaug[inaug$Party == "Republican",]
inaug <- rbind(democ,repub)

nrow(democ)
```

```
## [1] 21
```

```
nrow(repub)
```

```
## [1] 23
```

```
unique(democ$President)
```

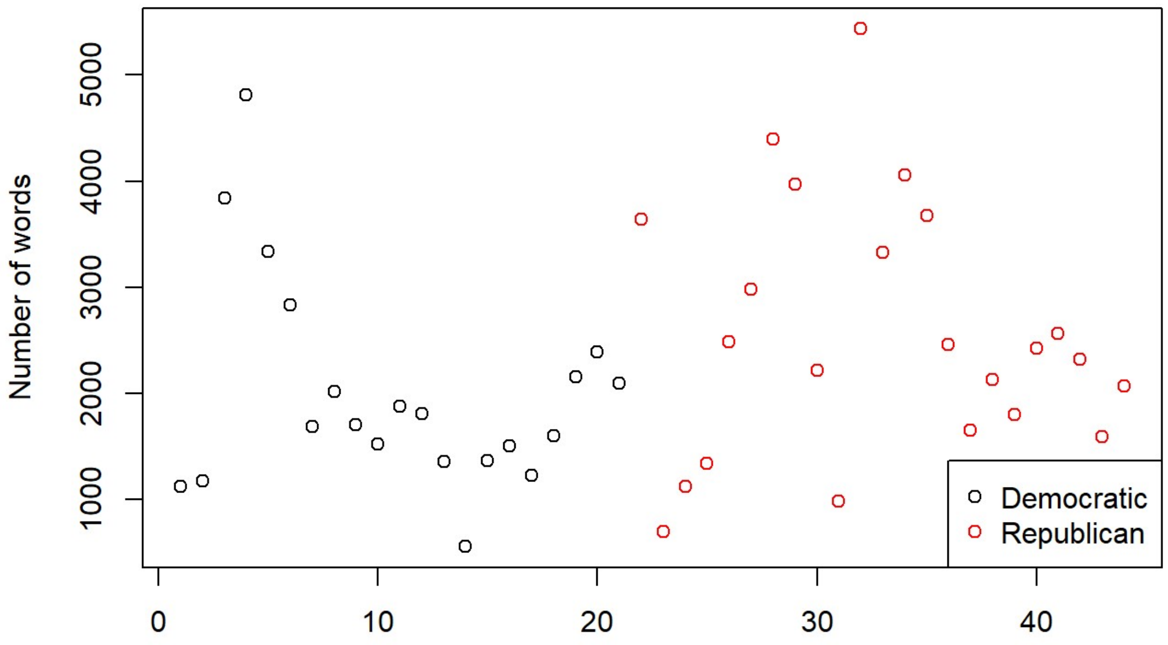
```
##  [1] "Andrew Jackson"        "Martin van Buren"
##  [3] "James K. Polk"         "Franklin Pierce"
##  [5] "James Buchanan"        "Grover Cleveland - I"
##  [7] "Grover Cleveland - II" "Woodrow Wilson"
##  [9] "Franklin D. Roosevelt" "John F. Kennedy"
## [11] "Lyndon B. Johnson"     "Jimmy Carter"
## [13] "William J. Clinton"    "Barack Obama"
```

```
unique(repub$President)
```

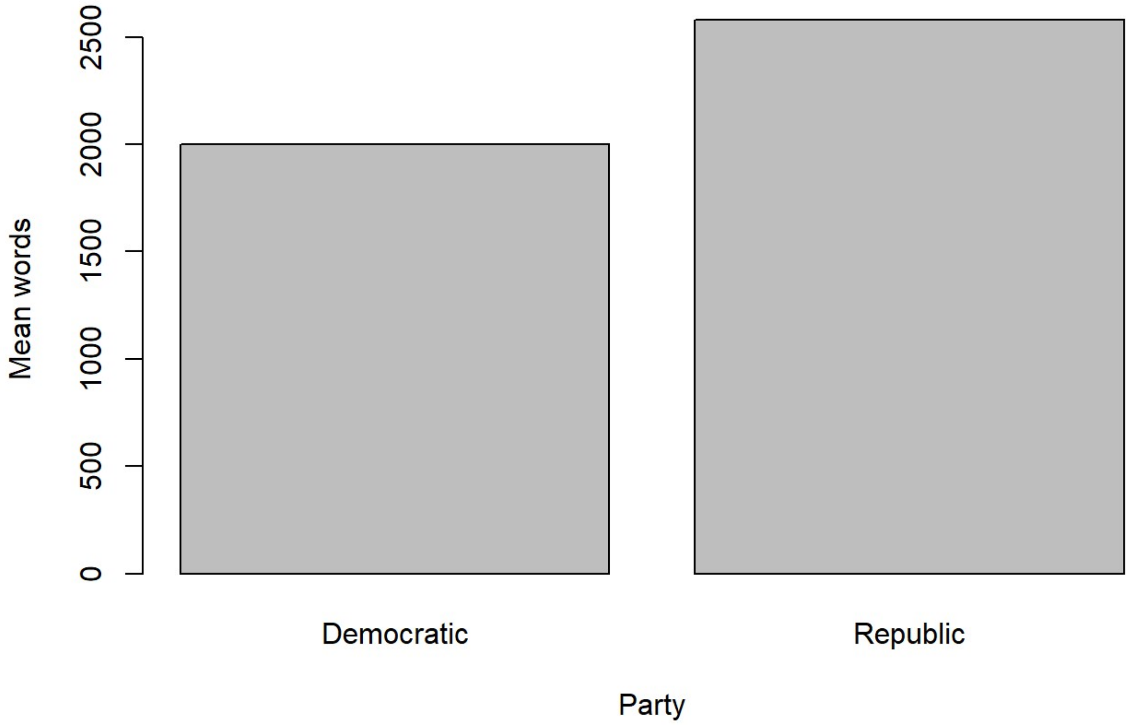
```
##  [1] "Abraham Lincoln"      "Ulysses S. Grant"     "Rutherford B. Hayes"
##  [4] "James Garfield"       "Benjamin Harrison"    "William McKinley"
##  [7] "Theodore Roosevelt"   "William Howard Taft"  "Warren G. Harding"
## [10] "Calvin Coolidge"      "Herbert Hoover"       "Dwight D. Eisenhower"
## [13] "Richard Nixon"        "Ronald Reagan"        "George Bush"
## [16] "George W. Bush"
```

We can see there are 14 democratic presidents giving 21 speeches in total,which for republic,there are 16 presidents giving 23 speeches.

```
#a rough scan on number of words.

plot(inaug$Words,col = factor(inaug$Party),xlab = "",ylab = "Number of words")
legend("bottomright", legend = levels(factor(inaug$Party)), col=1:length(levels(fac
tor(inaug$Party))), pch=1)
```



```
barplot(c(mean(democ$Words),mean(repub$Words,na.rm = TRUE)),names.arg = c("Democrat
ic","Republic"),xlab = "Party",ylab = "Mean words")
```



The plot shows that presidents from republic party tend to give a longer speech.Also, republic presidents’ speeches are more dispersed,while democratic presidents’ speeches are centered around 2000 words,generally.

Next,we generate list of sentences respectively.

We will use sentences as units of analysis for this project, as sentences are natural languge units for organizing thoughts and ideas. For each extracted sentence, we apply sentiment analysis using NRC sentiment lexion (http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm). “The NRC Emotion Lexicon is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive).

```
sentence.list.dem <- NULL
sentence.list.rep <- NULL

for(i in 1:nrow(democ)){
  sentences=sent_detect(democ$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.dem=rbind(sentence.list.dem, cbind(democ[i,],
                              sentences=as.character(sentences),
                              word.count,
                              emotions,
                              sent.id=1:length(sentences))
    )
  }
}

i<-24
for(i in 1:nrow(repub)){
  sentences=sent_detect(repub$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.rep=rbind(sentence.list.rep, cbind(repub[i,],
                              sentences=as.character(sentences),
                              word.count,
                              emotions,
                              sent.id=1:length(sentences))
    )
  }
}
```

Some non-sentences exist in raw data due to erroneous extra end-of-sentence marks.

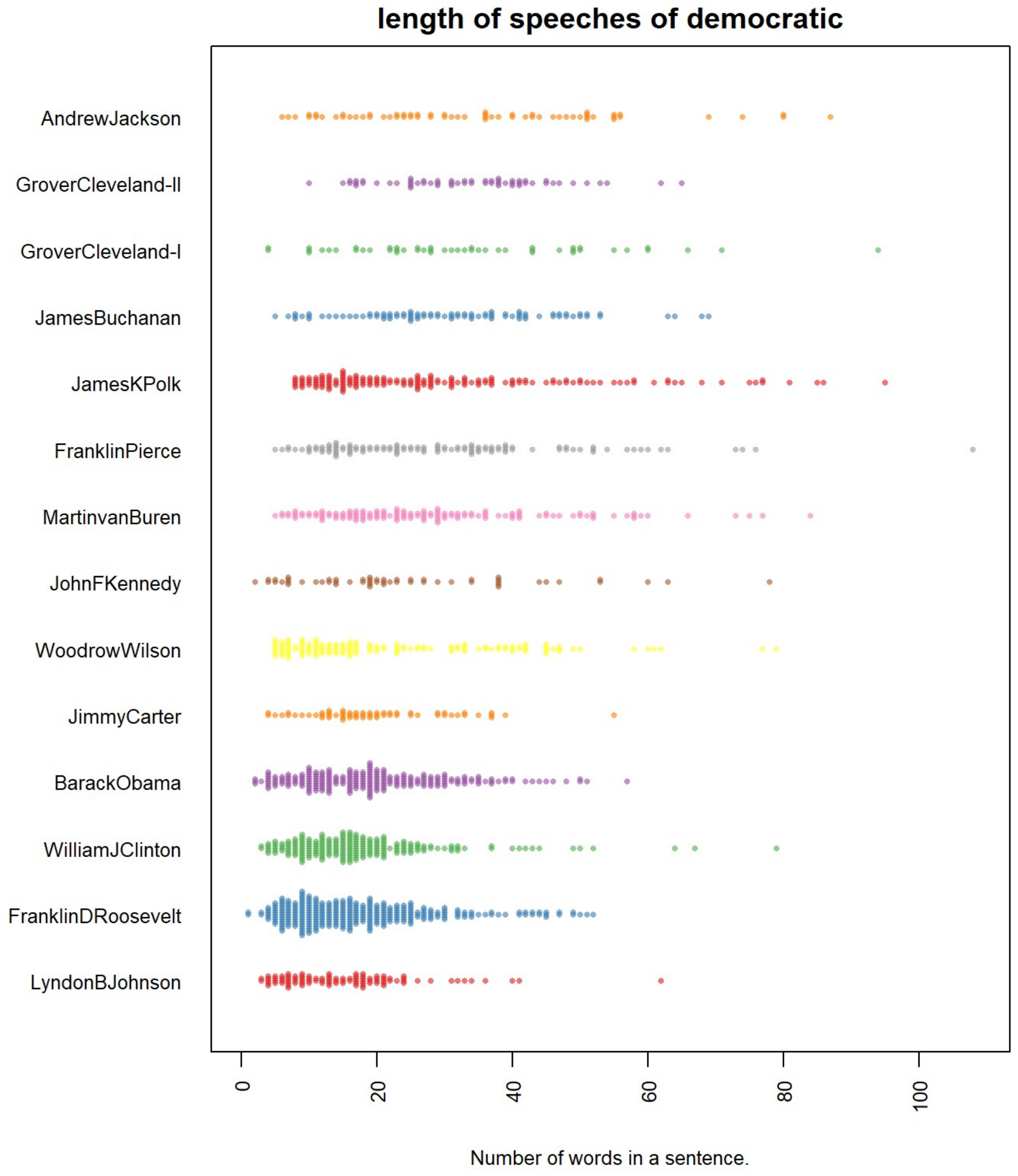
```
sentence.list.dem=
  sentence.list.dem%>%
  filter(!is.na(word.count))

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

Step 3:Data Analysis

length of speeches

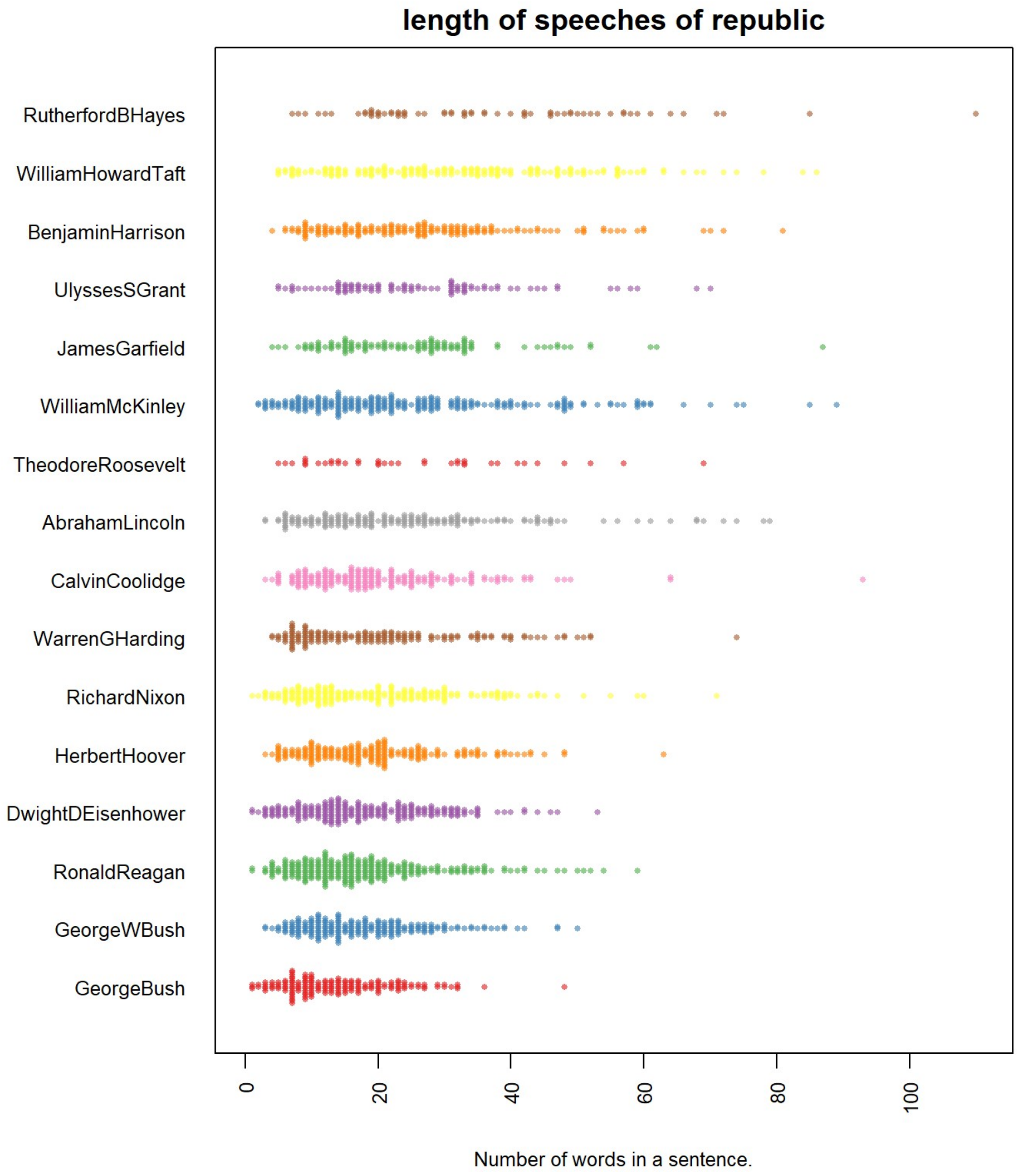
democratic

```
par(mar=c(4, 11, 2, 2))

sentence.list.dem$File=factor(sentence.list.dem$File)

sentence.list.dem$FileOrdered=reorder(sentence.list.dem$File,
                                  sentence.list.dem$word.count,
                                  mean,
                                  order=T)

beeswarm(word.count~FileOrdered,
         data=sentence.list.dem,
         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.dem$FileOrdered),
         las=2, xlab="Number of words in a sentence.", ylab="",
         main="length of speeches of democratic")
```


length of speeches of democratic
AndrewJackson
GroverCleveland-II
GroverCleveland-I
JamesBuchanan
JamesKPolk
FranklinPierce
MartinvanBuren
JohnFKennedy
WoodrowWilson
JimmyCarter
BarackObama
WilliamJClinton
FranklinDRoosevelt
LyndonBJohnson
0
20
40
60
80
100
Number of words in a sentence.

###republic

```
par(mar=c(4, 11, 2, 2))

sentence.list.rep$File=factor(sentence.list.rep$File)

sentence.list.rep$FileOrdered=reorder(sentence.list.rep$File,
                                     sentence.list.rep$word.count,
                                     mean,
                                     order=T)

beeswarm(word.count~FileOrdered,
         data=sentence.list.rep,
         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.rep$FileOrdered),
         las=2, xlab="Number of words in a sentence.", ylab="",
         main="length of speeches of republic")
```



The picture shows that in generall, republic’s sentences are shorter than democratics’.

sentiment analysis

Sentence length variation over the course of the speech, with emotions.

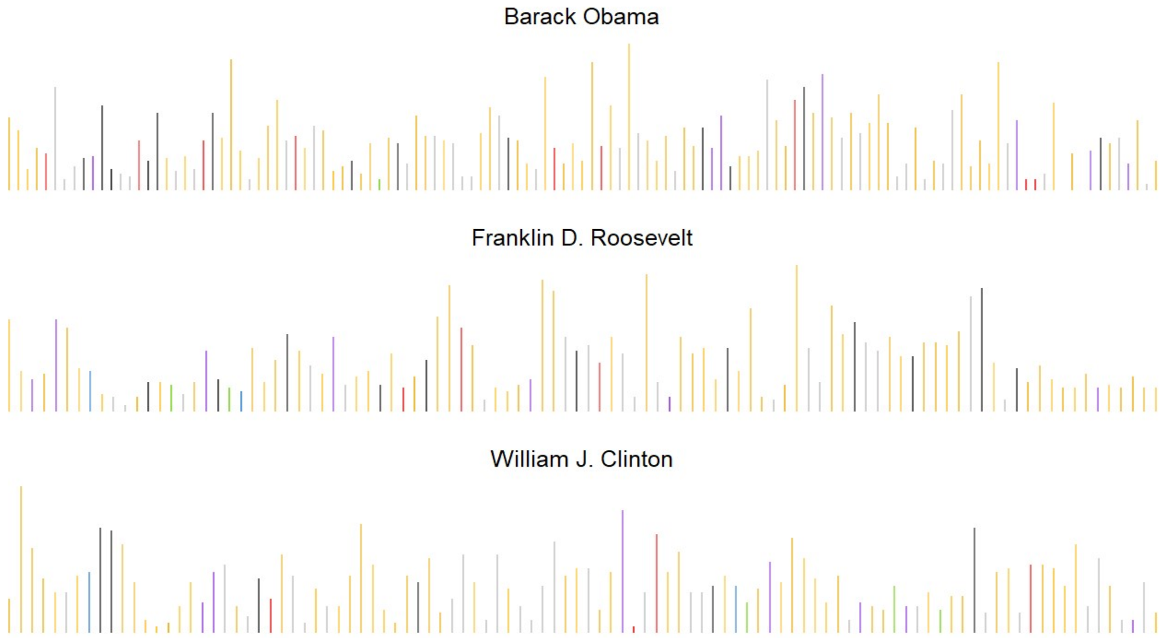
First,let’s define a function that shows sentence length variation over the course of the speech with emotions.

```
f.plotsent.len=function(In.list, InFile,InTerm, President){

  col.use=c("lightgray", "red2", "darkgoldenrod1",
            "chartreuse3", "blueviolet",
            "darkgoldenrod2", "dodgerblue3",
            "darkgoldenrod1", "darkgoldenrod1",
            "black", "darkgoldenrod2")

  In.list$topemotion=apply(select(In.list,
                                        anger:positive),
                                 1, which.max)
  In.list$topemotion.v=apply(select(In.list,
                                          anger:positive),
                                   1, max)
  In.list$topemotion[In.list$topemotion.v<0.05]=0
  In.list$topemotion=In.list$topemotion+1

  temp=In.list$topemotion.v
  In.list$topemotion.v[temp<0.05]=1

  df=In.list%>%filter(File==InFile,
                      Term==InTerm)%>%
    select(sent.id, word.count,
           topemotion, topemotion.v)

  ptcol.use=alpha(col.use[df$topemotion], sqrt(sqrt(df$topemotion.v)))

  plot(df$sent.id, df$word.count,
       col=ptcol.use,
       type="h", #ylim=c(-10, max(In.list$word.count)),
       main=President)
}
```

Then,we choose three presidents from each category and use the function above to show the sentence length variation over the course of the speech with emotions.

democratic

```
par(mfrow=c(4,1), mar=c(1,0,2,0), bty="n", xaxt="n", yaxt="n", font.main=1)

f.plotsent.len(In.list=sentence.list.dem, InFile="BarackObama",InTerm=1,President="Barack Obama")

f.plotsent.len(In.list=sentence.list.dem,InFile="FranklinDRoosevelt",InTerm=1,President="Franklin D. Roosevelt")

f.plotsent.len(In.list=sentence.list.dem,InFile="WilliamJClinton",InTerm=1,President="William J. Clinton")
```



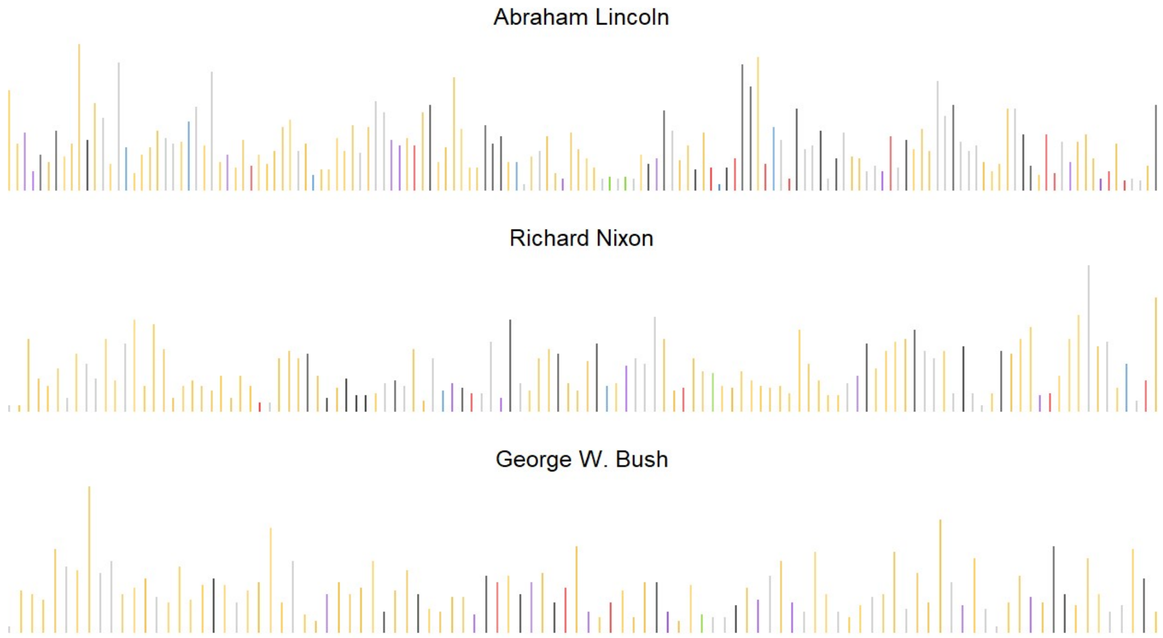
Barack Obama

Franklin D. Roosevelt

William J. Clinton

republic

```
par(mfrow=c(4,1), mar=c(1,0,2,0), bty="n", xaxt="n", yaxt="n", font.main=1)

f.plotsent.len(In.list=sentence.list.rep,InFile="AbrahamLincoln",InTerm=1,President="Abraham Lincoln")

f.plotsent.len(In.list=sentence.list.rep,InFile="RichardNixon",InTerm=1,President="Richard Nixon")

f.plotsent.len(In.list=sentence.list.rep,InFile="GeorgeWBush",InTerm=1,President="George W. Bush")
```



Abraham Lincoln

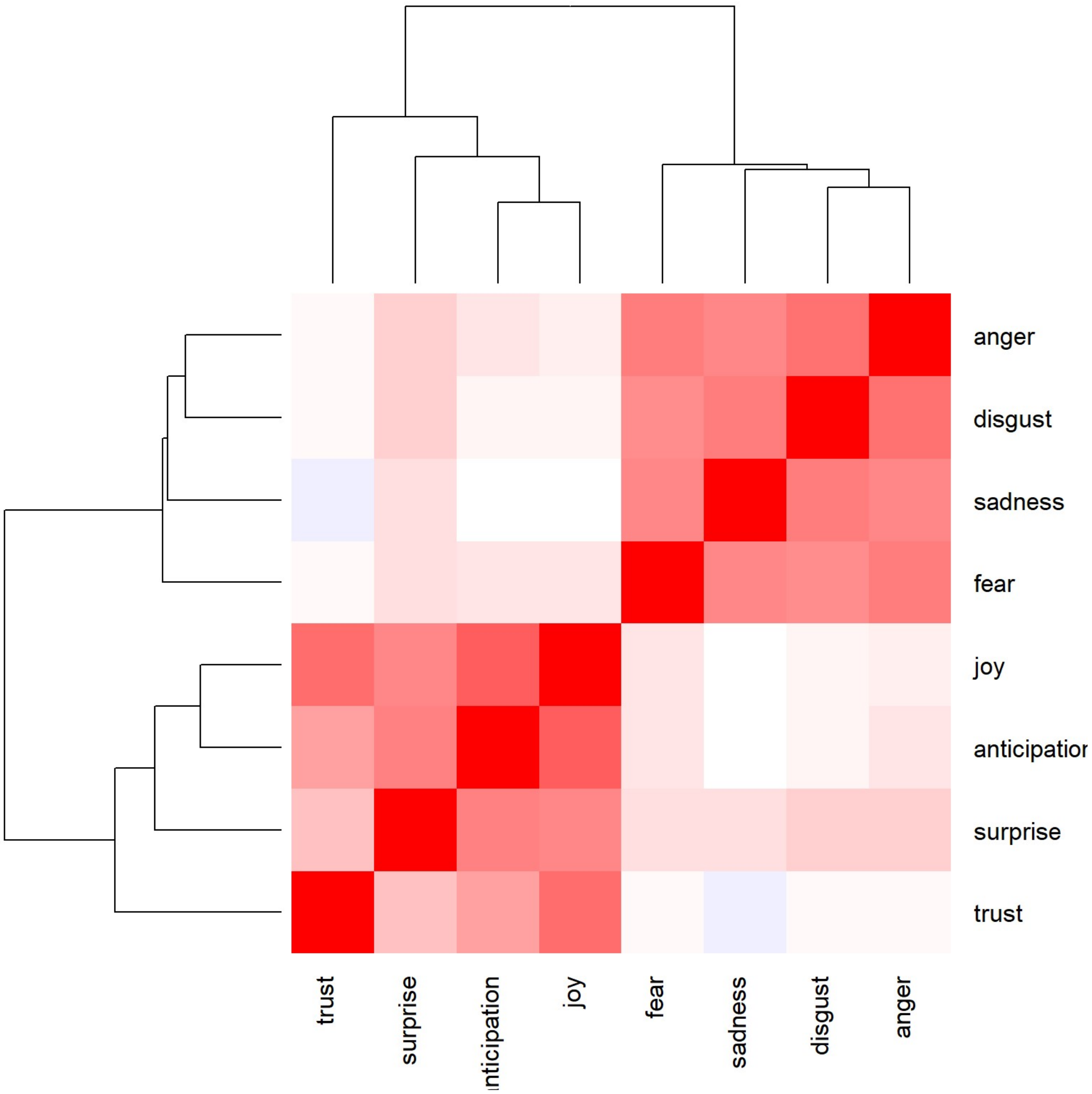
Richard Nixon

George W. Bush

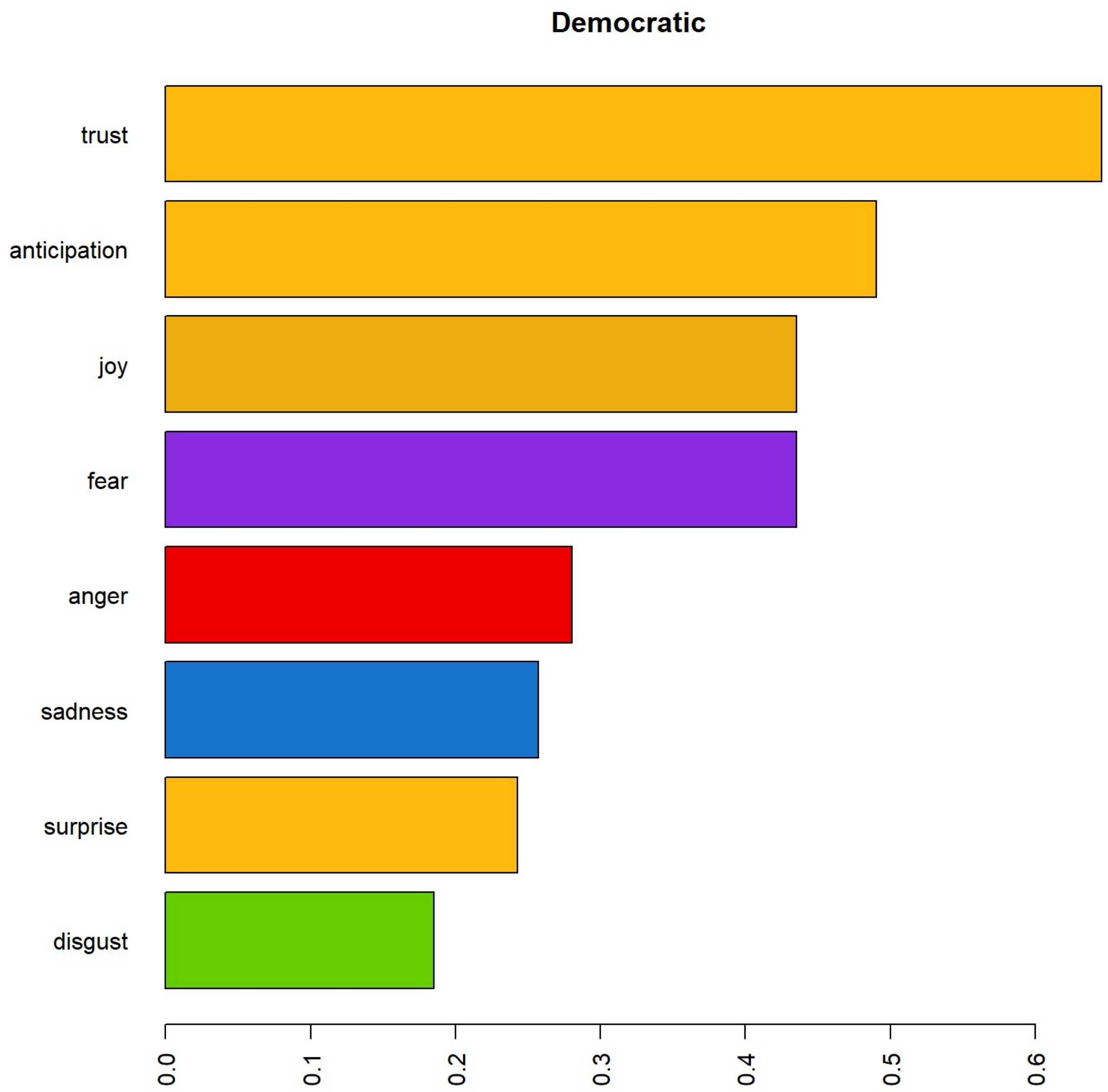
###clustering of emotions

Democratic

```
heatmap.2(cor(sentence.list.dem%>%select(anger:trust)),
          scale = "none",
          col = bluered(100), , margin=c(6, 6), key=F,
          trace = "none", density.info = "none")
```

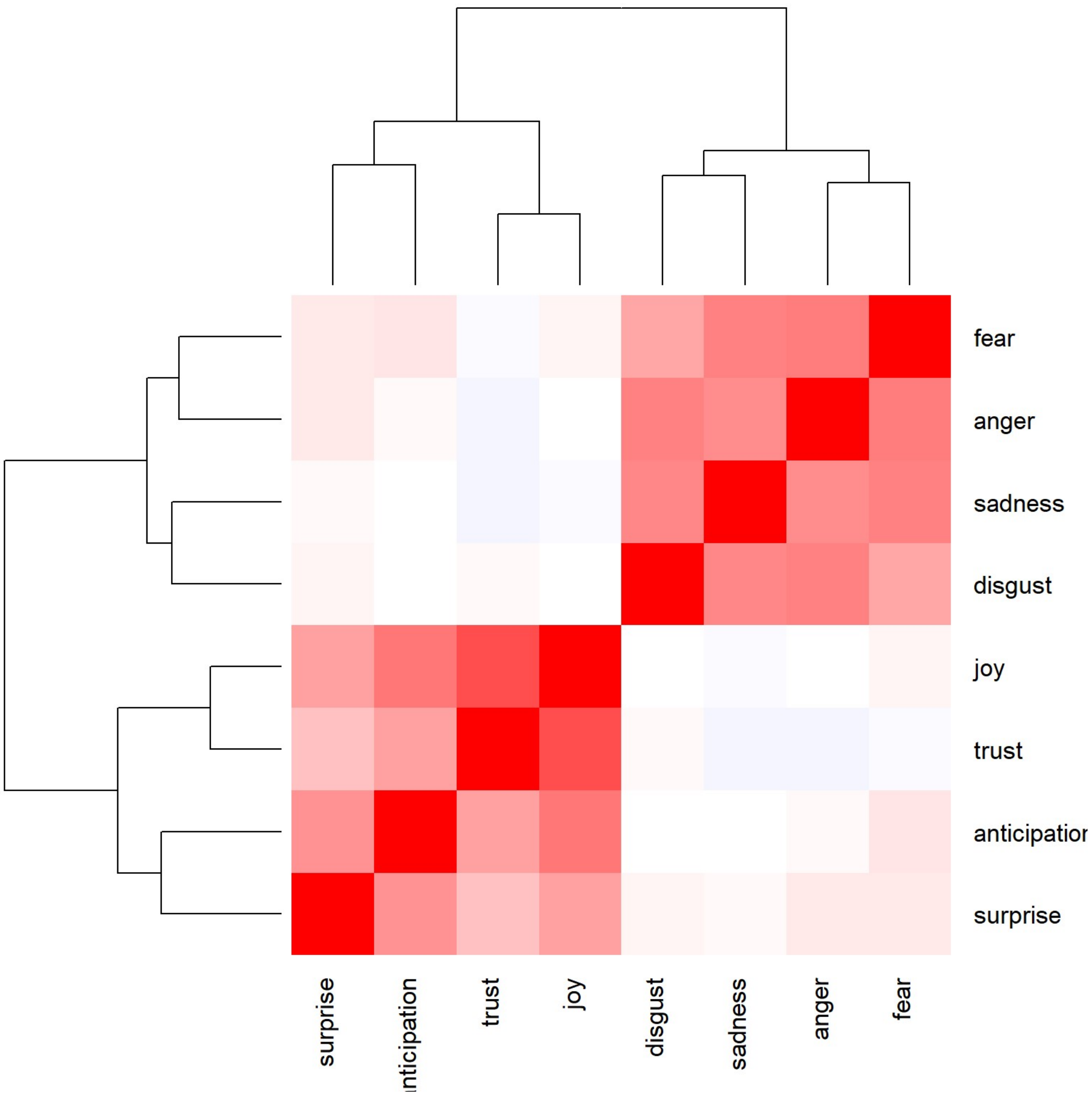


```
par(mar=c(4, 6, 2, 1))
emo.means=colMeans(select(sentence.list.dem, anger:trust)>0.01)
col.use=c("red2", "darkgoldenrod1",
            "chartreuse3", "blueviolet",
            "darkgoldenrod2", "dodgerblue3",
            "darkgoldenrod1", "darkgoldenrod1")
barplot(emo.means[order(emo.means)], las=2, col=col.use[order(emo.means)], horiz=
T, main="Democratic")
```

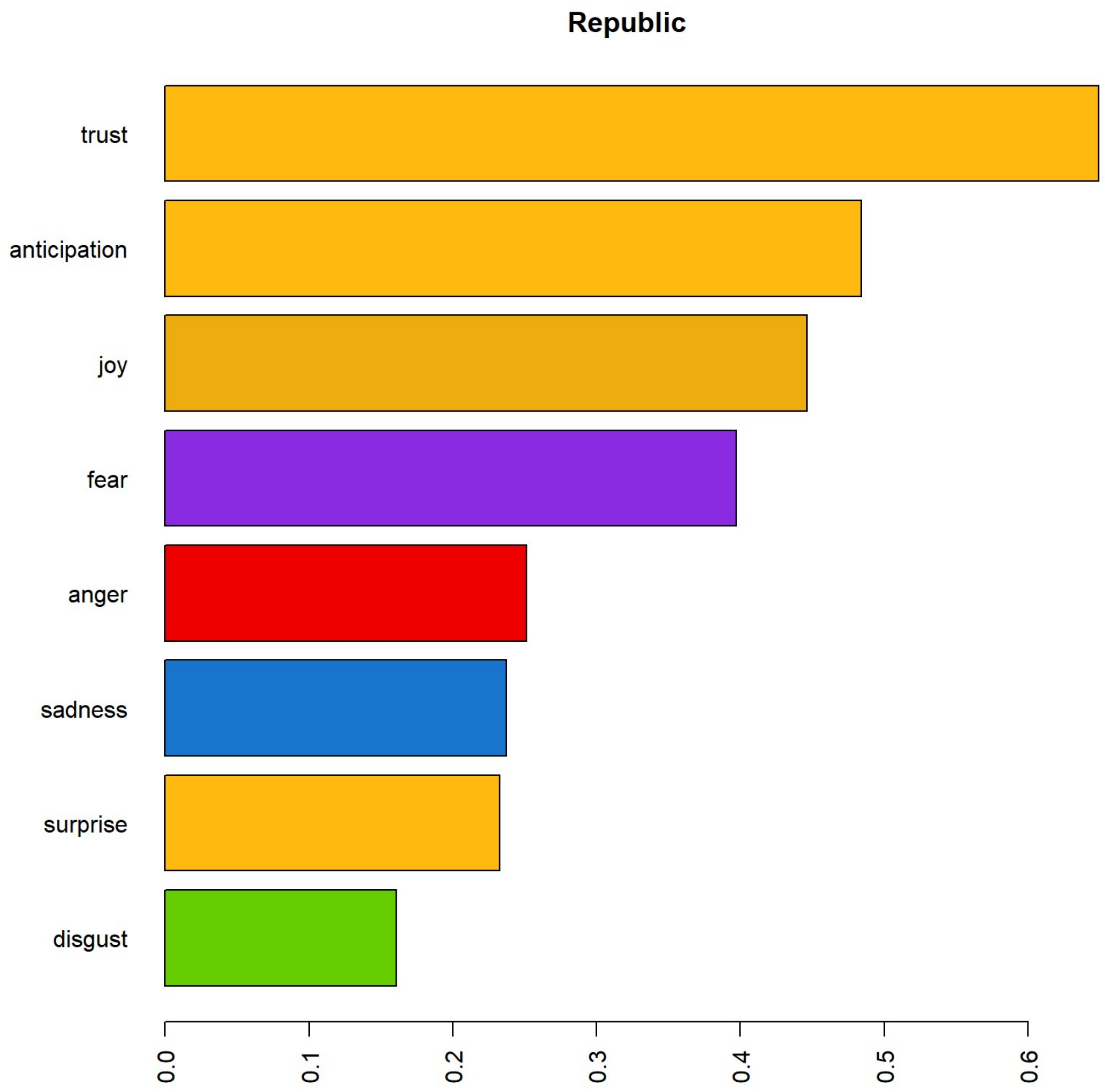


Republic

```
heatmap.2(cor(sentence.list.rep%>%select(anger:trust)),
          scale = "none",
          col = bluered(100), , margin=c(6, 6), key=F,
          trace = "none", density.info = "none")
```



```
par(mar=c(4, 6, 2, 1))
emo.means=colMeans(select(sentence.list.rep, 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="Republic")
```

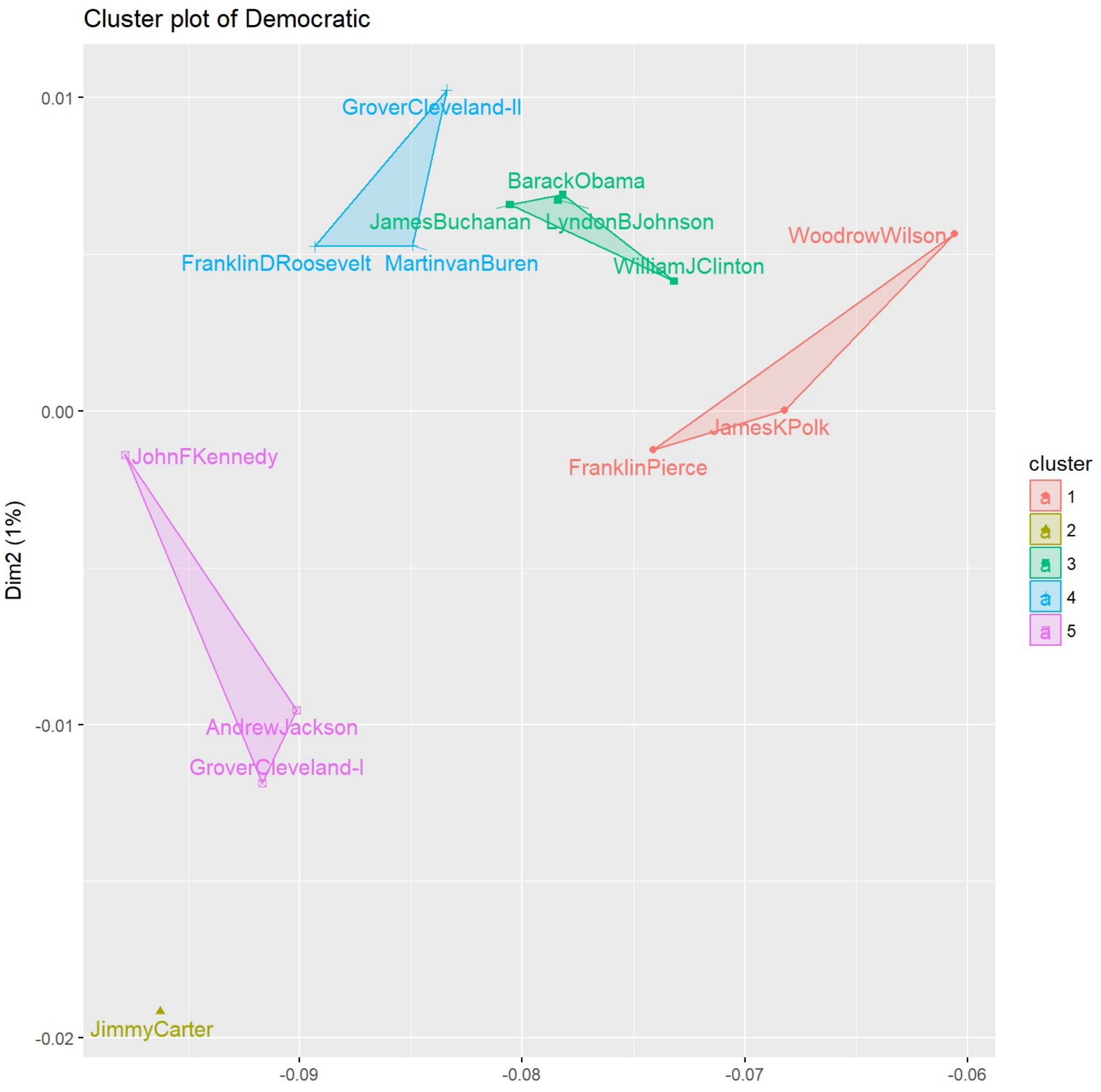


From the picture,we can see that the sort of emotions appears to be almost the same for two parties,which implies that there may be some consistency among these speeches.

Next we plot the cluster plot for both parties.

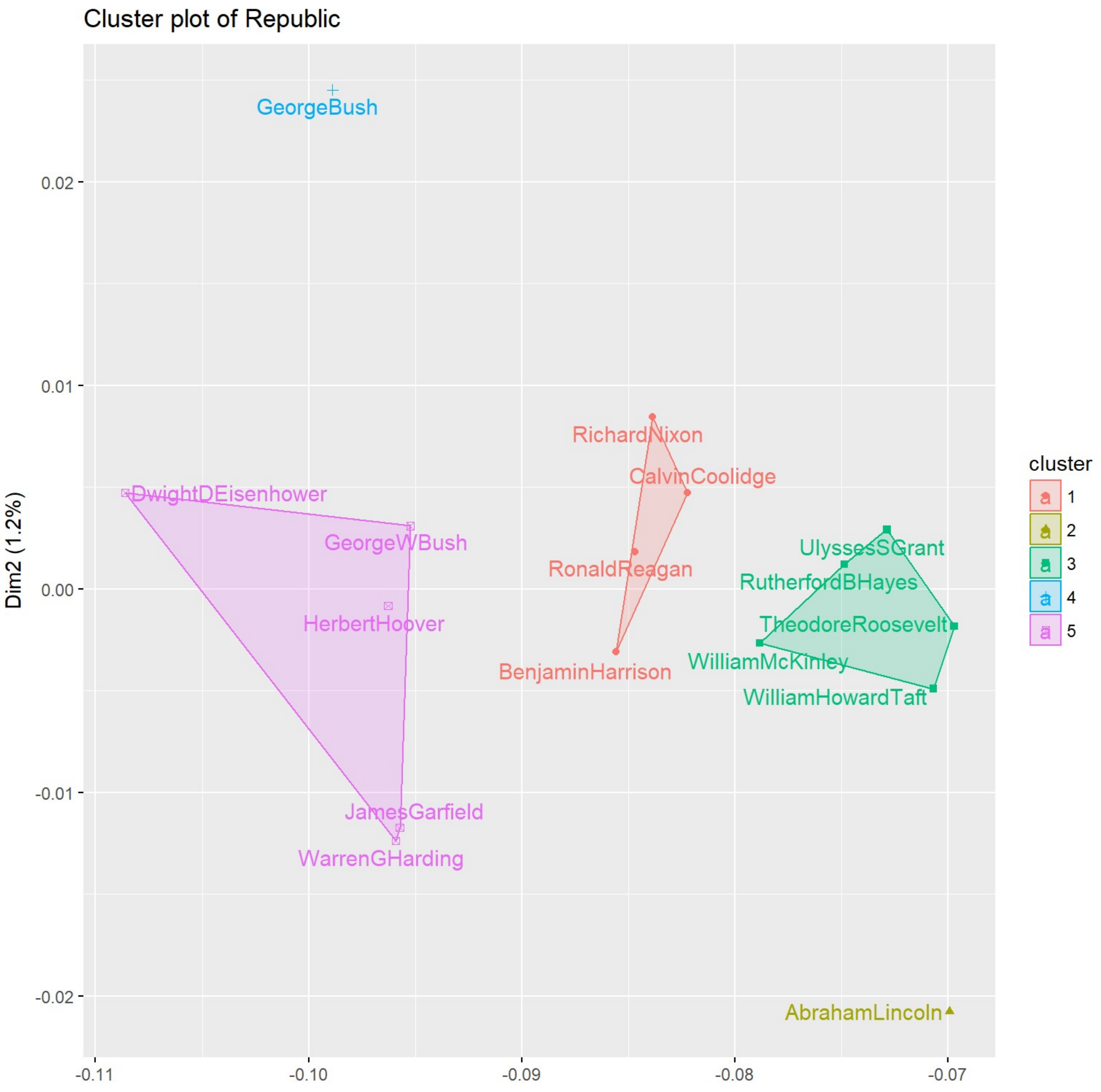
Democratic

```
presid.summary.dem=tbl_df(sentence.list.dem)%>%
  group_by(File)%>%
  summarise(
    anger=mean(anger),
    anticipation=mean(anticipation),
    disgust=mean(disgust),
    fear=mean(fear),
    joy=mean(joy),
    sadness=mean(sadness),
    surprise=mean(surprise),
    trust=mean(trust)
  )

presid.summary.dem=as.data.frame(presid.summary.dem)
rownames(presid.summary.dem)=as.character((presid.summary.dem[,1]))
km.res=kmeans(presid.summary.dem[,-1], iter.max=200,
              5)
fviz_cluster(km.res,
             stand=F, repel= TRUE,
             data = presid.summary.dem[,-1], xlab="", xaxt="n",
             show.clust.cent=FALSE,main = "Cluster plot of Democratic")
```


Cluster plot of Democratic
Dim2 (1%)
0.01
0.00
-0.01
-0.02
-0.09
-0.08
-0.07
-0.06
GroverCleveland-II
BarackObama
JamesBuchanan
LyndonBJohnson
WoodrowWilson
FranklinDRoosevelt
MartinvanBuren
WilliamJClinton
JamesKPolk
FranklinPierce
JohnFKennedy
AndrewJackson
GroverCleveland-I
JimmyCarter
cluster
1
2
3
4
5

Republic

```
presid.summary.rep=tbl_df(sentence.list.rep)%>%
  group_by(File)%>%
  summarise(
    anger=mean(anger),
    anticipation=mean(anticipation),
    disgust=mean(disgust),
    fear=mean(fear),
    joy=mean(joy),
    sadness=mean(sadness),
    surprise=mean(surprise),
    trust=mean(trust)
  )

presid.summary.rep=as.data.frame(presid.summary.rep)
rownames(presid.summary.rep)=as.character((presid.summary.rep[,1]))
km.res=kmeans(presid.summary.rep[,-1], iter.max=200,
              5)
fviz_cluster(km.res,
             stand=F, repel= TRUE,
             data = presid.summary.rep[,-1], xlab="", xaxt="n",
             show.clust.cent=FALSE,main = "Cluster plot of Republic")
```


Cluster plot of Republic
GeorgeBush
RichardNixon
CalvinCoolidge
DwightDEisenhower
GeorgeWBush
RonaldReagan
HerbertHoover
UlyssesSGrant
RutherfordBHayes
TheodoreRoosevelt
WilliamMcKinley
BenjaminHarrison
WilliamHowardTaft
JamesGarfield
WarrenGHarding
AbrahamLincoln
Dim2 (1.2%)
0.02
0.01
0.00
-0.01
-0.02
-0.11
-0.10
-0.09
-0.08
-0.07
cluster
1
2
3
4
5