Joaquim Lyrio

9/9/2017

Analysis of the Inaugural Speeches of the American presidents

insert image here

Introduction

In this notebook, we will perform different analyses of the inaugural speeches of all the previous presidents of the United States of America. Different aspects of the speeches will be analysed, such as main basic differences between democrats and republicans, how the general sentiment evolvs over time, differences between speeches of first time elected and reelected presidents, and so on.

Analysis

This notebook was prepared with the following environmental settings.

```
print(R.version)
```

```
##                 _
## platform       x86_64-apple-darwin15.6.0
## arch           x86_64
## os             darwin15.6.0
## system         x86_64, darwin15.6.0
## status
## major          3
## minor          4.2
## year           2017
## month          09
## day            28
## svn rev        73368
## language       R
## version.string R version 3.4.2 (2017-09-28)
## nickname       Short Summer
```

Import data

First, let's install all necessary packages (if necessary) and load them.

```
# list of packages needed
pckgsNeeded <- c("dplyr", "tidyr", "tidytext", "readtext", "ggplot2",
                "readxl", "data.table", "scales", "wordcloud", "RColorBrewer",
                "reshape2", "gdata")

# check packages that need to be installed.
pckgsToInstall <-  setdiff(pckgsNeeded,
                        intersect(installed.packages()[,1],
                                  pckgsNeeded))
```

```r
# install additional packages
if( length( pckgsToInstall ) > 0 ){
  install.packages( pckgsToInstall, dependencies = TRUE, repos='http://cran.us.r-project.org' )
}

# load necessary libraries
library(dplyr)
library(tidyr)
library(tidytext)
library(readtext)
library(ggplot2)
library(readxl)
library(data.table)
library(scales)
library(wordcloud)
library(RColorBrewer)
library(reshape2)
library(gdata)
```

Now, let's read the input data.

```r
# data path
setwd('..')
projPath  <- getwd();

# read inauguration info
inaugInfo <- as.data.table( read.xls( paste( projPath, "/data/InaugurationInfo.xlsx", sep = ""), sheet =
inaugInfo$Party <- as.character(inaugInfo$Party)


# read inauguration date
dateFile <-  readtext( paste( projPath, "/data/InauguationDates.txt", sep = "") )
dateFile  <- strsplit( x = dateFile$text, split = "\n")
dateFile  <- sapply( dateFile[[1]][2:47], function(x) strsplit( x, split = "\t" ) )
inaugDate <- c(dateFile[[2]],"")
for( i in 3:46 ){
  if( length( dateFile[[i]] ) == 4 ){
    inaugDate <- rbind( inaugDate, c( dateFile[[i]], "" ) )
  } else{
    inaugDate <- rbind( inaugDate, dateFile[[i]] )
  }
}
colnames( inaugDate ) <- dateFile[[1]]

# initialize variables
speechesList <- list();
counts     <- list();
presidents <- c();
term       <- c();
party      <- c();
date       <- c();
i <- 1;
for( iFile in list.files( paste( projPath, "/data/InauguralSpeeches", sep = "") ) ){
```

```
# store president's name
aux <- strsplit( x = iFile, split = "inaug" )
aux <- strsplit( x = aux[[1]][2], split = "-" )
presidents[ i ] <- aux[[1]][1]

# store term information
term[ i ] <- strsplit( x = aux[[1]][length(aux[[1]])], split = ".txt" )

# conditionals deal with special name differences between files
if( presidents[ i ] == "GroverCleveland" ){

  # store party information
  party[ i ] <- "Democratic"

  # store date information
  if( iFile == "inaugGroverCleveland-I-1.txt" ){
    date <- "3/4/1885"
  } else {
    date <- "3/4/1893"
  }

} else if( presidents[ i ] == "JamesGarfield" ) {

  party[ i ] <- "Republican"
  date       <- "3/4/1881"

} else if( presidents[ i ] == "JamesKPolk" ) {

  party[ i ] <- "Democratic"
  date       <- "3/4/1845"

} else if( presidents[ i ] == "MartinvanBuren" ) {

  party[ i ] <- "Democratic"
  date       <- "3/4/1837"

} else if( presidents[ i ] == "RichardNixon" ) {

  party[ i ] <- "Republican"

  if( iFile == "inaugRichardNixon-1.txt" ){
    date       <- "1/20/1969"
  } else {
    date       <- "1/20/1973"
  }

} else {

    # store party information
    party[ i ] <- inaugInfo[ File == presidents[i] ]$Party[ 1 ]

    # store date information
    nameLong <- inaugInfo[ File == presidents[i] ]$President[1]
```

```r
    date     <- inaugDate[ which( inaugDate[,1] == nameLong ), 1 + as.numeric(term[[i]]) ]

  }

  # read speech
  speech <- readtext( paste( paste( projPath, "/data/InauguralSpeeches", sep = ""),
                                    iFile, sep = "/") );

  # put speech in tidy format
  speech <- speech %>%
    unnest_tokens(word, text);

  # remove stop words from speech
  speech <- speech %>%
    anti_join( stop_words[ stop_words$lexicon == "snowball", ] )

  # create dataframe
  speechesList[[ i ]] <- data.frame( "doc_id" = speech$doc_id,
                                     "president" = rep( presidents[i], nrow(speech) ) ,
                                     "term" = rep( term[[i]][1], nrow(speech) ),
                                     "date" = rep( date, nrow(speech) ),
                                     "party" = rep( party[i], nrow(speech) ),
                                     "word" = speech$word );

  i <- i + 1;

}

# Now, merge all speeches into a datatable
speechesDt <- as.data.table( speechesList[[1]] );
for( i in 2:length(speechesList) ){
  speechesDt <- rbind( speechesDt, speechesList[[ i ]] );
}

# convert word from factor to character vector
speechesDt$word <- as.character( speechesDt$word )
# convert dates from factor to date
speechesDt$date <- as.character( speechesDt$date )
speechesDt$date <- as.Date(speechesDt$date, "%m/%d/%Y")
# convert term to numeric
speechesDt$term <- as.numeric( speechesDt$term )
# convert president to character
speechesDt$president <- as.character( speechesDt$president )
# convert party to character
speechesDt$party <- as.character( speechesDt$party )
```

Now that we have all the speech data stored into a data.table, we can start making some analyses.

Repulicans vs Democrats First, let's analyse the difference between the most frequent words in the Republican and Democratic speeches.

```r
# perform word count
countRepublican <- speechesDt[ party == "Republican" ] %>% count(word, sort = TRUE);
countDemocrat   <- speechesDt[ party == "Democratic" ] %>% count(word, sort = TRUE);

par(mfrow=c(1,2))

# wordcloud for Republicans
wordcloud(words = countRepublican$word,
          freq = countRepublican$n,
          max.words=150, random.order=FALSE, rot.per=0.35, scale=c(5, .2),
          colors=brewer.pal(8, "Dark2"), main = "Republicans" )

# wordcloud for Democrats
wordcloud(words = countDemocrat$word,
          freq = countDemocrat$n,
          max.words=150, random.order=FALSE, rot.per=0.35, scale=c(5, .2),
          colors=brewer.pal(8, "Dark2"), main = "Democrats" )
```

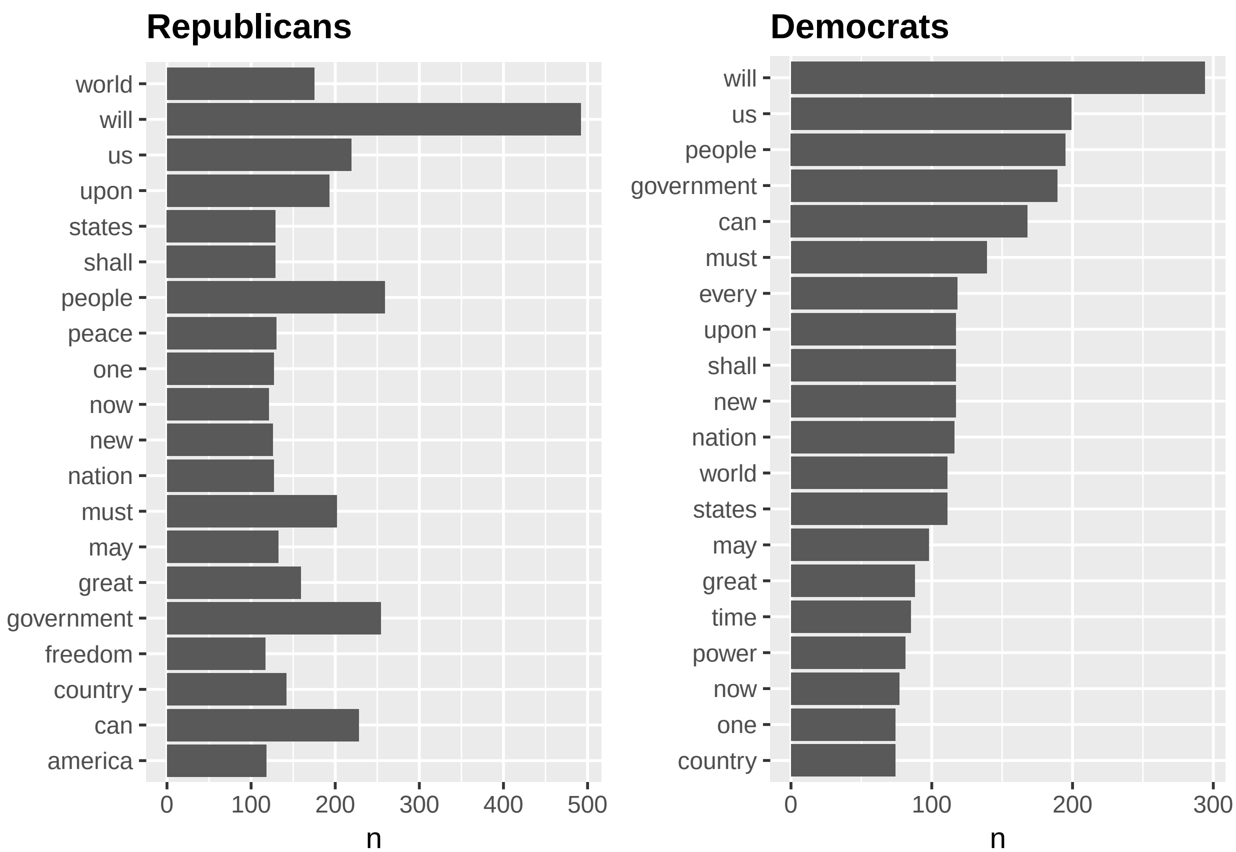

```r
par(mfrow=c(1,1))

plotR <- countRepublican[1:20,] %>%
  mutate(words = reorder(word, n)) %>%
  ggplot(aes(word, n)) +
  geom_col() +
  xlab(NULL) +
  coord_flip() +
  ggtitle("Republicans") +
  theme(plot.title = element_text(lineheight = .8, face = "bold"))

plotD <- countDemocrat[1:20,] %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n)) +
  geom_col() +
  xlab(NULL) +
```

```r
  coord_flip() +
  ggtitle("Democrats") +
  theme(plot.title = element_text(lineheight = .8, face = "bold"))

require(gridExtra)
grid.arrange(plotR, plotD, ncol=2)
```

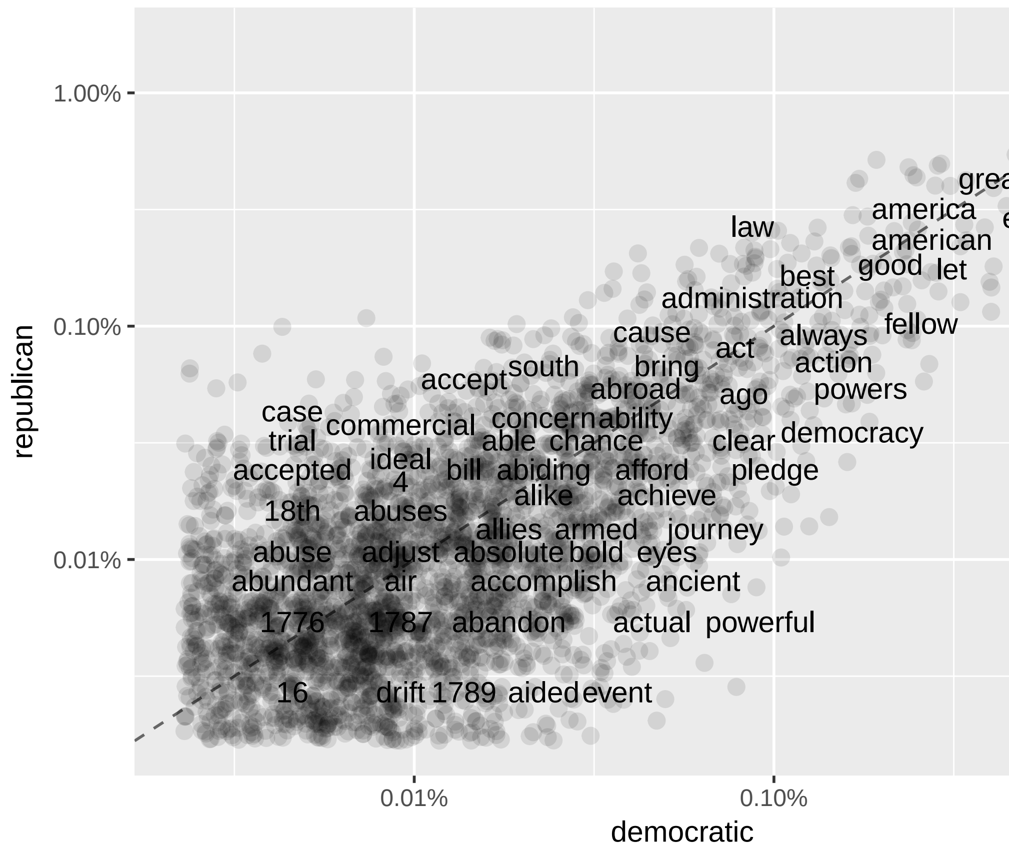


As we can see, "will" is the most common word in the speech from the presidents of both parties. This makes sense, since imaugural speeches are typically filled with promises and perspectives for the years to come. In general, the most common words among both speeched are relatively similar.

```r
frequency <- bind_rows(mutate(speechesDt[ party == "Democratic"], party = "democratic"),
                       mutate(speechesDt[ party == "Republican"], party = "republican")) %>%
  count(party, word) %>%
  group_by(party) %>%
  mutate(proportion = n / sum(n)) %>%
  select(-n) %>%
  spread(party, proportion)

ggplot( frequency, aes(x = democratic, y = republican ) ) +
  geom_abline(color = "gray40", lty = 2) +
  geom_jitter(alpha = 0.1, size = 2.5, width = 0.3, height = 0.3) +
  geom_text(aes(label = word), check_overlap = TRUE, vjust = 1.5) +
  scale_x_log10(labels = percent_format()) +
```

```
scale_y_log10(labels = percent_format()) +
scale_color_gradient(limits = c(0, 0.001), low = "darkslategray4", high = "gray75") +
theme(legend.position="none")
```



Scatter plot of word frequencies

The above plot can be interpreted the following way: points closer the vertical axis represent words that appear more frequently in republican speech, whereas the ones closer to the horizontal axis appear more in the democratic speech. The axis indicate the percentage of the aggregated speech of each party that is composed by each of the words. We can see that words such as "will", "can", "great" and "america" are very common on both speeches. On the other hand, we can notice that words such as "powerful" and "aided" are more common in the Democratic speech, whereas "trial" and "accepted" are more present in the republican one. Furthermore, we can note that the points are not so spread out so one could think that these speeches might not be so different. This might not be true since the meaning of a speech is highly dependable on how the words are put together (ie sentences), and not only by their frequency.

SENTIMENT ANALYSIS

One common aspect to be analysed in texts is the general sentiment that the author conveyed. Is the author transmitting joy to the reader? Is it anger? Various analyses can be made in order to address this issue. This way, we will perform a sentiment analysis of the inaugural speeches of the past presidents of the United States.

Sentiments Wordcloud - Positive/Negative Let's see the most common words that appear in each party's speeches, categorized by sentiment (positive/negative).

```
speechesDt %>%
  filter( party == "Republican" ) %>%
  inner_join( get_sentiments("bing") ) %>%
  count( word, sentiment, sort = TRUE ) %>%
  acast( word ~ sentiment, fill = 0 ) %>%
  comparison.cloud(colors = c("#F8766D", "#00BFC4"),
                   max.words = 100)
```

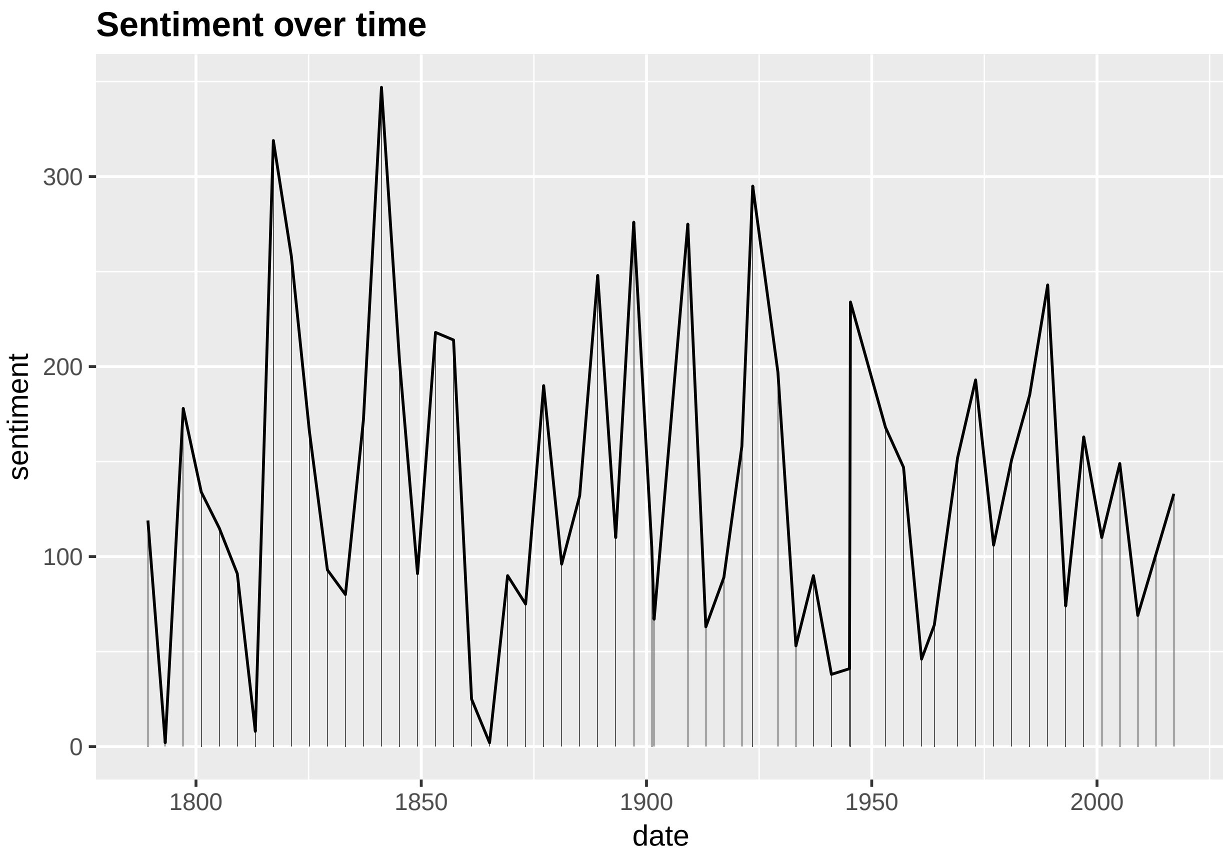
Republican

```
speechesDt %>%
  filter( party == "Democratic" ) %>%
  inner_join( get_sentiments("bing") ) %>%
  count( word, sentiment, sort = TRUE ) %>%
  acast( word ~ sentiment, fill = 0 ) %>%
  comparison.cloud(colors = c("#F8766D", "#00BFC4"),
                   max.words = 100)
```

Democratic

Sentiment over time Now, we can analyse how sentiment evolve in the presidential speech over time. For such, we will use the "afinn" sentiment dataset, which rates words from -5 to +5, indicating how negative or positive they are.

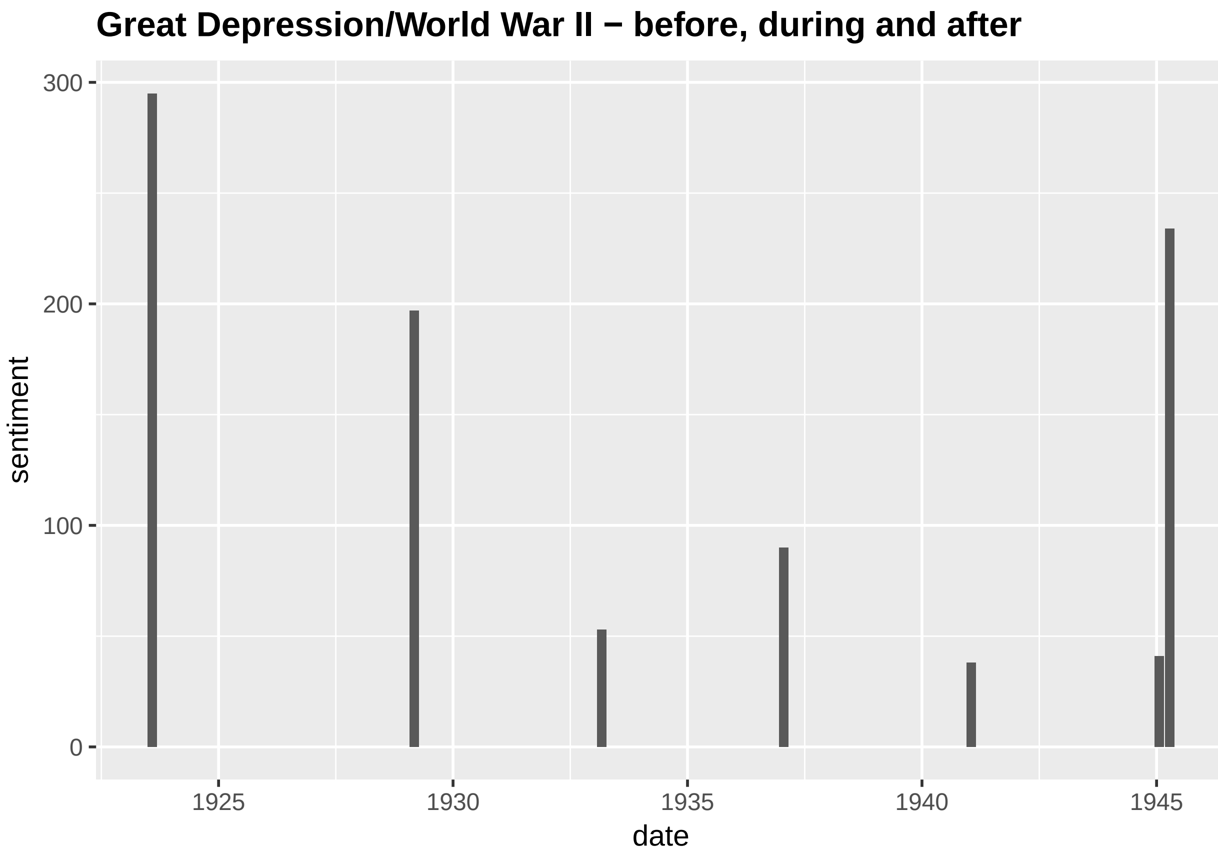
```
speechesDt %>%
  inner_join( get_sentiments("afinn") ) %>%
  group_by( date ) %>%
  summarise( sentiment = sum(score) ) %>%
  ggplot( aes(date, sentiment) ) +
  geom_bar(stat = "identity" ) +
  geom_line() +
  scale_x_date( ) +
  ggtitle("Sentiment over time") +
  theme(plot.title = element_text(lineheight = .8, face = "bold"))
```



We can notice a very low point on the graph which refers to the second speech of Abraham Lincoln, in March 4, 1865. In this speech, Abraham Lincoln talked constantly about the American civil war which was coming to an end. This is a discourse of hope, but it is loaded with mentions about all the terrible things that the war incurred in the American population.

Another interesting time frame to analyse is during the Great Depression, which went roughly from 1929 to the beginning of the 1940s.

```
speechesDt %>%
  inner_join( get_sentiments("afinn") ) %>%
  group_by( date ) %>%
  summarise( sentiment = sum(score) ) %>%
  filter( date > "1923-01-01" & date < "1945-05-01" ) %>%
  ggplot( aes(date, sentiment) ) +
  geom_bar(stat = "identity" ) +
  scale_x_date( ) +
  ggtitle("Great Depression/World War II - before, during and after") +
  theme(plot.title = element_text(lineheight = .8, face = "bold"))
```

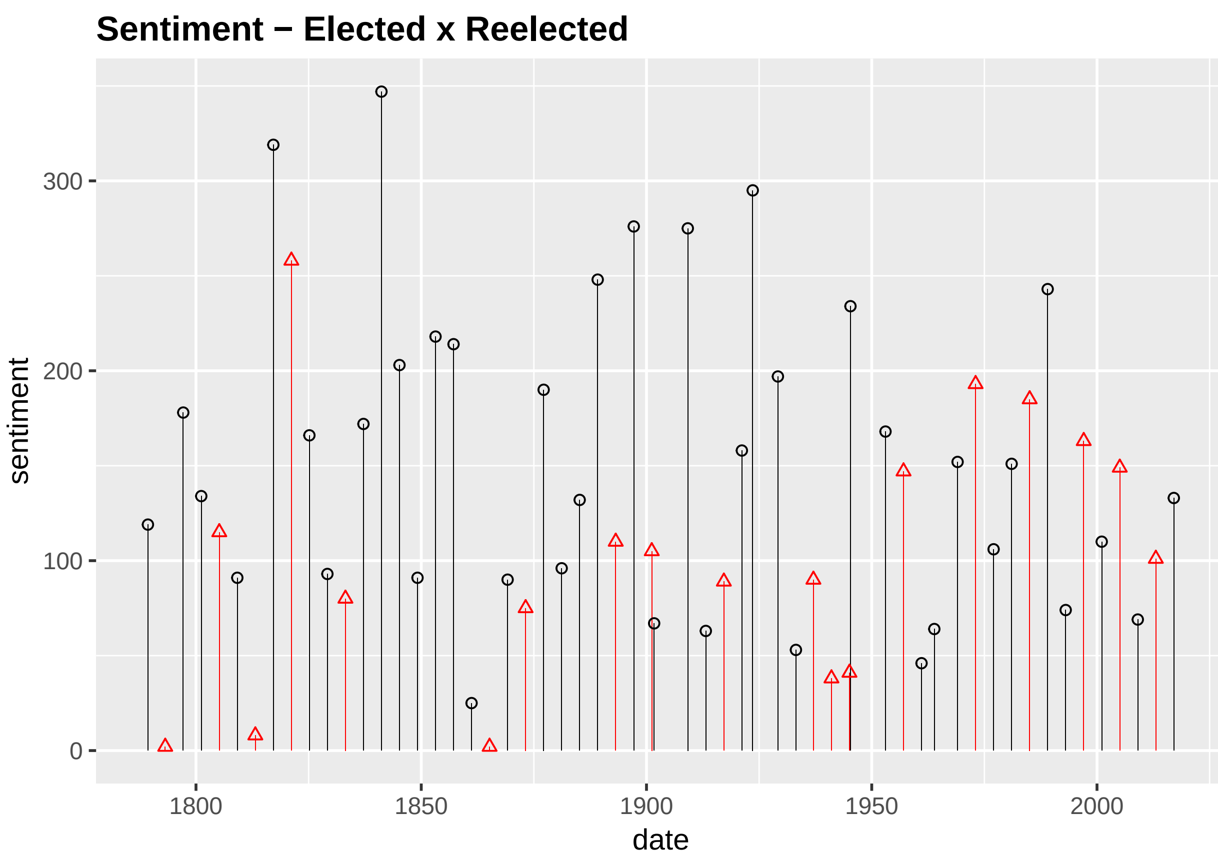


We can see that the inaugural speech of 1923, from Calvin Coolidge, was a very optimistic and positive one. It was a speech that didn't take into account the terrible times that were about to come. Following him, came Herbert Hoover, which was the one who led the US through a great part of the economic depression, from 1929-1933. His speech was rather optimistic in face of the events that were about to come and, within months after taking the office, the Stock Market Crash of 1929 happened, also known as Black Tuesday. We can notice that the level of "positivity" of the following speeches given by Franklin D. Roosevelt were very low. The world during this period was not only facing the harsh consequences of the Great Depression, but it was also passing through the World War II (1939-1945). After president Roosevelt's death, Harry Truman became the president. His inaugural speech was given at a time where the WWII was already waning, so it was full of high hopes for the times to come, with a general positive sentiment. This can be noted by the spike on the graph on his speech.

Differences in speeches between first time elected/reelected candidates Now, let's analyse if there is any difference in the speech between when a candidate is elected for the first time or reelected. One can expect that the speech of a reelected candidate must be more positive than the other one. Usually, in their inaugural speeches, presidents tend to talk about events of the last few years and, mostly, about expectations for the years to come. Naturally, it is less likely that someone would talk negatively about his own administration.

```
auxTerm <- sapply( inaugInfo$Term, function(x) ifelse( x>2, 2, x) )
speechesDt %>%
  inner_join( get_sentiments("afinn"), width = 1 ) %>%
  group_by( date ) %>%
  summarise( sentiment = sum(score) ) %>%
  ggplot( aes(date, sentiment) ) +
  geom_bar(stat = "identity", position = "identity", fill = auxTerm ) +
```

```
  geom_point( col = auxTerm, shape = auxTerm ) +
  scale_x_date( ) +
  ggtitle("Sentiment - Elected x Reelected") +
  theme(plot.title = element_text(lineheight = .8, face = "bold"))
```



In the above graph, the black lines represent first speeches and the red second speeches. We can notice on the above graph that the expectations were not completely met. Although, we can notice that there's a tendency after Woodrow Wilson that the second speech of a candidate is usually more positive than the first one. Another factor that tends to influence more this pattern is whether the elected president was preceeded by someone from the same party or not.

First term

```
speechesDt %>%
  filter( term == 1 ) %>%
  inner_join( get_sentiments("bing") ) %>%
  count( word, sentiment, sort = TRUE ) %>%
  acast( word ~ sentiment, fill = 0 ) %>%
  comparison.cloud(colors = c("#F8766D", "#00BFC4"),
                   max.words = 100)
```

Second term

```r
speechesDt %>%
  filter( term == 2 ) %>%
  inner_join( get_sentiments("bing") ) %>%
  count( word, sentiment, sort = TRUE ) %>%
  acast( word ~ sentiment, fill = 0 ) %>%
  comparison.cloud(colors = c("#F8766D", "#00BFC4"),
                   max.words = 100)
```

Furthermore, we can notice that the most common words from both first and second terms speeches are similar to each other.

Mentions about gender equality Now, let's compare the differences between mentions to women between the Republican and Democratic parties. First, we must define a list of words that indicate that the president is concerned with equal rights.

```
femaleList      <- c( "female", "woman", "women", "feminine", "maternal", "girl", "feminism", "feminist
                      "equality", "sex", "empower", "empowerment", "equal", "rigths",
                      "parent", "parenthood", "mother", "motherhood", "children", "kids",
                      "emancipation", "gender" )
```

Now, let's perform some analyses. Let's count the number of times each of the words appear in the speech of each of the Democratic presidents.

```
femaleCountDemoc  <- matrix( 0 ,
                         nrow = length( unique( speechesDt[party == "Democratic"]$president ) ),
                         ncol = length( femaleList) )
colnames(femaleCountDemoc) <- femaleList
rownames(femaleCountDemoc) <- unique( speechesDt[party == "Democratic"]$president )

i <- 1
for( iPres in unique( speechesDt[party == "Democratic"]$president ) ){

  a <- table( speechesDt[party == "Democratic"][ president == iPres ]$word )
  aux <- sapply( femaleList, function(x) a[names(a) == x] )
  femaleCountDemoc[i,] <- as.vector(sapply( aux, function(x) ifelse( length(x) == 0, 0, x ) ))
```

```
  i <- i + 1

}
```

We can perform the same analysis to the Republican party.

```
femaleCountRepub  <- matrix( 0 ,
                            nrow = length( unique( speechesDt[party == "Republican"]$president ) ),
                            ncol = length( femaleList) )
colnames(femaleCountRepub) <- femaleList
rownames(femaleCountRepub) <- unique( speechesDt[party == "Republican"]$president )

i <- 1
for( iPres in unique( speechesDt[party == "Republican"]$president ) ){

  a <- table( speechesDt[party == "Republican"][ president == iPres ]$word )
  aux <- sapply( femaleList, function(x) a[names(a) == x] )
  femaleCountRepub[i,] <- as.vector(sapply( aux, function(x) ifelse( length(x) == 0, 0, x ) ))
  i <- i + 1

}
```

We can compare the average time that any of those words appear in the speech of both parties.

Democratic

```
sum( femaleCountDemoc )/ nrow( femaleCountDemoc )
```

```
## [1] 6
```

Republican

```
sum( femaleCountRepub )/ nrow( femaleCountRepub )
```

```
## [1] 4.882353
```

As we can notice, that list of words tend to appear approximately 6 times in a Democratic speech in comparison to 4.8 times in a Republican speech. This might indicate that Democratic presidents tend to mention more frequently themes related to gender equality.

Furthermore, we can perform an analysis of how this topic mention evolve over time in presidential speeches.

```
femaleCountDates  <- matrix( 0 ,
                            nrow = length( unique( speechesDt$date ) ),
                            ncol = length( femaleList) )
colnames(femaleCountDates) <- femaleList
rownames(femaleCountDates) <- unique( speechesDt$date )

i <- 1
for( iDate in unique( speechesDt$date ) ){

  a <- table( speechesDt[ date == iDate ]$word )
```

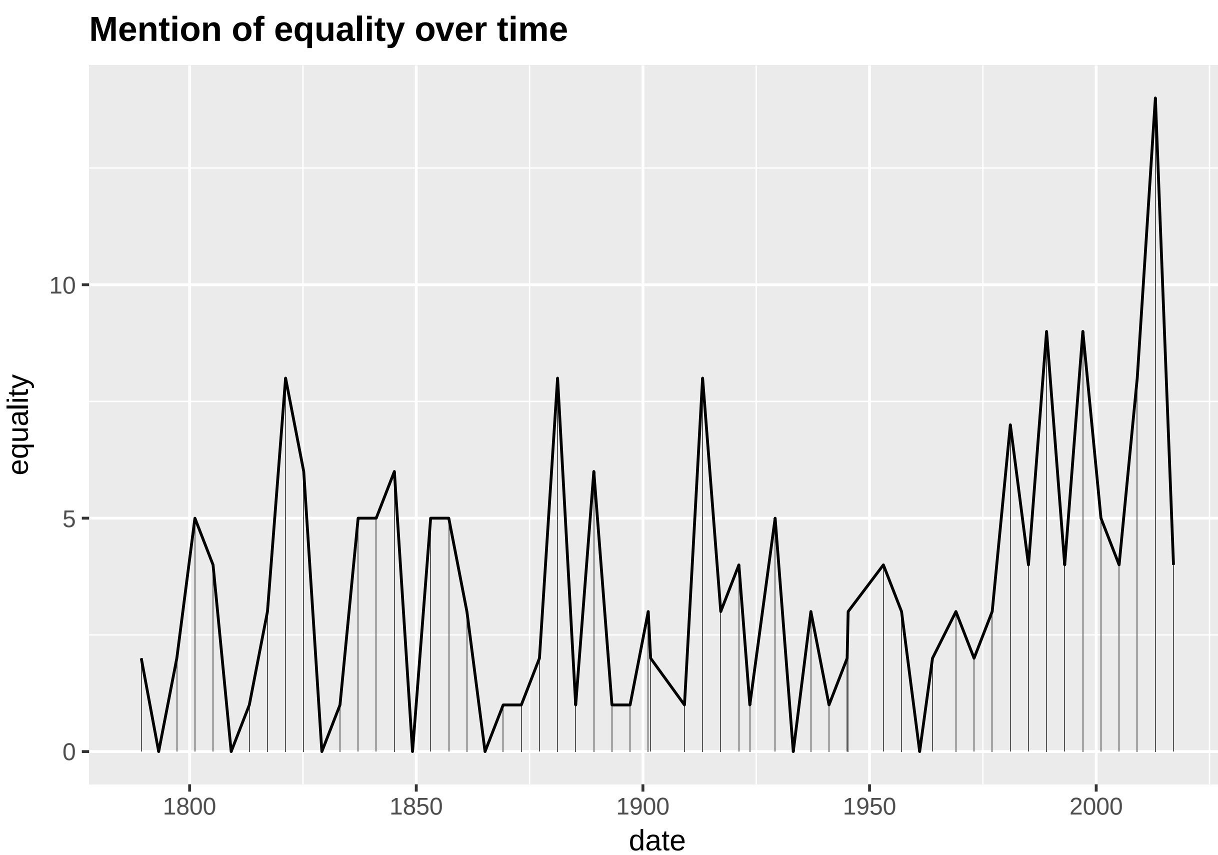
```
  aux <- sapply( femaleList, function(x) a[names(a) == x] )
  femaleCountDates[i,] <- as.vector(sapply( aux, function(x) ifelse( length(x) == 0, 0, x ) ))
  i <- i + 1

}

plotDf <- data.frame( "date" = unique( speechesDt$date ),
                      "equality" = rowSums( femaleCountDates ),
                      "party" = rep(-1, length( unique( speechesDt$date ) ) ) )

plotDf       <- plotDf[ order( plotDf$date), ]

ggplot( plotDf, aes(date, equality) ) +
  geom_bar(stat = "identity" ) +
  geom_line() +
  scale_x_date( ) +
  ggtitle("Mention of equality over time") +
  theme(plot.title = element_text(lineheight = .8, face = "bold"))
```



We can notice that the mention of equality related words had its peak on the second speech of president Barack Obama. Additionally, we can note that the mention to equality stabilized at a higher level from around the 1970s on and there seems to be some kind of cyclic behaviour in the frequency of these words.

Conclusion

As we can see, text mining represents a very powerful tool to discover patterns among texts. From our analysis, we can see that there are some basic differences between the Republican and Democratic speeches. The whole analysis made was based on single-word comparisons, rather than sentences. Topic modelling and the analysis of sentences would represent a next step to the discovery of underlying patterns in the inaugural presidential speeches of the United States of America.