The Secret to Greatness: A Look Into Historians' Presidential Rankings

Wyatt Thompson

September 18, 2017

Abe Lincoln—a jedi of rhetoric

Introduction

When we read the history books, certain presidents stand out as more important than others: Abe Lincoln, George Washington, Franklin Roosevelt, John F Kennedy, Barack Obama. But what makes a great president so great? The state of the economy, the policies they passed, their leadership style, undoubtedly play a role. But when the job description includes addressing the public on a regular basis, one would expect that a president's speaking style and rhetoric certainly matter as well! In this project, I explore the the text from presidential

inaugeration speeches with an eye for sentence length and the breakdown of their sentiment. By pairing this with the C-SPAN historical ranking of presidents, I hope to gain some insight into what kind of a speaker the historians tend to find appealing.

The main goal of this project is to explore the text mining packages in R and survey some methods of text analysis. However, maybe we'll find that the best presidents were longwinded optimists, or perhaps punctuated pessimists!

A note here about the assumptions of this analysis. First, I'm assuming that the sentiment of the inaugeration speech captures a general sentiment of the speaker, and I'm leaving out the obviously important variables of their impact on society.

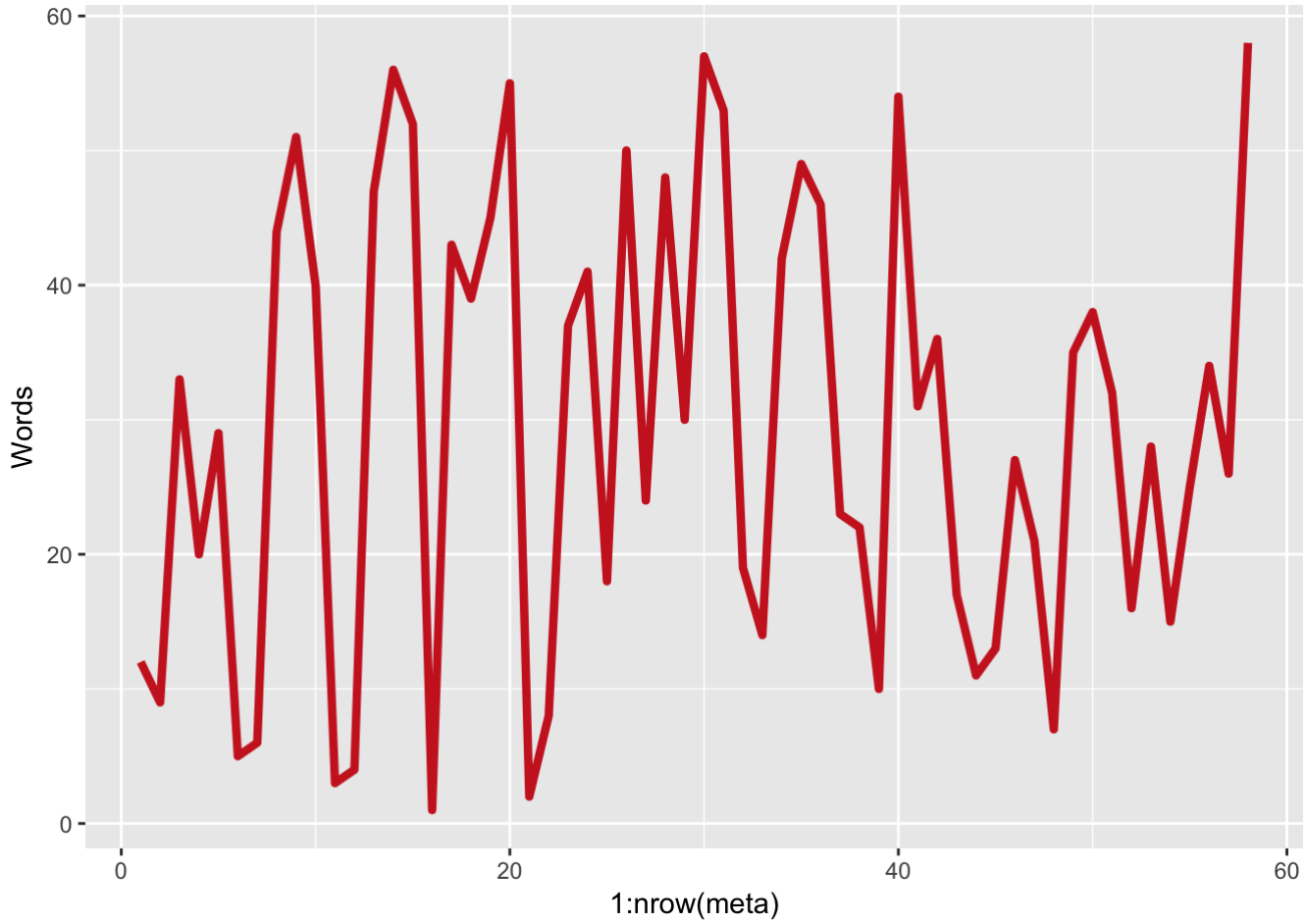
Load Packages

First, let's load in the right packages and start reading in some data!

Format and Import Data

```
ranking<-read.csv("../data/CSPANRanking.csv",header = F)
dates<-read.table("../data/InauguationDates.txt", sep="\t", header=TRUE)
#dates$PRESIDENT<-gsub(" ", "", dates$PRESIDENT)
meta<-read.csv("../data/InaugurationInfo.csv")
files<- list.files(path="../data/InauguralSpeeches")
speeches<-trimws(files)
speeches<- paste("../data/InauguralSpeeches/",speeches, sep="")
speechlist<- lapply(speeches, read_file)
speechlist<-as.vector(speechlist)
```

```
meta$Words<-as.numeric(meta$Words)
ggplot(meta, aes(x=1:nrow(meta), y= Words))+geom_line(colour="firebrick3", size=1.5)
```




```
  xlab("President Number") + theme_bw()
```

```
## NULL
```

```
  ylab("Total Number of Words in Speech")
```

```
## $y
## [1] "Total Number of Words in Speech"
##
## attr(,"class")
## [1] "labels"
```

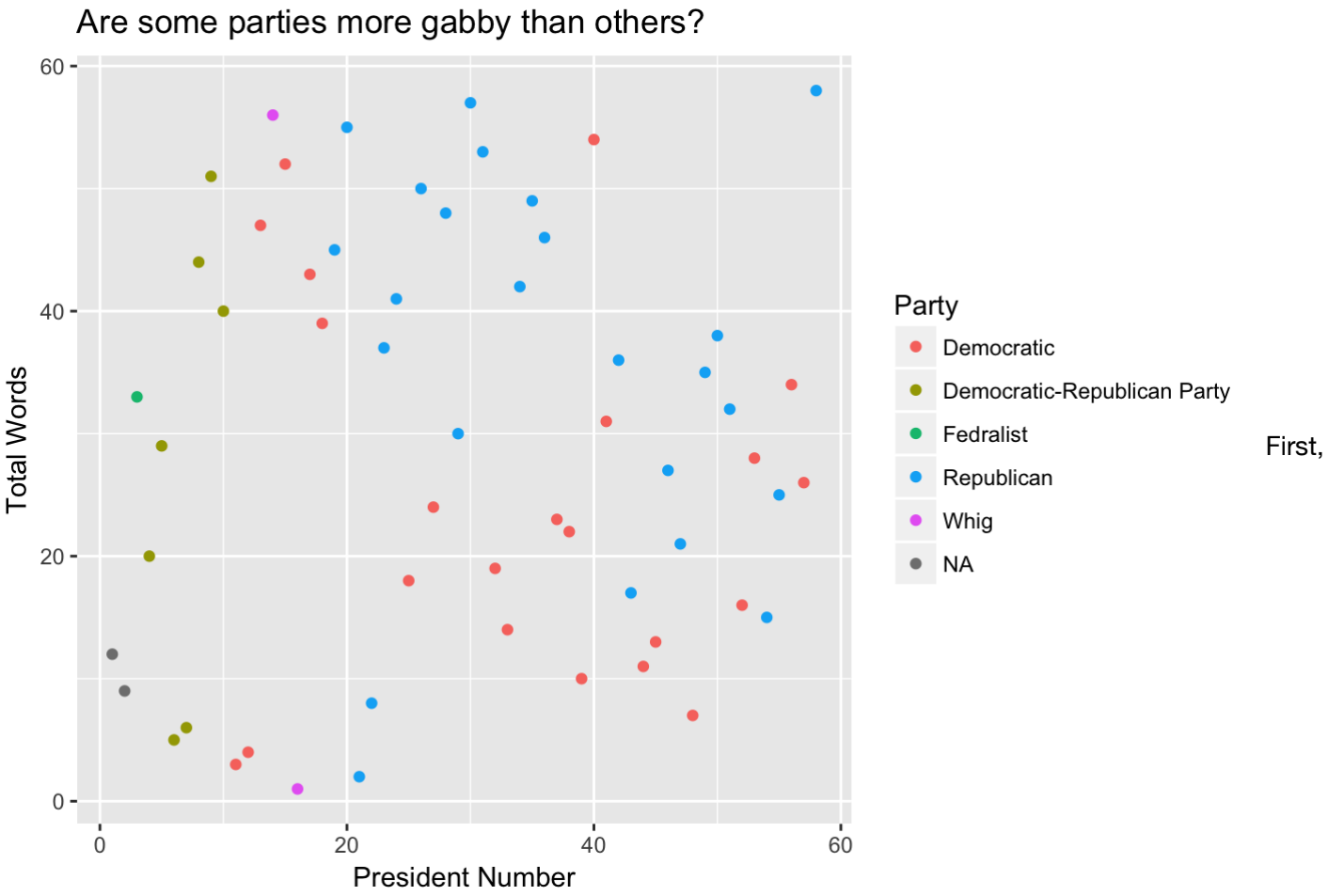
```
paste("The second time speakers give speeches averaging,
",mean(meta$Words[meta$Term==2])," words")
```

```
## [1] "The second time speakers give speeches averaging,  23.9411764705882  words"
```

```
paste("But the first time speakers give speeches averaging, ",mean(meta$Words[meta$Term=
=1])," words")
```

```
## [1] "But the first time speakers give speeches averaging,  31.7948717948718  words"
```

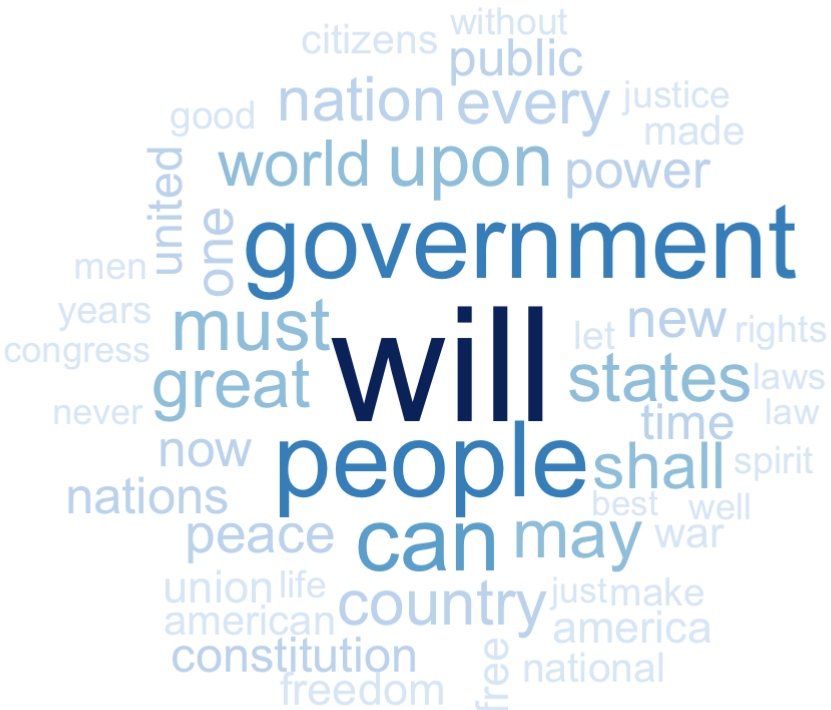
```
ggplot(meta,aes(x=1:nrow(meta), y= Words))+geom_point(aes(col=Party))+ylab("Total
Words")+xlab("President Number")+ggtitle("Are some parties more gabby than others?")
```



First, we should see if president's are using a different volume of words across time. Contrary to what I expected, there's not a trend of longer speeches. But we do note some different behaviors among the presidents. The Republicans appear to give longer speeches, and the second time presidents give shorter inaugaration addresses. Let's take a look at what words the presidents used the most often.

In this chunk of code, I take the speeches and put them into a Document-Term Matrix so we can see what words the presidents are using most frequently. I then create a wordcloud across all presidential speeches to visualize the relative frequency of words and get a handle on what this text mining analysis will look like.

```
files<-gsub("inaug","",files)
files<-gsub(".txt","",files)
df<-data.frame(
  pres=speeches
)

df$text<-speechlist
df$text<-as.character(df$text)
speech.corp<-Corpus(VectorSource(df$text))
speech.corp<-tm_map(speech.corp,stripWhitespace)
speech.corp<-tm_map(speech.corp,content_transformer(tolower))
speech.corp<-tm_map(speech.corp,removeWords, stopwords("english"))
speech.corp<-tm_map(speech.corp,removePunctuation)
speech.corp<-tm_map(speech.corp, removeWords, character(0))
dtm<-TermDocumentMatrix(speech.corp)
dtm.tidy<-tidy(dtm)
dtm.overall<-summarise(group_by(dtm.tidy,term), sum(count))

wordcloud(dtm.overall$term,dtm.overall$`sum(count)`,
          scale=c(5,.5),
          max.words=50,
          min.freq=1,
          random.order=FALSE,
          colors=brewer.pal(9,"Blues"))
```

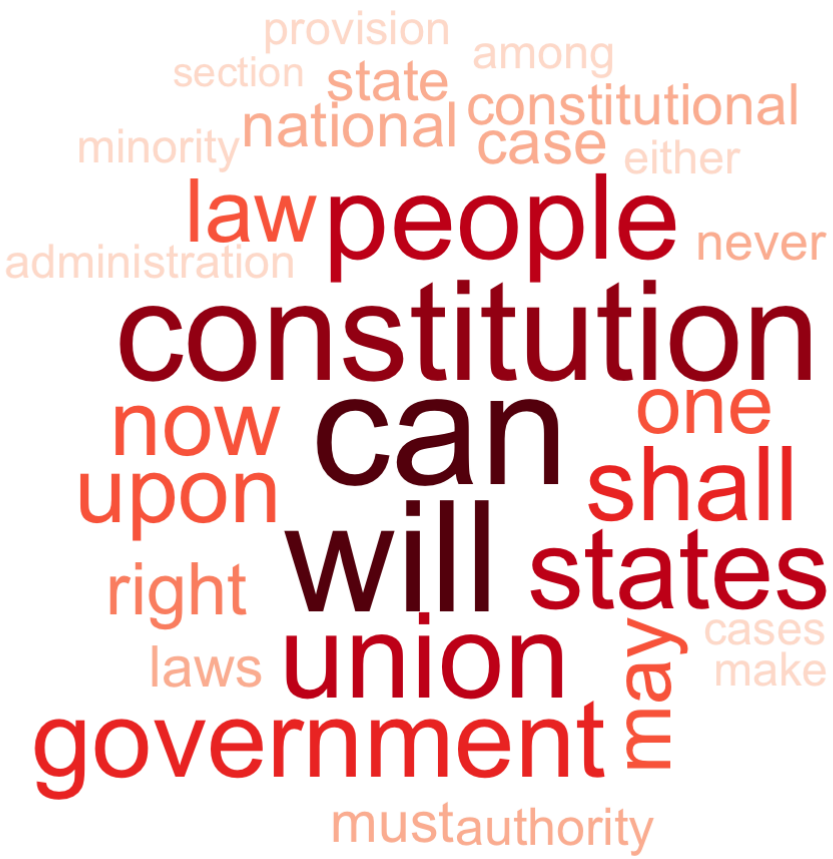


```
# for(i in length(df$pres)){
#   temp<-df$text[i]
#   speech.list[[i]]<-(VectorSource(temp))
#   corp.list[[i]]<-Corpus(speech.list[[i]])
# }
```

Not surprisingly, the presidents are using the word "will" a lot. The speeches we are analyzing are the start of their term in office, so this makes sense. Let's take a look at Abe Lincoln and George W. Bush to see if there are any glaring differences in word choice from century to century.

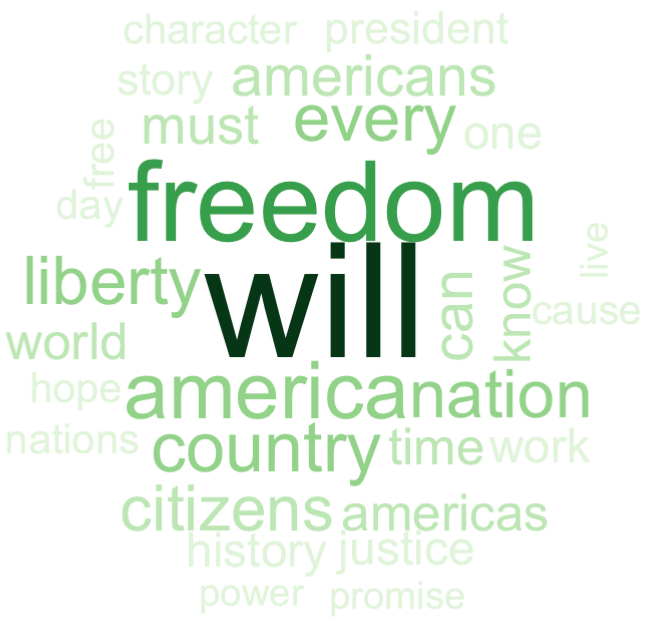
Word Frequency

```
dtm.Abraham<-summarise(group_by(dtm.tidy[dtm.tidy$document ==1,],term), sum(count))
wordcloud(dtm.Abraham$term,dtm.Abraham$`sum(count)`,
          scale=c(5,.5),
          max.words=30,
          min.freq=1,
          random.order=FALSE,
          colors=brewer.pal(9,"Reds"))
```



Honest Abe is concerned with the ability of the government to solve social problems, and the power of the consitution. We see that as compared with the overall word cloud, Abe emphasizes the capacity for change with the frequent use of the word "can". What about George Jr?

```
dtm.GeorgeBush<-summarise(group_by(dtm.tidy[dtm.tidy$document %in% c(20,21),],term),
sum(count))
wordcloud(dtm.GeorgeBush$term,dtm.GeorgeBush$`sum(count)`,
          scale=c(5,.5),
          max.words=30,
          min.freq=1,
          random.order=FALSE,
          colors=brewer.pal(9,"Greens"))
```



George is a big fan of "Will", like the other presidents, but true to his Republic party, he uses the word "freedom" as much as he can.

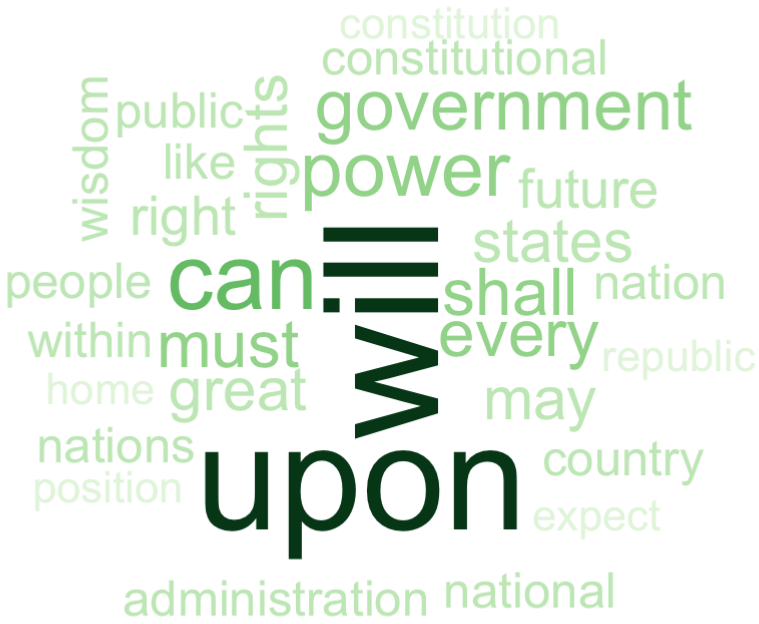
Let's see if the historian's favorite and least favorite presidents wordclouds give us any insight. First, the 3 least favorite full-term presidents. Let's look at Buchanan, Pierce, and Harding.

```
colnames(ranking)<-c("President", "Ranking")
ranking[c(1:4,40:43),]
```

```
##                President Ranking
## 1         Abraham Lincoln       1
## 2       George Washington       2
## 3  Franklin D. Roosevelt       3
## 4      Theodore Roosevelt       4
## 40      Warren G. Harding      40
## 41        Franklin Pierce      41
## 42         Andrew Johnson      42
## 43         James Buchanan      43
```

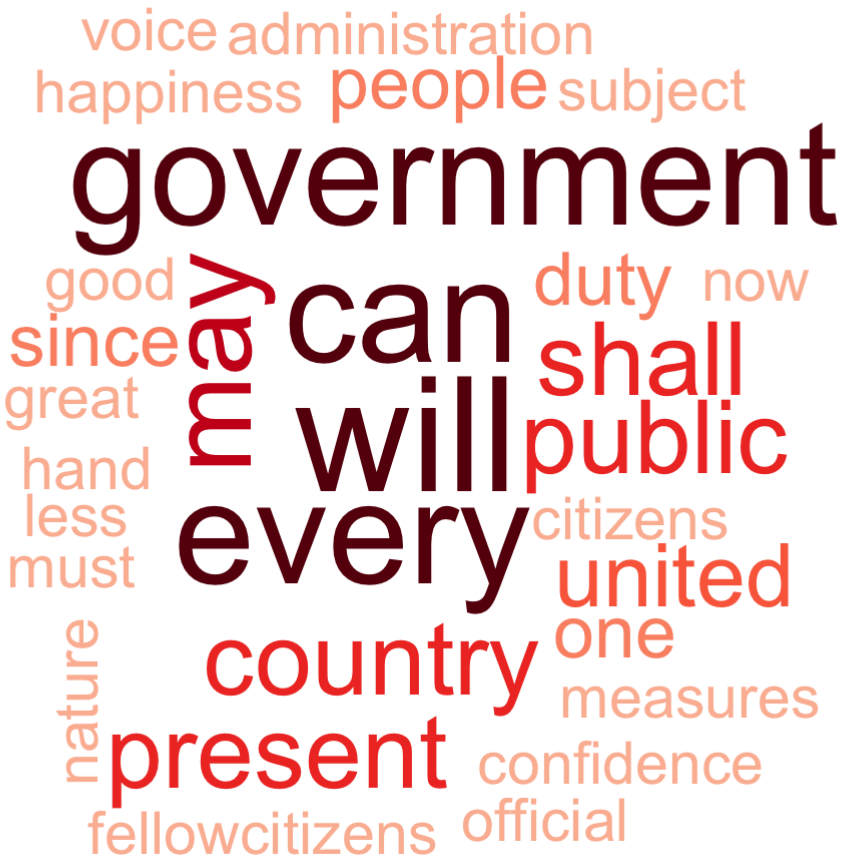
```
dtm.Buchanan<-summarise(group_by(dtm.tidy[dtm.tidy$document ==26,],term), sum(count))
wordcloud(dtm.Buchanan$term,dtm.Buchanan$`sum(count)`,
          scale=c(5,.5),
          max.words=30,
          min.freq=1,
          random.order=FALSE,
          colors=brewer.pal(9,"Blues"))
```

```
dtm.Pierce<-summarise(group_by(dtm.tidy[dtm.tidy$document ==16,],term), sum(count))
wordcloud(dtm.Pierce$term,dtm.Pierce$`sum(count)`,
          scale=c(5,.5),
          max.words=30,
          min.freq=1,
          random.order=FALSE,
          colors=brewer.pal(9,"Greens"))
```




```
dtm.Harding<-summarise(group_by(dtm.tidy[dtm.tidy$document ==49,],term), sum(count))
wordcloud(dtm.Harding$term,dtm.Harding$`sum(count)`,
          scale=c(5,.5),
          max.words=30,
          min.freq=1,
          random.order=FALSE,
          colors=brewer.pal(9,"Blues"))
```


```
dtm.Washington<-summarise(group_by(dtm.tidy[dtm.tidy$document %in% c(18,19),],term),
sum(count))
wordcloud(dtm.Washington$term,dtm.Washington$`sum(count)`,
          scale=c(5,.5),
          max.words=30,
          min.freq=1,
          random.order=FALSE,
          colors=brewer.pal(9,"Reds"))
```




```
dtm.FRoos<-summarise(group_by(dtm.tidy[dtm.tidy$document %in% c(12,13,14,15),],term), su
m(count))
wordcloud(dtm.FRoos$term,dtm.FRoos$`sum(count)`,
          scale=c(5,.5),
          max.words=30,
          min.freq=1,
          random.order=FALSE,
          colors=brewer.pal(9,"Blues"))
```


```
dtm.TRoos<-summarise(group_by(dtm.tidy[dtm.tidy$document ==44,],term), sum(count))
wordcloud(dtm.TRoos$term,dtm.TRoos$`sum(count)`,
          scale=c(5,.5),
          max.words=30,
          min.freq=1,
          random.order=FALSE,
          colors=brewer.pal(9,"Reds"))
```

```
## Warning in wordcloud(dtm.TRoos$term, dtm.TRoos$`sum(count)`, scale = c(5, :
## conditions could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(dtm.TRoos$term, dtm.TRoos$`sum(count)`, scale = c(5, :
## regards could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(dtm.TRoos$term, dtm.TRoos$`sum(count)`, scale = c(5, :
## relations could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(dtm.TRoos$term, dtm.TRoos$`sum(count)`, scale = c(5, :
## responsibility could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(dtm.TRoos$term, dtm.TRoos$`sum(count)`, scale = c(5, :
## spirit could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(dtm.TRoos$term, dtm.TRoos$`sum(count)`, scale = c(5, :
## success could not be fit on page. It will not be plotted.
```

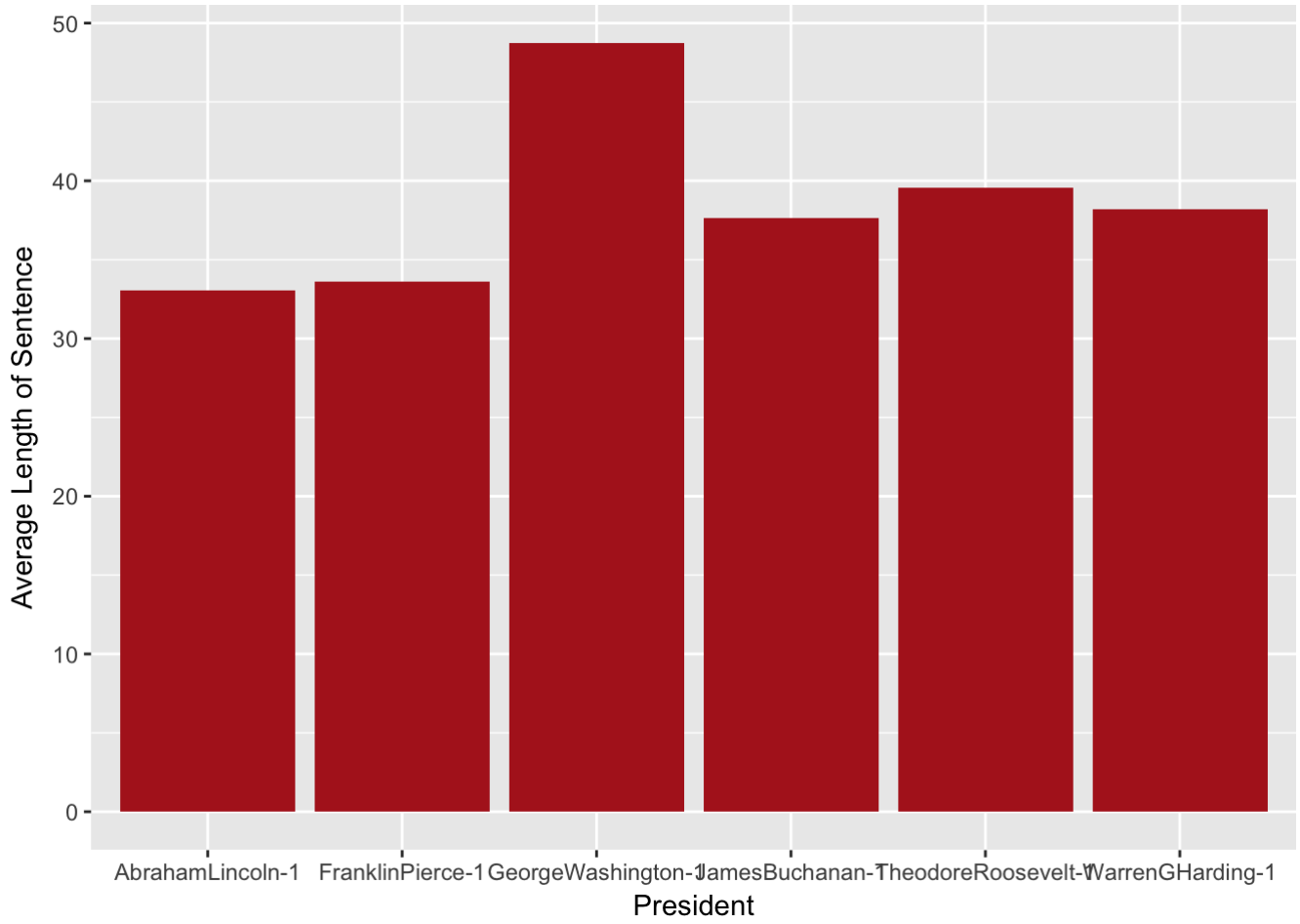
```
## Warning in wordcloud(dtm.TRoos$term, dtm.TRoos$`sum(count)`, scale = c(5, :
## tasks could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(dtm.TRoos$term, dtm.TRoos$`sum(count)`, scale = c(5, :
## world could not be fit on page. It will not be plotted.
```

The favorite presidents and the least favorite use a lot of the same words, but it's interesting to note the greater use of virtues in the favorite presidents. For example, note Theodore Roosevelt's findness for the words "Life" and "Great" and Franklin Roosevelt's use of "People" and "Democracy". ##Sentence Length Now let's look at these speeches from a different angle, how longwinded are these speeches? Are some speakers using longer sentences than others?

```
sentence.list=NULL
for(i in 1:nrow(df)){
  sentences=sent_detect(df$text[i],
                        endmarks = c("?", ".", "!", "|",";"))
  if(length(sentences)>0){
    emotions=get_nrc_sentiment(sentences)
    word.count=word_count(sentences)
    # colnames(emotions)=paste0("emo.", colnames(emotions))
    # in case the word counts are zeros?
    emotions=diag(1/(word.count+0.01))%*%as.matrix(emotions)
    sentence.list=rbind(sentence.list,
                        cbind(df[i,-ncol(df)],
                              sentences=as.character(sentences),
                              word.count,
                              emotions,
                              sent.id=1:length(sentences)
                              )
    )
  }
}
sentence.list<-as.data.frame(sentence.list)
sentence.list$speaker<-NA
for(i in 1:58){
sentence.list$speaker[sentence.list$V1==i]<-files[i]
}

sentence.length<-data.frame(
  name=files
)
sentence.length$numwords<-NA
sentence.list$word.count<-as.numeric(sentence.list$word.count)
for(names in files){
sentence.length$numwords[as.character(sentence.length$name)==names]<- mean(as.numeric(se
ntence.list$word.count[as.character(sentence.list$speaker)==names]))
}
```

```
bestworst<- c(1,16,18,26,44,49)
ggplot(sentence.length[bestworst,])+geom_col(aes(x=name,y=numwords),fill="firebrick")+yl
ab("Average Length of Sentence")+ xlab("President")
```

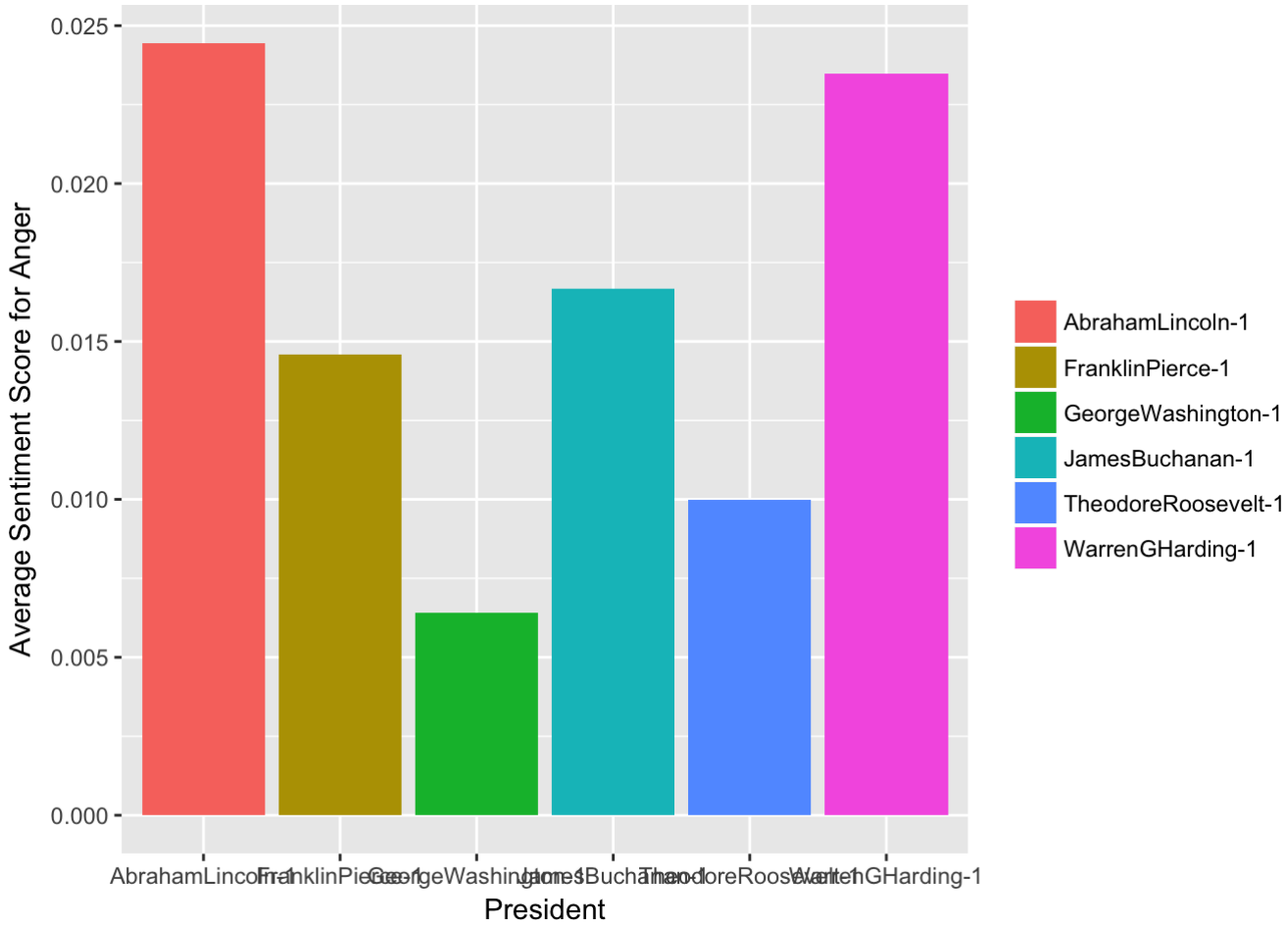


Using our subset of the best and worst presidents, we can see there's not a whole lot of variation in the number of words in a sentence. This makes sense, because these are speeches. The length of a sentence in speech would tend to be random. What about the content of the speeches?

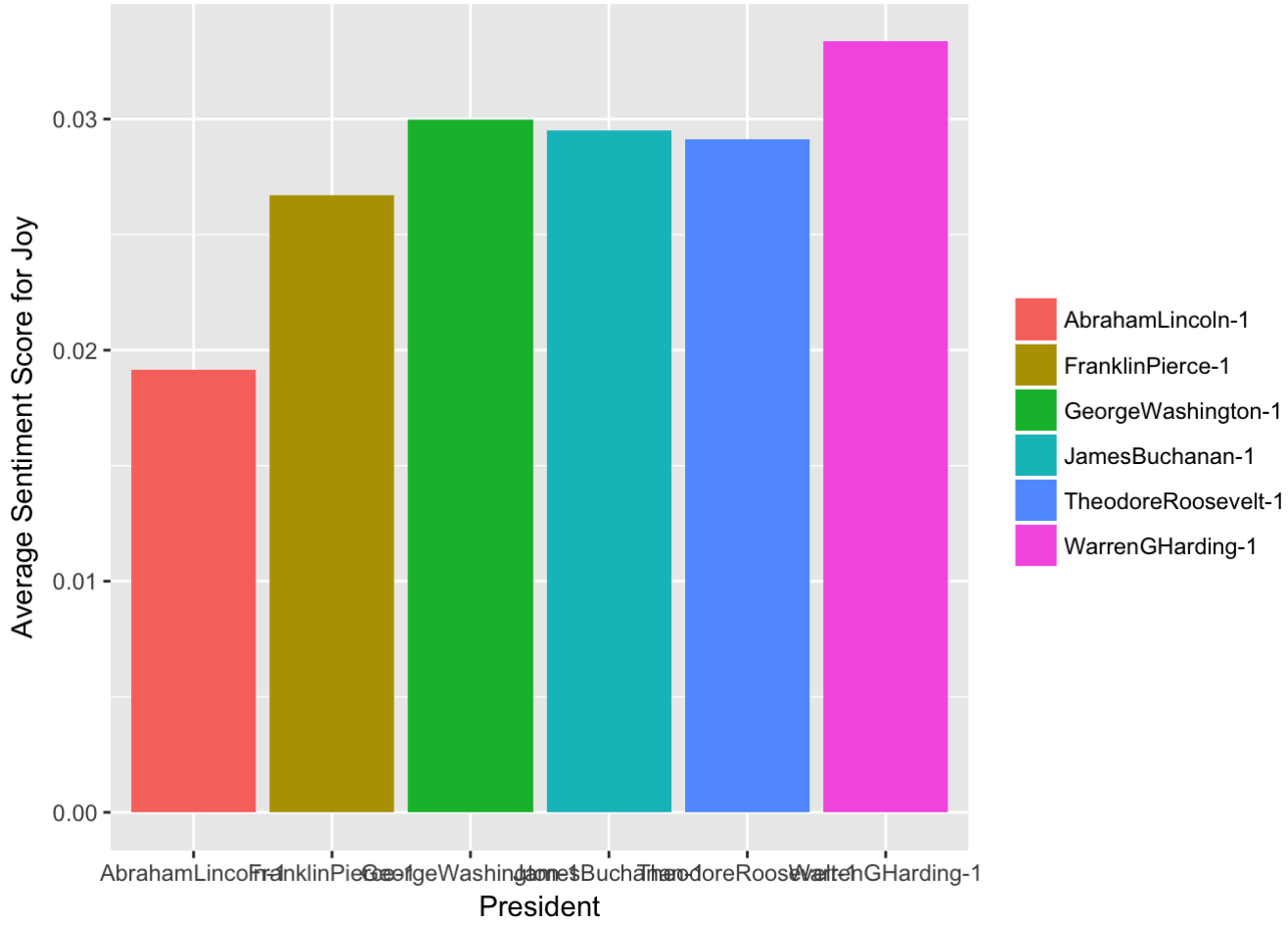
Sentiment Analysis

What tends to be the sentiment in an inaugeral speech? If we can capture the general "vibe" of a speech as positive or negative, maybe this will give us some insight into the rhetoric of a great leader.

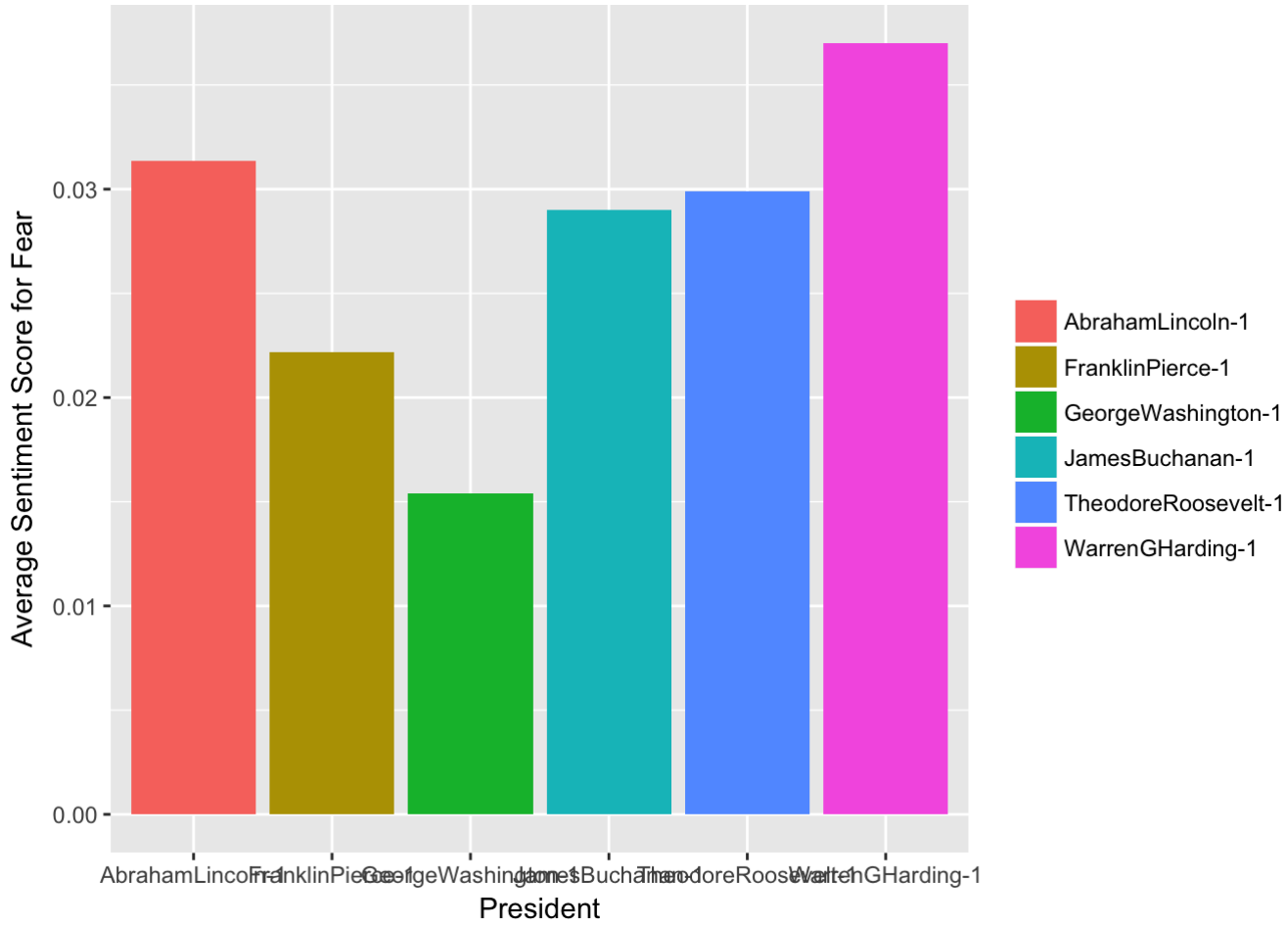
```
sentence.length$anger<-NA
sentence.length$anticipation<-NA
sentence.length$disgust<-NA
sentence.length$fear<-NA
sentence.length$joy<-NA
sentence.length$sadness<-NA
sentence.length$surprise<-NA
sentence.length$trust<-NA
sentence.length$negative<-NA
sentence.length$positive<-NA
for(names in files){

sentence.length$anger[as.character(sentence.length$name)==names]<- mean(as.numeric(as.ch
aracter(sentence.list$anger[as.character(sentence.list$speaker)==names])))
sentence.length$anticipation[as.character(sentence.length$name)==names]<- mean(as.numeri
c(as.character(sentence.list$anticipation[as.character(sentence.list$speaker)==names])))
sentence.length$disgust[as.character(sentence.length$name)==names]<- mean(as.numeric(as.
character(sentence.list$disgust[as.character(sentence.list$speaker)==names])))
sentence.length$fear[as.character(sentence.length$name)==names]<- mean(as.numeric(as.cha
racter(sentence.list$fear[as.character(sentence.list$speaker)==names])))
sentence.length$joy[as.character(sentence.length$name)==names]<- mean(as.numeric(as.char
acter(sentence.list$joy[as.character(sentence.list$speaker)==names])))
sentence.length$sadness[as.character(sentence.length$name)==names]<- mean(as.numeric(as.
character(sentence.list$sadness[as.character(sentence.list$speaker)==names])))
sentence.length$surprise[as.character(sentence.length$name)==names]<- mean(as.numeric(a
s.character(sentence.list$surprise[as.character(sentence.list$speaker)==names])))
sentence.length$trust[as.character(sentence.length$trust)==names]<- mean(as.numeric(as.c
haracter(sentence.list$trust[as.character(sentence.list$speaker)==names])))
sentence.length$negative[as.character(sentence.length$name)==names]<- mean(as.numeric(a
s.character(sentence.list$negative[as.character(sentence.list$speaker)==names])))
sentence.length$positive[as.character(sentence.length$name)==names]<- mean(as.numeric(a
s.character(sentence.list$positive[as.character(sentence.list$speaker)==names])))
}


ggplot(sentence.length[bestworst,])+geom_col(aes(x=name,y=anger,fill=name))+ylab("Averag
e Sentiment Score for Anger")+ xlab("President")+theme(legend.title=element_blank())
```



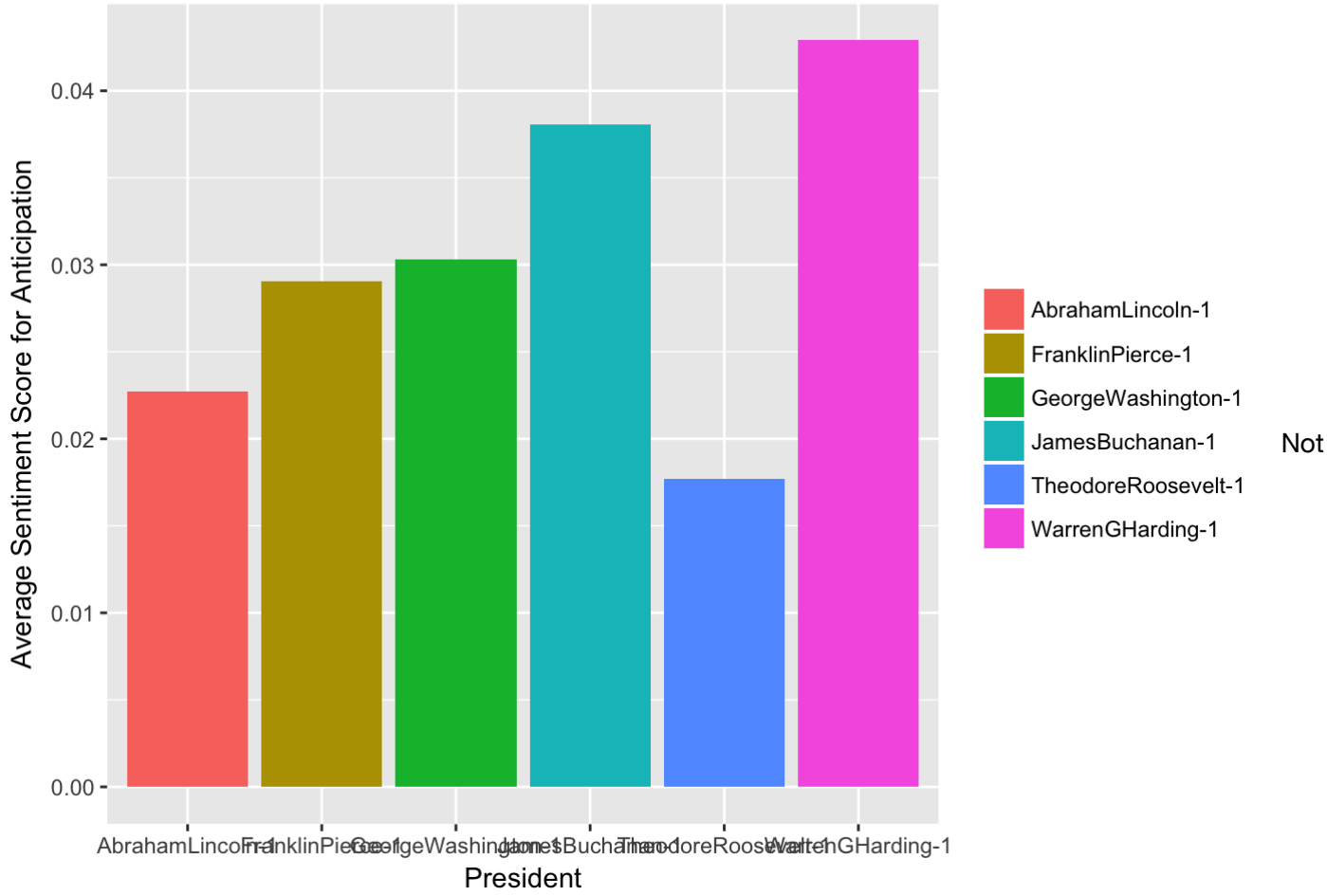

```
ggplot(sentence.length[bestworst,])+geom_col(aes(x=name,y=joy,fill=name))+ylab("Average
 Sentiment Score for Joy")+ xlab("President")+theme(legend.title=element_blank())
```




```
ggplot(sentence.length[bestworst,])+geom_col(aes(x=name,y=fear,fill=name))+ylab("Average
 Sentiment Score for Fear")+ xlab("President")+theme(legend.title=element_blank())
```

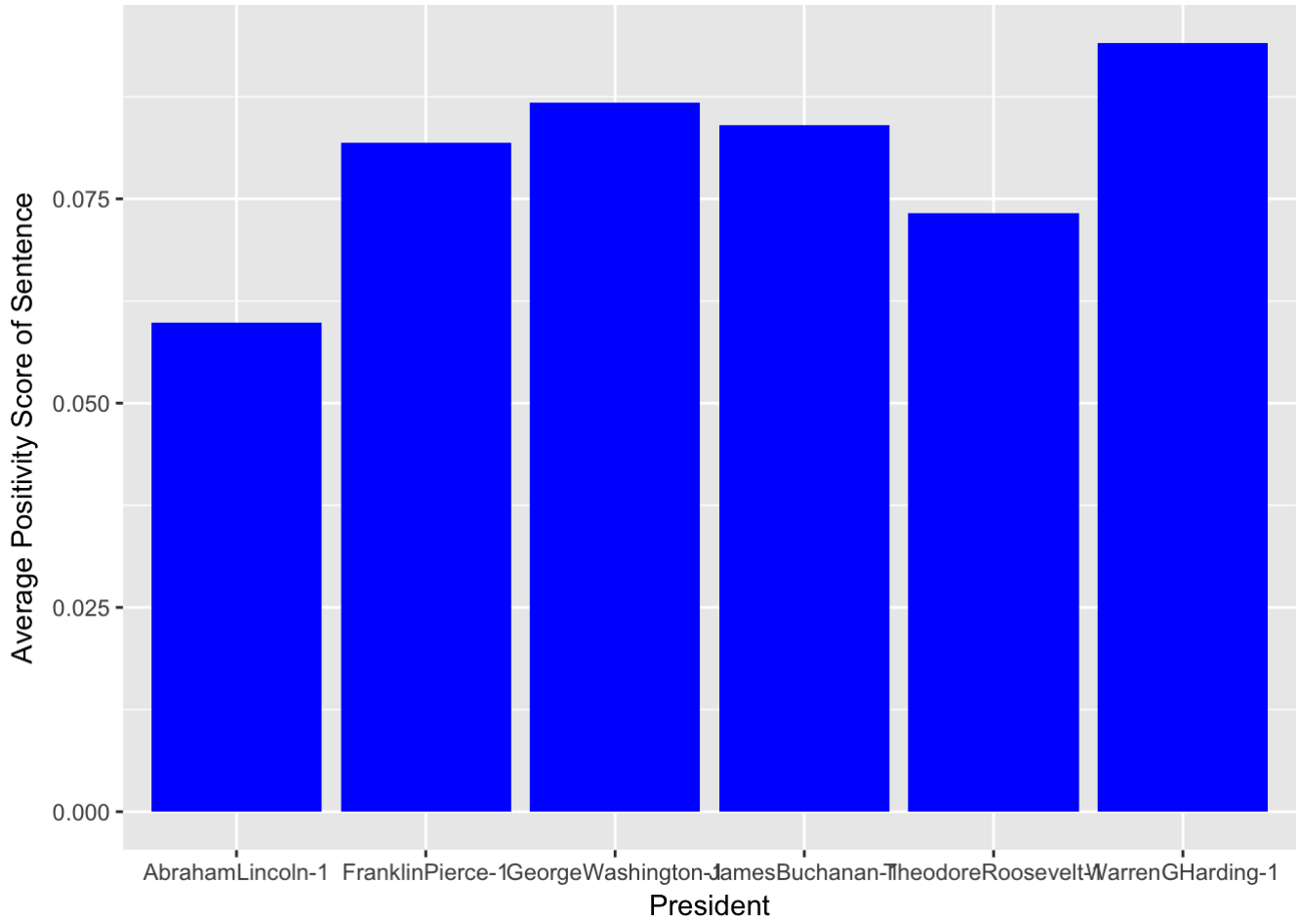



```
ggplot(sentence.length[bestworst,])+geom_col(aes(x=name,y=anticipation,fill=name))+ylab("A
verage Sentiment Score for Anticipation")+ xlab("President")+theme(legend.title=element_
blank())
```

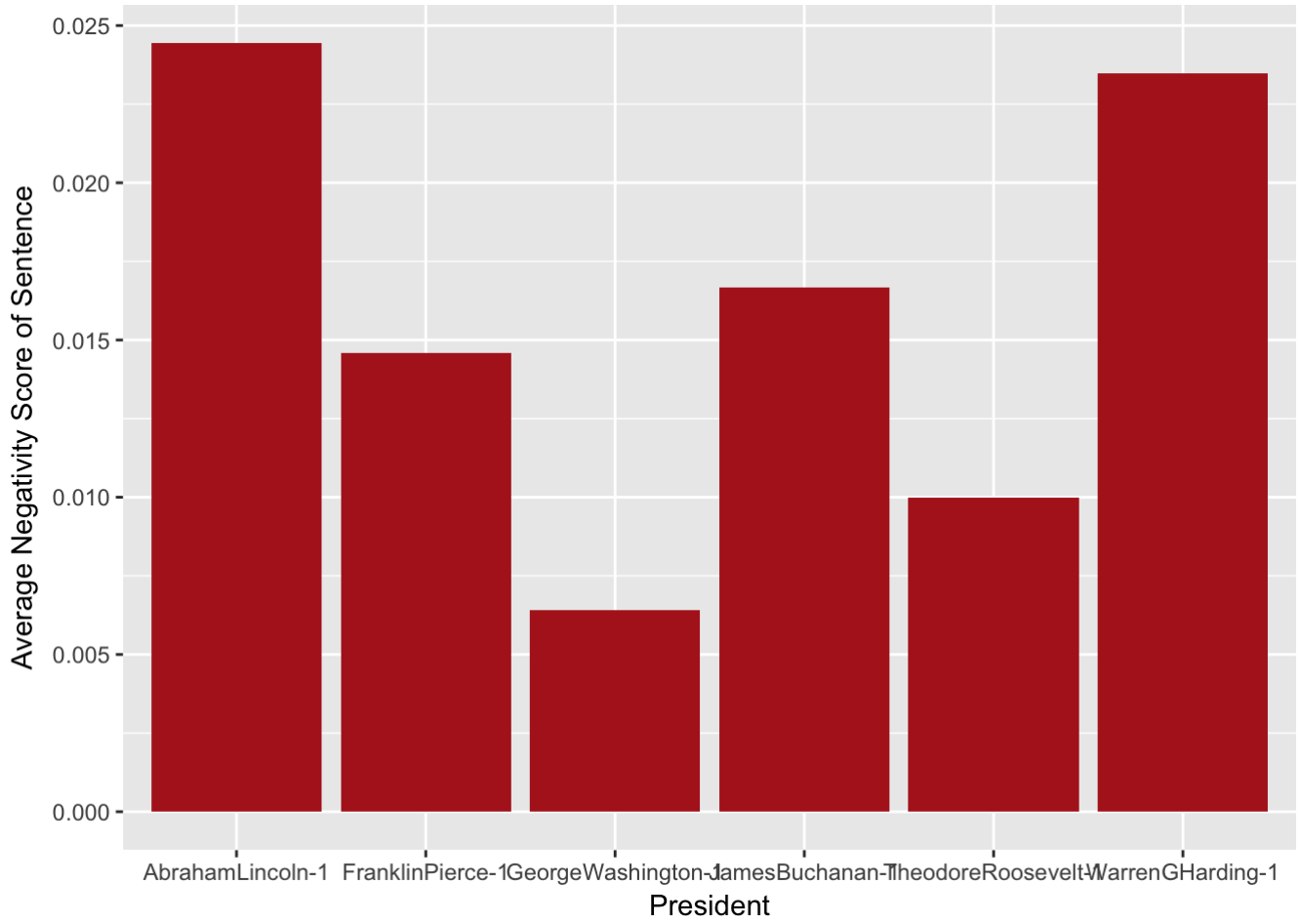


surprisingly, Lincoln's speech appears to be fearful and relatively lacking in joy. A country on the brink of a civil war!

```
ggplot(sentence.length[bestworst,])+geom_col(aes(x=name,y=positive),fill="blue")+ylab("A
verage Positivity Score of Sentence")+ xlab("President")+theme(legend.title=element_blan
k())
```




```
ggplot(sentence.length[bestworst,])+geom_col(aes(x=name,y=anger),fill="firebrick")+ylab("A
verage Negativity Score of Sentence")+ xlab("President")+theme(legend.title=element_blan
k())
```



In most of the sentiment analysis, we've seen no relation between the sentiments of the celebrated and the not so celebrated presidents of history. Although in the final graph on negativity, we see that the historians' least favorite presidents made everyone grumpy at their inaugaration speech.

Conclusions

We looked at history's favorite and least favorite presidents, and explored some of the powerful tools available to analyze text data. In the future, it'd be fun to further flesh out the historian's favroite sentiments of a president, and pull in more speech data to make sure we're getting a fair picture of the sentiment of a presidential candidate.

```
```