Project1

jl4756

9/18/2017

Overview:

```
library("rvest")
library("qdap")
library("syuzhet")
library("dplyr")
library("beeswarm")
library("tibble")
library("sentimentr")
library("factoextra")
library("scales")
library("RColorBrewer")
library("RANN")
source("../lib/plotstacked.R")
source("../lib/speechFuncs.R")
```

Read in url.

```
### Inauguaral speeches
main.page <- read_html(x = "http://www.presidency.ucsb.edu/inaugurals.php")
inaug=f.speechlinks(main.page)
as.Date(inaug[,1], format="%B %e, %Y")
```

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

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

#### Nomination speeches
main.page=read_html("http://www.presidency.ucsb.edu/nomination.php")
nomin <- f.speechlinks(main.page)
nomin<-nomin[-47,] # remove the irrelevant line.

#### Farewell speeches
```

```
main.page=read_html("http://www.presidency.ucsb.edu/farewell_addresses.php")
farewell <- f.speechlinks(main.page)
```

Read in list.

```
inaug.list=read.csv("inauglist.csv", stringsAsFactors = FALSE)
nomin.list=read.csv("nominlist.csv", stringsAsFactors = FALSE)
farewell.list=read.csv("farewelllist.csv", stringsAsFactors = FALSE)
```

Combine list and url.

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

Write in Full Text.

```
# Loop over each row in speech.list
speech.list$fulltext=NA
for(i in seq(nrow(speech.list))) {
  text <- read_html(speech.list$urls[i]) %>% # load the page
    html_nodes(".displaytext") %>% # isloate the text
    html_text() # get the text
  speech.list$fulltext[i]=text
  # Create the file name
  filename <- paste0("../data/fulltext/",
                     speech.list$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
}
```

Write in Trump's speeches.

```
speech1=paste(readLines("../data/fulltext/SpeechDonaldTrump-NA.txt",
                  n=-1, skipNul=TRUE),
              collapse=" ")
speech2=paste(readLines("../data/fulltext/SpeechDonaldTrump-NA2.txt",
                  n=-1, skipNul=TRUE),
```

```
              collapse=" ")
speech3=paste(readLines("../data/fulltext/PressDonaldTrump-NA.txt",
                  n=-1, skipNul=TRUE),
              collapse=" ")

Trump.speeches=data.frame(
  X...President=rep("Donald J. Trump", 3),
  File=rep("DonaldJTrump", 3),
  Term=rep(0, 3),
  Party=rep("Republican", 3),
  Date=c("August 31, 2016", "September 7, 2016", "January 11, 2017"),
  Words=c(word_count(speech1), word_count(speech2), word_count(speech3)),
  Win=rep("yes", 3),
  type=rep("speeches", 3),
  links=rep(NA, 3),
  urls=rep(NA, 3),
  fulltext=c(speech1, speech2, speech3)
)
speech.list=rbind(speech.list, Trump.speeches)
```

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

Remove non-sentences.

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

Choose only "Democratic" and "Republican" party to compare.

```
newlist=sentence.list%>%filter(!is.na(Party))
newlist<-rbind(newlist[newlist$Party=="Democratic",],newlist[newlist$Party=="Republican",])
```

Compare only between Parties.

```
party.summary<-aggregate(newlist[,13:22],list(newlist$Party),mean)
party.summary=as.data.frame(party.summary)
party.summary$ratio=(party.summary$negative)/(party.summary$positive)
# Negative/Positive rate, the smaller the better
party.summary$ratio
```

```
## [1] 0.4650764 0.5031626
```

```
# Conclude: The ratio of "Democratic" is smaller than the one of "Republican",
# which means, the speech of Democratic presidents tend to use positive words.
rownames(party.summary)<-party.summary[,1]
party.summary<-party.summary[,-1]
# Create a plot to compare through different emotions
{x<-c(1:8)
plot(x,party.summary[1,1:8],type="l",col=1,xaxt = "n")
lines(x,party.summary[2,1:8],type="l",col=2)
legend("topright",c("Democratic","Republican"),lty=1,col=c("black","red"))
axis(1,at=1:8,labels=c("anger","anticipation","disgust","fear","joy","sadness","surprise",
                       "trust"))}
```


```
# Conclude: In all, Republican Presidents are more willing to use emotional words in their research.
```

Compare through president, order by party.

```
president.summary<-aggregate(newlist[,13:22],list(newlist$File,newlist$Party),mean)
president.summary=as.data.frame(president.summary)
president.summary$ratio=(president.summary$negative)/(president.summary$positive)
rownames(president.summary)<-president.summary[,1]
president.summary<-president.summary[,-1]
colnames(president.summary)[1]<-c("Party")
# Compare ratio in numbers of presidents from different parties.
compare.ratio<-president.summary[order(president.summary$ratio),]
table(compare.ratio$Party[compare.ratio$ratio<0.5])
```

```
##
## Democratic Republican
##         14         13
```

```
# Compare positive word rates in numbers of presidents from different parties.
compare.pos<-president.summary[order(president.summary$positive),]
table(compare.pos$Party[compare.pos$positive>0.075])
```

```
##
## Democratic Republican
##         12         14
```

```
# Conclude: There isn't too much difference between the number of presidents who
# likes to use positive words.
head(compare.ratio)
```

```
##                       Party       anger anticipation     disgust
## MittRomney        Republican 0.010108867   0.02881149 0.005445982
## MichaelDukakis    Democratic 0.008360423   0.02910973 0.007290123
## ThomasEDewey      Republican 0.010546652   0.03645069 0.005603096
## GroverCleveland-I Democratic 0.011344617   0.03547421 0.007977169
## LyndonBJohnson    Democratic 0.011397235   0.03244827 0.009296228
## DwightDEisenhower Republican 0.017754288   0.03826237 0.007085148
##                         fear        joy     sadness    surprise      trust
## MittRomney        0.01182217 0.03507623 0.009398545 0.010073210 0.05916725
## MichaelDukakis    0.01250071 0.03536254 0.007380922 0.014317782 0.05354074
## ThomasEDewey      0.02666675 0.04084009 0.013935480 0.014018369 0.06632922
## GroverCleveland-I 0.02623541 0.03135520 0.009625197 0.011077800 0.07245122
## LyndonBJohnson    0.01814459 0.02789763 0.010706120 0.009720445 0.05496985
## DwightDEisenhower 0.02800416 0.03568016 0.013302523 0.012677093 0.05883946
##                     negative   positive     ratio
## MittRomney        0.01710144 0.07797955 0.2193067
## MichaelDukakis    0.02094632 0.08803742 0.2379252
## ThomasEDewey      0.03217418 0.09330396 0.3448318
## GroverCleveland-I 0.03584374 0.09900025 0.3620571
## LyndonBJohnson    0.02735448 0.07287860 0.3753431
## DwightDEisenhower 0.03427288 0.08850180 0.3872564
```

```
tail(compare.ratio)
```

```
##                       Party      anger anticipation     disgust
## WilliamHowardTaft Republican 0.01484824   0.02475088 0.010755806
## RobertDole        Republican 0.01693974   0.02925358 0.008943524
## CharlesEHughes    Republican 0.01915577   0.03311182 0.010375591
```

```
## WarrenGHarding    Republican 0.01919575   0.03487362 0.011578204
## JohnMcCain        Republican 0.01953016   0.02749144 0.008639174
## AbrahamLincoln    Republican 0.02214610   0.02538016 0.010713349
##                         fear        joy    sadness    surprise      trust
## WilliamHowardTaft 0.02446268 0.01936605 0.01407274 0.009177716 0.04933131
## RobertDole        0.02001206 0.02187510 0.01722532 0.012102994 0.04439208
## CharlesEHughes    0.02540871 0.01968673 0.01778762 0.010495798 0.05615888
## WarrenGHarding    0.03682913 0.02908338 0.01871388 0.013631688 0.05740112
## JohnMcCain        0.03245849 0.02672097 0.01671050 0.007914201 0.04521050
## AbrahamLincoln    0.03262857 0.02011536 0.02021784 0.010240371 0.04327443
##                     negative   positive     ratio
## WilliamHowardTaft 0.03917012 0.06559809 0.5971229
## RobertDole        0.03477580 0.05729055 0.6070077
## CharlesEHughes    0.05210225 0.08534035 0.6105231
## WarrenGHarding    0.05573557 0.08649557 0.6443748
## JohnMcCain        0.04679076 0.06423233 0.7284613
## AbrahamLincoln    0.04717292 0.06228920 0.7573210
```

```
# Conclude: It is not surprising that President Lincoln is the least likely to use
# positive words in his speeches, this is mainly because the US was in "Civil War"
# at that time.
```

Compare between Terms.

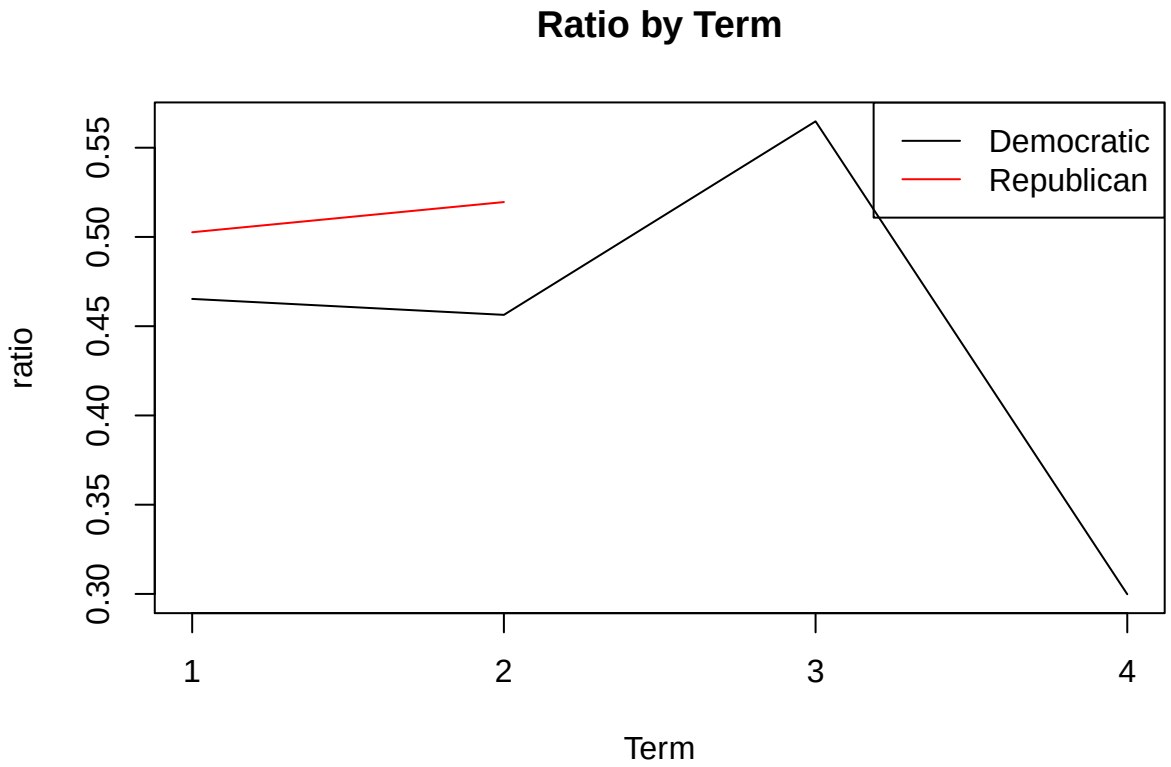
```
term.summary<-aggregate(newlist[,13:22],list(newlist$Term),mean)
term.summary=as.data.frame(term.summary)
term.summary$ratio=(term.summary$negative)/(term.summary$positive)
rownames(term.summary)<-term.summary[,1]
term.summary<-term.summary[,-1]
term.summary$ratio
```

```
## [1] 0.5365998 0.4869489 0.4946730 0.5647641 0.2998310
```

```
# Conclude: As we can see, presidents tend to use more possitive words when
# they become a president.
```

Compare between Terms an Parties.

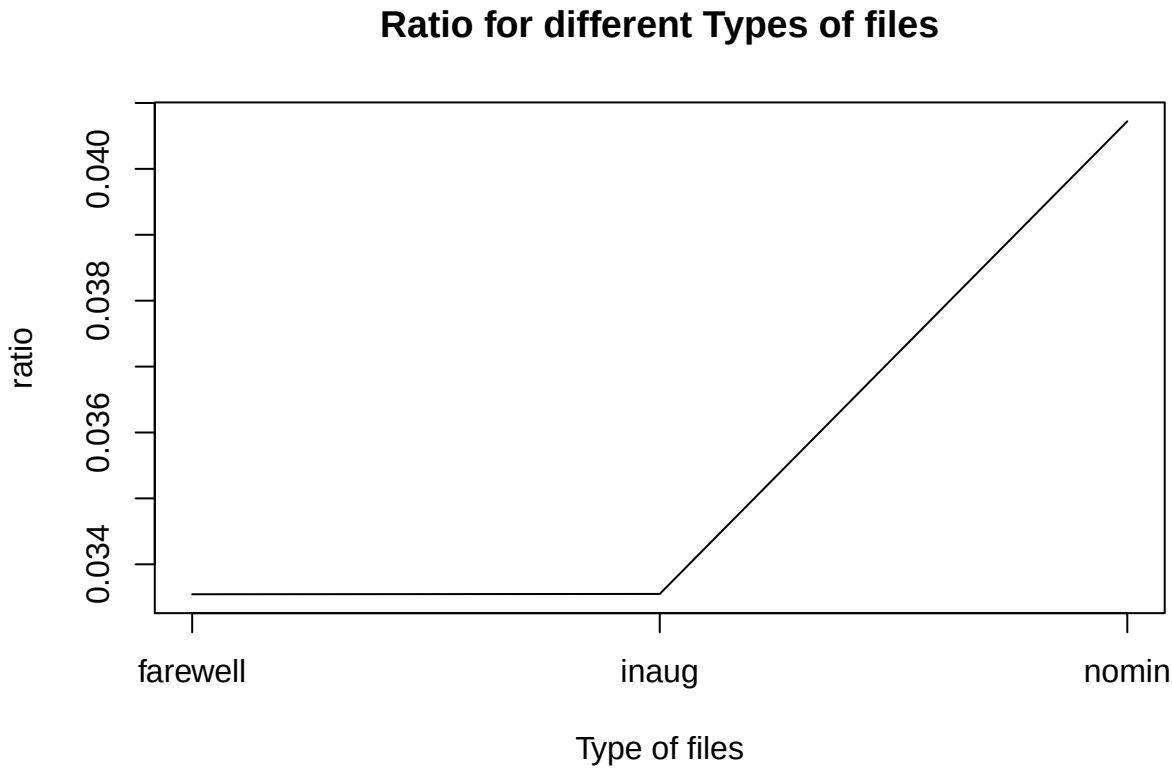
```
tp.summary<-aggregate(newlist[,13:22],list(newlist$Term,newlist$Party),mean)
tp.summary=as.data.frame(tp.summary)
tp.summary$ratio=(tp.summary$negative)/(tp.summary$positive)
colnames(tp.summary)[1:2]<-c("Term","Party")

{x<-c(1:4)
plot(x,tp.summary[1:4,13],type="l",col=1,xlab="Term",ylab="ratio",
     main="Ratio by Term",xaxt = "n")
lines(tp.summary[6:7,13],type="l",col=2)
axis(1,at=1:4,labels=c(1,2,3,4))
legend("topright",c("Democratic","Republican"),lty=1,col=c("black","red"))
}
```




```
# Conclude: We can see Democratic presidents tend to use more positive words from their
# Term1 to Term2, except for president FranklinDRoosevelt, he used more negative words
# during his 3rd Term because the USA is in WWII, and he used more positive words in his
# 4th Term because he wants to inspiring citizen.
# However we can see that Republican President tends to use more negative words in their
# 2nd Term.
```

Compare between Types of files.

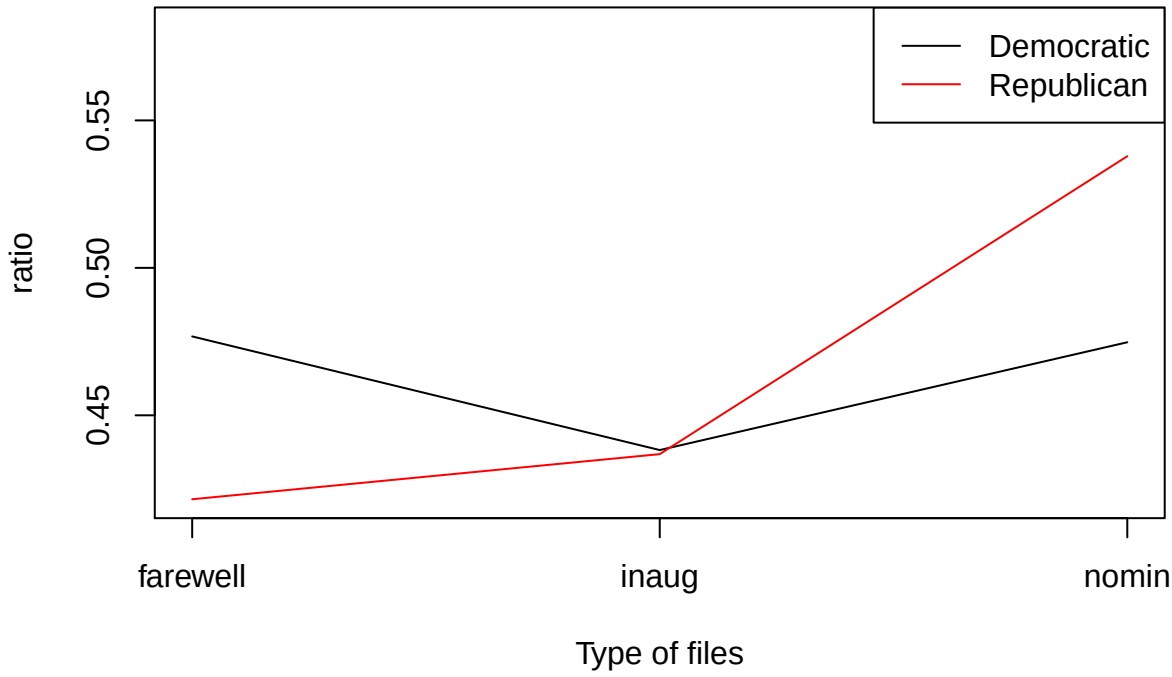
```
type.summary<-aggregate(newlist[,13:22],list(newlist$type),mean)
type.summary=as.data.frame(type.summary)
type.summary$ratio=(type.summary$negative)/(type.summary$positive)
rownames(type.summary)<-type.summary[,1]
type.summary<-type.summary[,-1]
{x<-c(1:3)
plot(x,tp.summary[1:3,11],type="l",col=1,xlab="Type of files",ylab="ratio",
     xaxt = "n", main="Ratio for different Types of files")
axis(1,at=1:3,labels=rownames(type.summary)[1:3])
}
```




```
# Conclude: We can see that there is a big decrease of negative words from nomin to inaug,
# I suppose this may because of pointing out the dark side of America's current situation,
# and after inaug, in order to make citizen believe that his policy is actually making some progress in
```

Compare between Types of files and Parties.

```
typ.summary<-aggregate(newlist[,13:22],list(newlist$type,newlist$Party),mean)
typ.summary=as.data.frame(typ.summary)
typ.summary$ratio=(typ.summary$negative)/(typ.summary$positive)
colnames(typ.summary)[1:2]<-c("Type","Party")
{x<-c(1:3)
plot(x,typ.summary[1:3,13],type="l",col=1,xlab="Type of files",ylab="ratio",
     xaxt = "n",ylim=c(min(typ.summary[,13]),max(typ.summary[,13])),
     main="Ratio for different Types of files and Parties ")
lines(typ.summary[4:6,13],type="l",col=2)
axis(1,at=1:3,labels=c("farewell","inaug","nomin"))
legend("topright",c("Democratic","Republican"),lty=1,col=c("black","red"))
}
```




```
# Conclude: "Democratic" president tends more to use negative words in their farewell and nomin speeche
```

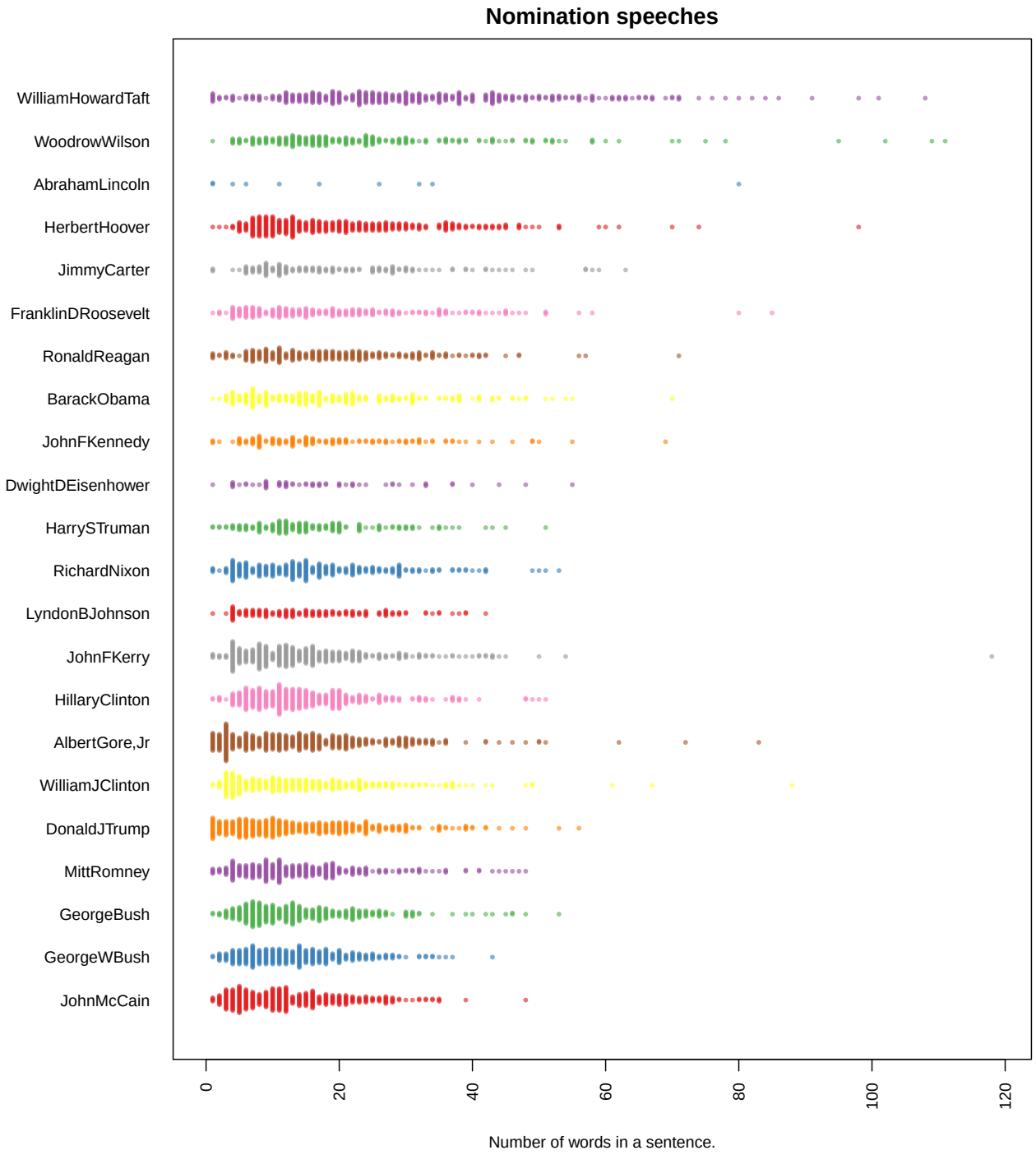
Choose those that are famous presidents.

```
sel.comparison=c("DonaldJTrump","JohnMcCain", "GeorgeBush", "MittRomney", "GeorgeWBush",
                 "RonaldReagan","AlbertGore,Jr", "HillaryClinton","JohnFKerry",
                 "WilliamJClinton","HarrySTruman", "BarackObama", "LyndonBJohnson",
                 "GeraldRFord", "JimmyCarter", "DwightDEisenhower", "FranklinDRoosevelt",
                 "HerbertHoover","JohnFKennedy","RichardNixon","WoodrowWilson",
                 "AbrahamLincoln", "TheodoreRoosevelt", "JamesGarfield",
                 "JohnQuincyAdams", "UlyssesSGrant", "ThomasJefferson",
                 "GeorgeWashington", "WilliamHowardTaft", "AndrewJackson",
                 "WilliamHenryHarrison", "JohnAdams")
```

First term.

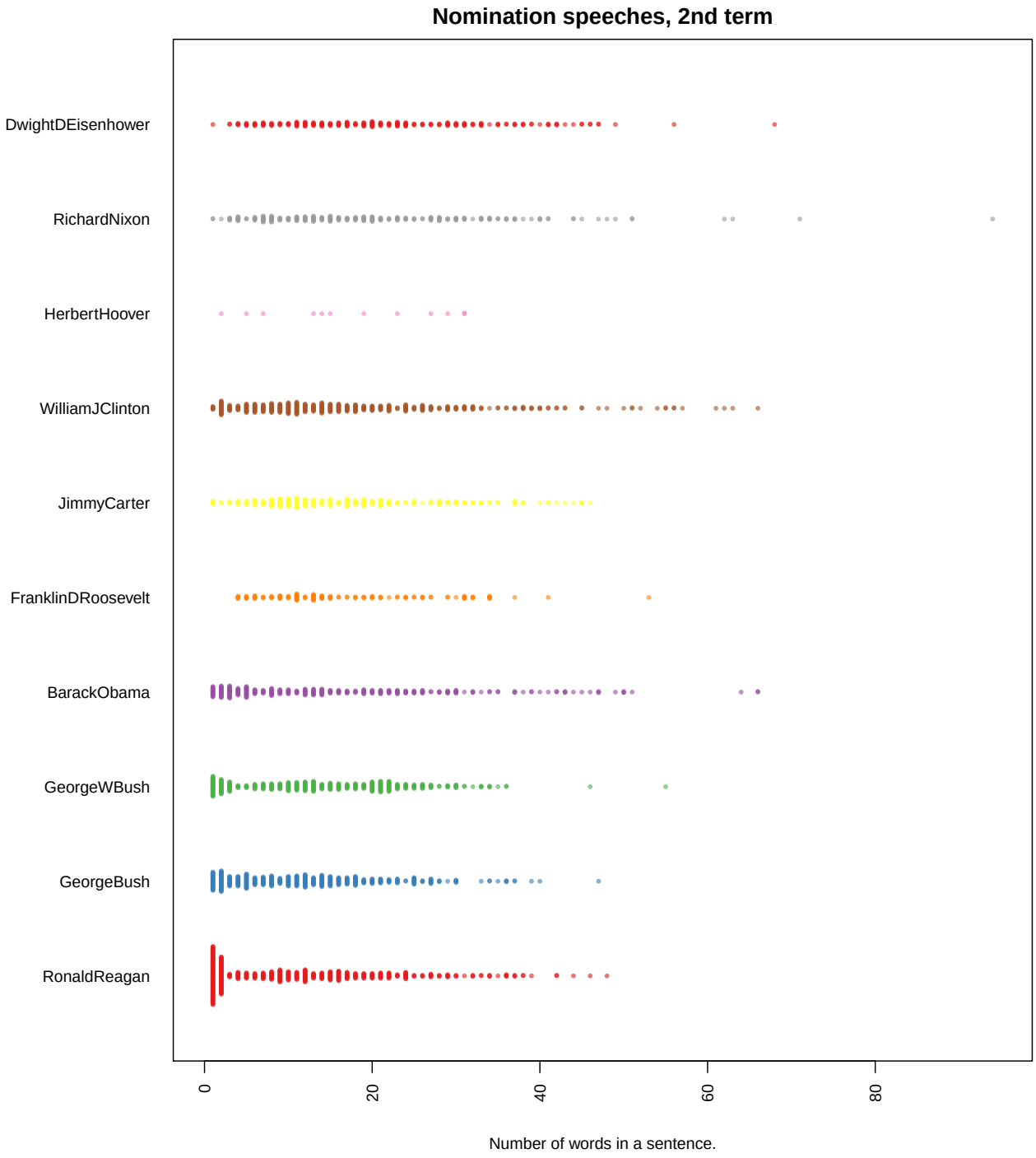
```
par(mar=c(4, 11, 2, 2))
#sel.comparison=levels(sentence.list$FileOrdered)
sentence.list.sel=filter(sentence.list,
                         type=="nomin", Term==1, File%in%sel.comparison)
sentence.list.sel$File=factor(sentence.list.sel$File)

sentence.list.sel$FileOrdered=reorder(sentence.list.sel$File,
                                      sentence.list.sel$word.count,
                                      mean,
```

```
                                    order=T)
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, xlab="Number of words in a sentence.", ylab="",
         main="Nomination speeches")
```



Second term.

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

#sel.comparison=levels(sentence.list$FileOrdered)
sentence.list.sel=filter(sentence.list,
                         type=="nomin", Term==2, File%in%sel.comparison)
sentence.list.sel$File=factor(sentence.list.sel$File)

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

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=1.2/nlevels(sentence.list.sel$FileOrdered),
         las=2, xlab="Number of words in a sentence.", ylab="",
         main="Nomination speeches, 2nd term")
```



Find the longest length of a word in each sentence.

```
word<-matrix(0,nrow=21326,ncol=max(sentence.list$word.count)+1)
for (i in 1:21326){
  word[i,1:length(nchar(strsplit(as.character(
    sentence.list$sentences[i]),split="\\, |\\,| |\\:|\\-|\\>")[[1]])))]<-nchar(strsplit(
                  as.character(sentence.list$sentences[i]),split="\\, |\\,| |\\:|\\-|\\>")[[1]])
```

```
}
word[,1:124]<-as.numeric(word[,1:124])
rownames(word)<-sentence.list$File

# Find the longest word in each sentence.
maxlength.word<-matrix(NA,nrow=nrow(word),ncol=2)
colnames(maxlength.word)<-c("President","max length of word")
maxlength.word[,1]<-sentence.list$File
president<-as.matrix(as.data.frame(table(sentence.list$File))[,1])
maxlength.word[,2]<-apply(word[,1:124],1,max)
table(as.numeric(maxlength.word[,2]))
```

```
##
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
##  169  105   91  365  711  924 1915 2461 3182 3917 2818 2134 1213  990  181
##   16   17   18   19
##  117   26    6    1
```

```
# Find the number of complex words in each president's speeches.
complexword<-matrix(0,ncol=2,nrow=length(president))
complexword[,1]<-as.matrix(president)
# Assume that a word with length larger than 9 is defined as a complex word.
for(i in 1:length(president)){
  complexword[i,2]<-sum(as.numeric(word[rownames(word)==president[i],1:124])>=11)
}
complexword.ordered<-complexword[order(as.numeric(complexword[,2])),]
head(complexword.ordered,10)
```

```
##       [,1]                [,2]
##  [1,] "TheodoreRoosevelt" "25"
##  [2,] "WalterFMondale"    "28"
##  [3,] "GeorgeMcGovern"    "33"
##  [4,] "JohnFKennedy"      "49"
##  [5,] "CharlesEHughes"    "52"
##  [6,] "MichaelDukakis"    "52"
##  [7,] "MittRomney"        "52"
##  [8,] "ZacharyTaylor"     "55"
##  [9,] "HubertHHumphrey"   "62"
## [10,] "JohnMcCain"        "66"
```

```
tail(complexword.ordered,10)
```

```
##       [,1]                   [,2]
## [49,] "WilliamHenryHarrison" "371"
## [50,] "AlSmith"              "380"
## [51,] "GeorgeWashington"     "387"
## [52,] "AndrewJackson"        "398"
## [53,] "DonaldJTrump"         "452"
## [54,] "FranklinDRoosevelt"   "478"
## [55,] "HerbertHoover"        "554"
## [56,] "WarrenGHarding"       "561"
## [57,] "BenjaminHarrison"     "625"
## [58,] "WilliamHowardTaft"    "808"
```

```
# Conclude: Is it said that number of complex words in a sentence reflect the
# complexity of a sentence, further reflect the literature attainment of the speaker.
```

```
# Surprisingly, President Trump is the highest 10 presidents of using complex words.
```