Social Science Testimony

Julia Tache

03/02/2021

While looking through the Data is Plural archive, I was interested in the dataset on social scientists testifying in congress. I always wondered about the efficacy of "expert witnesses" in court cases as well as in congressional hearings. People who have very extensive experience in a certain topic could provide an excellent resource in the consultation of legislation, especially if the bill or other initiative they are sharing their testimony for is directly related to their field of study. They may be able to persuade the members of congress to vote in a certain way if their testimony is strong enough. Because the individuals in the dataset are social scientists, one can assume that they would prioritize social benefits of whatever legislative process the congressional committee is going through. For example, an economist might make suggestions on the budget based on the financial needs of the country, or a sociologist might suggest ways to make a bill about housing more equitable and fairer to those affected by structural inequities. The presence of social scientists in congress could have a very positive impact overall.

Of course, the opposite could very well be true: what if their testimonies are not listened to and respected, even if their advice is beneficial? What if these scholars, academics, thinktank researchers, etc. have their own special interests they are trying to promote? What if some members of certain disciplines are given more credence than others? I was curious about the tangible effect that these testimonies could have on actual legislative change.

My research question is whether or not social science testimonies make an impact in congressional hearings? By impact, I decided somewhat arbitrarily on a few factors: the total presence of social scientists at hearings, the ability to get bills passed during related hearings, as well as the ability to incite change in specific committees.

```
library(tidyverse)
library(data.table)
library(DT)
library(stargazer)

sst <- read_csv("social science congressional testimony.csv")
gov_df <- read_csv("gov_df.csv")
bills_df <- read_csv("bills_df.csv")

testimonies <- readtext::readtext("*.txt")

# divide each group of testimonies based on discipline
econ <- testimonies[1:5, ]
polisci <- testimonies[6:10, ]
psych <- testimonies[11:13, ]
soc <- testimonies[14:18, ]
```

Using the Social Science Congressional Testimony dataset, I first looked at the most popular committees that social scientists attend. For all of time, the most popular is the Joint Economic Committee, but from 2015 onwards the most popular committee in the data was the House Committee on Financial Services. I looked at 2015 onwards because this represents the most recent sessions of congress in the data: the 114th to 116th

sessions. I figured that these data would be the most relevant to my research question because I would not have to take historical factors into consideration as much.

```
# most popular committee
table_com <- as.data.frame(table(sst$full_committee1))
head(table_com[order(-table_com$Freq), ])
```

```
##                                                    Var1 Freq
## 38                            committee on economic. joint 1154
## 125                        committee on ways and means. house  967
## 118                        committee on the judiciary. senate  686
## 21  committee on banking, housing, and urban affairs. senate  682
## 50                            committee on finance. senate  653
## 52                         committee on foreign affairs. house  565
```

```
# most common witness affiliations
table_aff <- as.data.frame(table(sst$witness_affiliation))
head(table_aff[order(-table_aff$Freq), ])
```

```
##                                                      Var1 Freq
## 1760                        chm, council of economic advisers  163
## 223  american enterprise institute for public policy research  144
## 1386                          chairman, federal reserve board  138
## 2694                                        director, cbo  129
## 2105                                                d -ny  122
## 8585                             sr fellow, brookings instn   92
```

```
sst_recent <- sst %>%
  filter(year >= 2015)

sst_recent
```

```
## # A tibble: 332 x 29
##       X1 hearing_id   year date_begin  date_end witness_name witness_affiliation
##    <dbl> <chr>       <dbl> <chr>       <chr>    <chr>        <chr>
##  1 15151 HRG-2015-P~  2015 Jul. 15, 2~ Jul. 15~ susan colli~ r -me
##  2 15152 HRG-2015-H~  2015 Jan. 21, 2~ Jan. 22~ alice rivlin economic studies, ~
##  3 15153 HRG-2015-H~  2015 13-May-15   13-May-~ robert joha~ usda
##  4 15154 HRG-2015-A~  2015 Feb. 13, 2~ Feb. 13~ marc lynch   institute for midd~
##  5 15155 HRG-2015-H~  2015 Jul. 23, 2~ Jul. 23~ sujit chakr~ clearing house ass~
##  6 15156 HRG-2015-H~  2015 Sep. 30, 2~ Sep. 30~ richard ved~ center for college~
##  7 15157 HRG-2015-H~  2015 Sep. 30, 2~ Sep. 30~ philip joyce school of public p~
##  8 15158 HRG-2015-N~  2015 Apr. 22, 2~ Apr. 22~ lynn scarle~ public policy , na~
##  9 15159 HRG-2015-H~  2015 Feb. 26, 2~ Feb. 26~ douglas hol~ american action fo~
## 10 15160 HRG-2015-F~  2015 Feb. 4, 20~ Feb. 4,~ danielle pl~ foreign and defens~
## # ... with 322 more rows, and 22 more variables: discipline1 <chr>,
## #   discipline2 <chr>, discipline3 <lgl>, sociologist <dbl>, economist <dbl>,
## #   anthropologist <dbl>, psychologist <dbl>, polscientist <dbl>, agency <chr>,
## #   census <dbl>, nsfsbe <dbl>, thinktank <dbl>, thinktank1 <chr>,
## #   thinktank2 <lgl>, full_committee1 <chr>, sub_committee1 <chr>,
## #   full_committee2 <chr>, sub_committee2 <chr>, title_description <chr>,
## #   filter <dbl>, HID <dbl>, univ_aff <dbl>
```

```
library(data.table)
table_df <- as.data.frame(table(sst_recent$full_committee1)) %>%
  rename("Frequency" = "Freq") %>%
  rename("Committee Name" = "Var1")
```

```
top5_comm <- table_df[order(-table_df$Frequency), ][0:5, ]

table_df$`Committee Name` <- str_to_title(table_df$`Committee Name`)

datatable(head(table_df[order(-table_df$Frequency), ]))
```

I created a similar frequency table for the witnesses who spoke, the most represented disciplines in the field, and witness affiliation.

```
table_df <- as.data.frame(table(sst_recent$witness_name)) %>%
  rename("Frequency" = "Freq") %>%
  rename("Witness" = "Var1")

table_df$Witness <- str_to_title(table_df$Witness)

datatable(table_df[order(-table_df$Frequency), ])
```

```
table_df <- as.data.frame(table(sst_recent$discipline1)) %>%
  rename("Frequency" = "Freq") %>%
  rename("Discipline" = "Var1")

table_df$Discipline <- str_to_title(table_df$Discipline)

datatable(table_df[order(-table_df$Frequency), ])
```

```
table_df <- as.data.frame(table(sst_recent$witness_affiliation)) %>%
  rename("Frequency" = "Freq") %>%
  rename("Witness Affiliation" = "Var1")

data.table(table_df[order(-table_df$Frequency), ])
```

```
##                                                                                Witness Affiliation
##   1:                                    american enterprise institute for public policy research
##   2:                                                                          american action forum
##   3:                                                                                       cbo
##   4:                                                                          federal reserve board
##   5:                                                            economics , stanford university
##  ---
## 185:                                                                trade policy , cornell university
## 186:                                                     tuck school of business , dartmouth college
## 187: ware center for security studies , american enterprise institute for public policy research
## 188:                                                     washington center for equitable growth
## 189:                                                                                wiley rein
##      Frequency
##   1:        19
##   2:        18
##   3:        14
##   4:        13
##   5:        11
##  ---
## 185:         1
## 186:         1
## 187:         1
## 188:         1
```

```
## 189:         1
```

```
df <- sst_recent %>%
  count(discipline1)

df[order(-df$n), ]
```

```
## # A tibble: 5 x 2
##   discipline1             n
##   <chr>               <int>
## 1 Economist             228
## 2 Political Scientist    89
## 3 Sociologist            10
## 4 Psychologist            4
## 5 Anthropologist          1
```

```
df <- sst_recent %>%
  filter(discipline1 == "Political Scientist") %>%
  count(discipline1, witness_name)

df[order(-df$n), ]
```

```
## # A tibble: 62 x 3
##    discipline1         witness_name         n
##    <chr>               <chr>            <int>
##  1 Political Scientist philip joyce         5
##  2 Political Scientist seth jones           5
##  3 Political Scientist donald kettl         4
##  4 Political Scientist james capretta       3
##  5 Political Scientist andrew kelly         2
##  6 Political Scientist daniel benjamin      2
##  7 Political Scientist danielle pletka      2
##  8 Political Scientist diana negroponte     2
##  9 Political Scientist graham allison       2
## 10 Political Scientist gregory mcneal       2
## # ... with 52 more rows
```

```
df <- sst_recent %>%
  filter(discipline1 == "Sociologist") %>%
  count(discipline1, witness_name)

df[order(-df$n), ]
```

```
## # A tibble: 7 x 3
##   discipline1 witness_name       n
##   <chr>       <chr>          <int>
## 1 Sociologist leon aron          3
## 2 Sociologist hal salzman        2
## 3 Sociologist amitai etzioni     1
## 4 Sociologist jeffrey passel     1
## 5 Sociologist nikos passas       1
## 6 Sociologist w. wilcox          1
## 7 Sociologist walden bello       1
```

```
df <- sst_recent %>%
  filter(discipline1 == "Psychologist") %>%
  count(discipline1, witness_name)
```

```
df[order(-df$n), ]
```

```
## # A tibble: 4 x 3
##   discipline1  witness_name        n
##   <chr>        <chr>           <int>
## 1 Psychologist frances deviney     1
## 2 Psychologist keith payne         1
## 3 Psychologist mark mahone         1
## 4 Psychologist tami decoteau       1
```

Next, I did some NLP and sentiment analysis to understand the nature of the hearings. I used the title_description variable to create a corpus of text, a document term matrix, and another variable calculating the sentiment value of the description using the syuzhet package. This process of sentiment analysis is concerned with "plot" development as well as sentiment by utilizing several different existing dictionaries. I felt that this was suitable and comprehensive enough to calculate the sentiments of the descriptions of the hearings because it would give an idea of the positive/negative attributes of these hearings while also giving us a sense of how "well-developed" the intention of the hearing was.

Code template from https://www.red-gate.com/simple-talk/sql/bi/text-mining-and-sentiment-analysis-with-r/

```
library(tidytext)
library(tm)
library(SnowballC)
library(RColorBrewer)
library(syuzhet)

TextDoc <- Corpus(VectorSource(sst_recent$title_description))

toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
TextDoc <- tm_map(TextDoc, toSpace, "/")
TextDoc <- tm_map(TextDoc, toSpace, "@")
TextDoc <- tm_map(TextDoc, toSpace, "\\|")

# Convert the text to lower case
TextDoc <- tm_map(TextDoc, content_transformer(tolower))

# Remove numbers
TextDoc <- tm_map(TextDoc, removeNumbers)

# Remove english common stopwords
TextDoc <- tm_map(TextDoc, removeWords, stopwords("english"))

# Remove punctuations
TextDoc <- tm_map(TextDoc, removePunctuation)

# Eliminate extra white spaces
TextDoc <- tm_map(TextDoc, stripWhitespace)
```

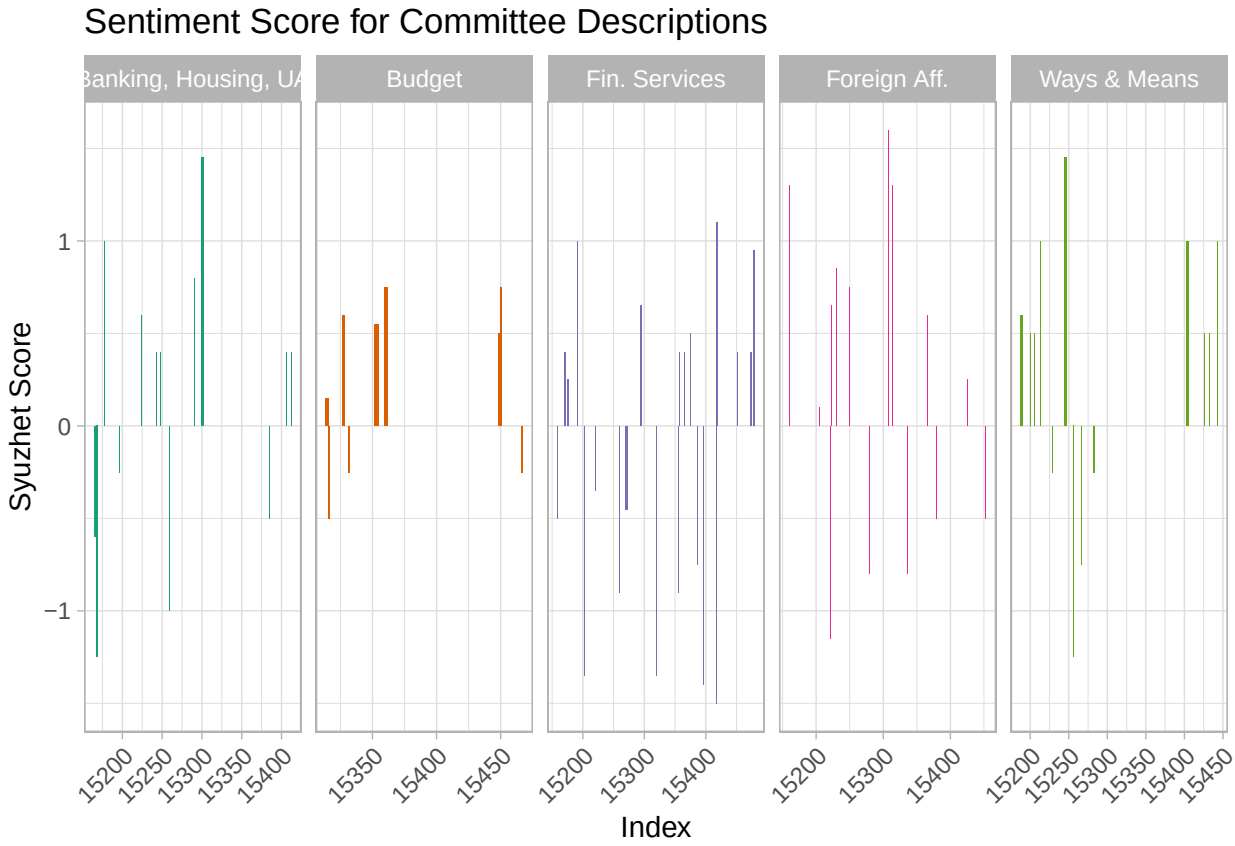
```
# Build a term-document matrix
TextDoc_dtm <- TermDocumentMatrix(TextDoc)
dtm_m <- as.matrix(TextDoc_dtm)

# Sort by descearing value of frequency
dtm_v <- sort(rowSums(dtm_m),decreasing=TRUE)
```

```r
dtm_d <- data.frame(word = names(dtm_v),freq=dtm_v)
dtm_d <- dtm_d %>%
  rename("Word" = "word") %>%
  rename("Frequency" = "freq")

# Display the top 5 most frequent words
head(dtm_d, 5)
```

```
##              Word Frequency
## federal   federal        27
## budget     budget        27
## policy     policy        26
## economic economic        25
## reform     reform        24
```

Looking at the five most common committees in the data, the sentiments ranged widely from very positive to very negative across hearings. This could suggest that these hearings were of a very pressing nature and required experts to speak on the subject immediately.

```r
sst_recent$syuzhet_vector <- get_sentiment(sst_recent$title_description, method="syuzhet")

top5_comm_graph <- sst_recent %>%
  filter(full_committee1 %in% top5_comm$`Committee Name`) %>%
  mutate(full_committee1 = replace(full_committee1, full_committee1 == "committee on financial services
  mutate(full_committee1 = replace(full_committee1, full_committee1 == "committee on foreign affairs. ho
  mutate(full_committee1 = replace(full_committee1, full_committee1 == "committee on banking, housing, a
  mutate(full_committee1 = replace(full_committee1, full_committee1 == "committee on ways and means. hou
  mutate(full_committee1 = replace(full_committee1, full_committee1 == "committee on budget. house", "Bu
  filter(syuzhet_vector != 0)

ggplot(top5_comm_graph, aes(X1, syuzhet_vector, fill = full_committee1)) +
  geom_col(show.legend = FALSE) +
  facet_grid(~full_committee1, scales = "free") +
  theme_light() +
  ggtitle("Sentiment Score for Committee Descriptions") +
  xlab("Index") +
  ylab("Syuzhet Score") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_brewer(palette = "Dark2")
```



Using binary variables in the data, I compared the syuzhet vector to the outcome variable of the field of the expert present. However, none of the results were significant.

```
fit <- glm(economist ~ syuzhet_vector, sst_recent, family = "binomial")
fit_2 <- glm(polscientist ~ syuzhet_vector, sst_recent, family = "binomial")
fit_3 <- glm(sociologist ~ syuzhet_vector, sst_recent, family = "binomial")

stargazer(fit, fit_2, fit_3, title="Results", align=TRUE, type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Wed, Mar 03, 2021 - 03:05:22 % Requires LaTeX packages: dcolumn

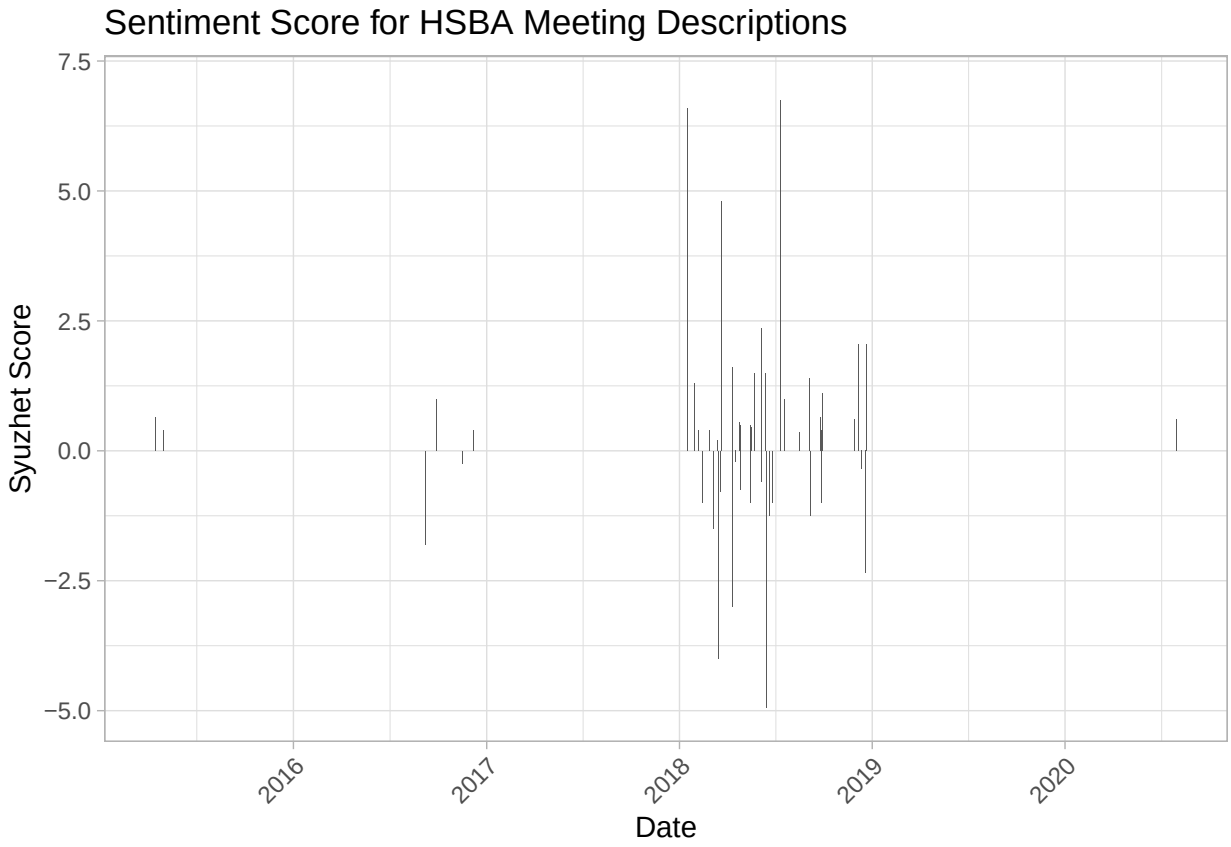
Because the data were a bit limited, I retrieved some additional data using the ProPublica API for data on congress. For the most popular committee in the original dataset where most social scientists were present, HSBA, the sentiment of meetings spike in intensity in 2018 both positively and negatively.

```
HSBA <- gov_df %>%
  filter(committee_code == "HSBA")

HSBA$syuzhet_vector <- get_sentiment(HSBA$description, method="syuzhet")

HSBA <- HSBA %>%
  filter(syuzhet_vector != 0)

ggplot(HSBA, aes(date, syuzhet_vector)) +
  geom_col(show.legend = FALSE) +
  theme_light() +
  ggtitle("Sentiment Score for HSBA Meeting Descriptions") +
  xlab("Date") +
```

Table 1: Results

	Dependent variable:		
	economist	polscientist	sociologist
	(1)	(2)	(3)
syuzhet_vector	0.211	−0.246	−0.245
	(0.163)	(0.169)	(0.439)
Constant	0.737***	−0.905***	−3.426***
	(0.123)	(0.127)	(0.326)
Observations	332	332	332
Log Likelihood	-205.552	-194.878	-44.717
Akaike Inf. Crit.	415.104	393.756	93.434
Note:			*p<0.1; **p<0.05; ***p<0.01

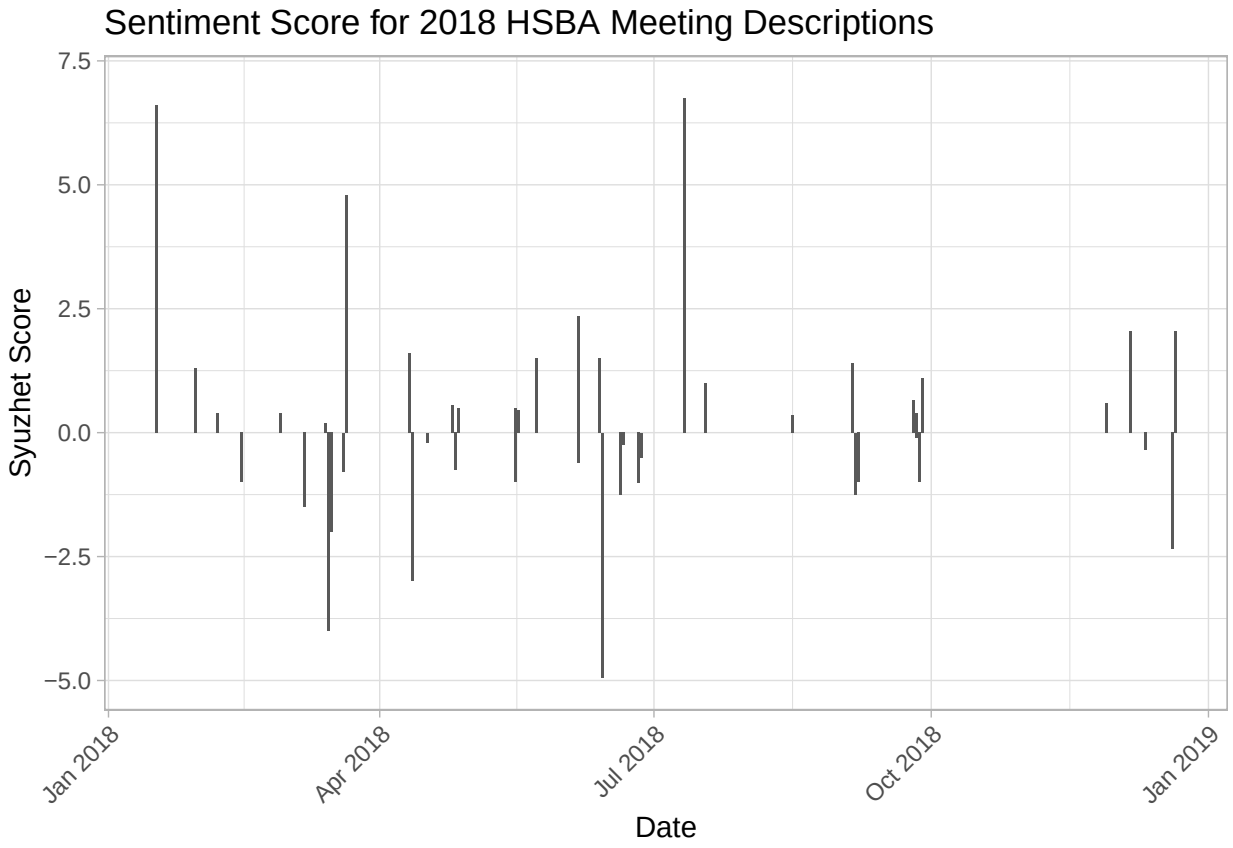
```
  ylab("Syuzhet Score") +
  ylim(-5, 7) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
HSBA$year <- str_extract(HSBA$date, "\\d{4}")

gov_df
```

```
## # A tibble: 5,867 x 11
##    chamber committee     committee_code api_uri        date       time  location
##    <chr>   <chr>         <chr>          <chr>          <date>     <tim> <chr>
##  1 House   Committee on~ HSBA           https://api.p~ 2016-12-08 09:30 RHOB 21~
##  2 House   Committee on~ HSGO           https://api.p~ 2016-12-08 09:00 RHOB 21~
##  3 House   Committee on~ HSGO           https://api.p~ 2016-12-07 14:00 RHOB 21~
##  4 House   Committee on~ HSIF           https://api.p~ 2016-12-07 10:00 RHOB 23~
##  5 House   Committee on~ HSBA           https://api.p~ 2016-12-07 10:00 RHOB 21~
##  6 House   Committee on~ HSPW           https://api.p~ 2016-12-07 10:00 RHOB 21~
##  7 House   Committee on~ HSFA           https://api.p~ 2016-12-07 10:00 RHOB 21~
##  8 House   Committee on~ HSAG           https://api.p~ 2016-12-07 10:00 LHOB 13~
##  9 House   Committee on~ HSGO           https://api.p~ 2016-12-07 09:00 RHOB 21~
## 10 House   Committee on~ HSRU           https://api.p~ 2016-12-06 15:00 CAPITOL~
## # ... with 5,857 more rows, and 4 more variables: description <chr>,
## #   bill_ids <chr>, url <chr>, meeting_type <chr>
```

```
HSBA_recent <- HSBA %>%
  filter(year == 2018)

ggplot(HSBA_recent, aes(date, syuzhet_vector)) +
  geom_col(show.legend = FALSE) +
  theme_light() +
  ggtitle("Sentiment Score for 2018 HSBA Meeting Descriptions") +
  xlab("Date") +
  ylab("Syuzhet Score") +
  ylim(-5, 7) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The document term matrix and word frequency table suggest that most of the hearings at the time were related to regulations on investment protection.

```
TextDoc <- Corpus(VectorSource(HSBA_recent$description))

toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
TextDoc <- tm_map(TextDoc, toSpace, "/")
TextDoc <- tm_map(TextDoc, toSpace, "@")
TextDoc <- tm_map(TextDoc, toSpace, "\\|")

# Convert the text to lower case
TextDoc <- tm_map(TextDoc, content_transformer(tolower))

# Remove numbers
TextDoc <- tm_map(TextDoc, removeNumbers)

# Remove english common stopwords
TextDoc <- tm_map(TextDoc, removeWords, stopwords("english"))

# Remove punctuations
TextDoc <- tm_map(TextDoc, removePunctuation)

# Eliminate extra white spaces
TextDoc <- tm_map(TextDoc, stripWhitespace)
```

```
# Build a term-document matrix
TextDoc_dtm <- TermDocumentMatrix(TextDoc)
dtm_m <- as.matrix(TextDoc_dtm)

# Sort by descearing value of frequency
dtm_v <- sort(rowSums(dtm_m),decreasing=TRUE)
dtm_d <- data.frame(word = names(dtm_v),freq=dtm_v)
dtm_d <- dtm_d %>%
  rename("Word" = "word") %>%
  rename("Frequency" = "freq")

# Display the top 5 most frequent words
head(dtm_d, 5)
```

```
##                  Word Frequency
## act               act       107
## regulatory regulatory        27
## small           small        23
## investment investment        19
## protection protection        18
```

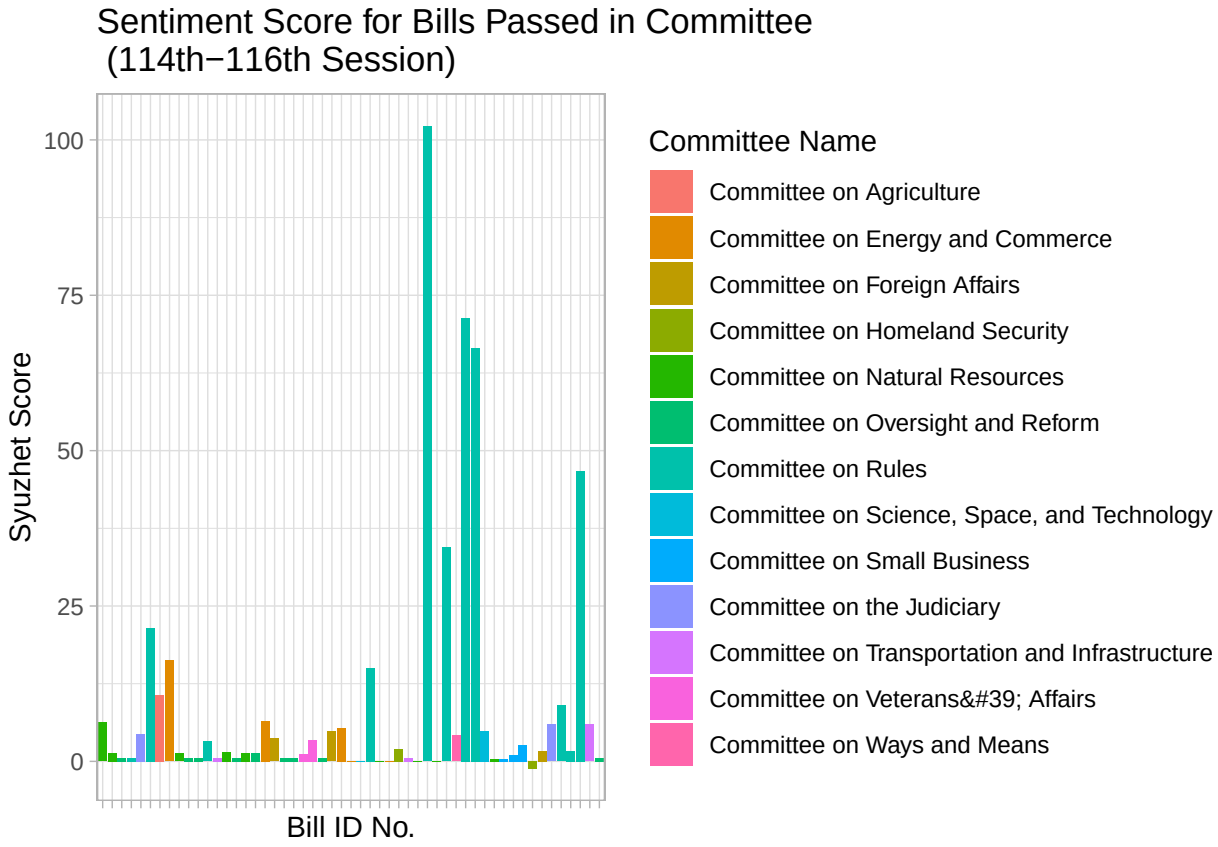
Combining the full API dataset on congressional hearings with one of bills that were passed and a sentiment analysis reveals that the committee most likely to pass bills is the Committee on Rules.

```
gov_df <- gov_df %>%
  rename("bill_id" = "bill_ids")

gov_bills <- gov_df %>% filter(bill_id != "n/a")
bills_df <- as.data.frame(bills_df)
```

```
joined_df <- inner_join(gov_bills, bills_df, by = "bill_id")
```

The chart created also shows that bills passed have overwhelmingly positive sentiments for their summaries.

```
joined_df$syuzhet_vector <- get_sentiment(joined_df$summary, method="syuzhet")

ggplot(joined_df, aes(bill_id, syuzhet_vector, fill = committee)) +
  geom_col(show.legend = TRUE) +
  theme_light() +
  ggtitle("Sentiment Score for Bills Passed in Committee \n (114th-116th Session)") +
  xlab("Bill ID No.") +
  ylab("Syuzhet Score") +
  labs(fill = "Committee Name") +
  theme(axis.text.x = element_blank())
```



This preliminary analysis is very limited with regards to the ability to answer my original research question. It was challenging finding ways to use the data from the government API in concert with the original dataset on testimonies when there were no related variables to join the datasets. However, through exploratory analysis, I did find some factors that do lead to "impactful" congressional hearings, most notably the sentiment and "intensity" of the hearing and/or bill description. The next steps for this project would be perhaps to find actual transcripts of the social scientists present to analyze the information shared and their persuasion tactics used. A hypothesis test can be completed on whether or not hearings with an expert present make a difference or not in the ultimate conclusion of the hearing. If we understand what makes a strong testimony, social scientists hoping to make a beneficial social impact can follow these guidelines to make sure their voices are heard in the elected body.

Testimony Analysis

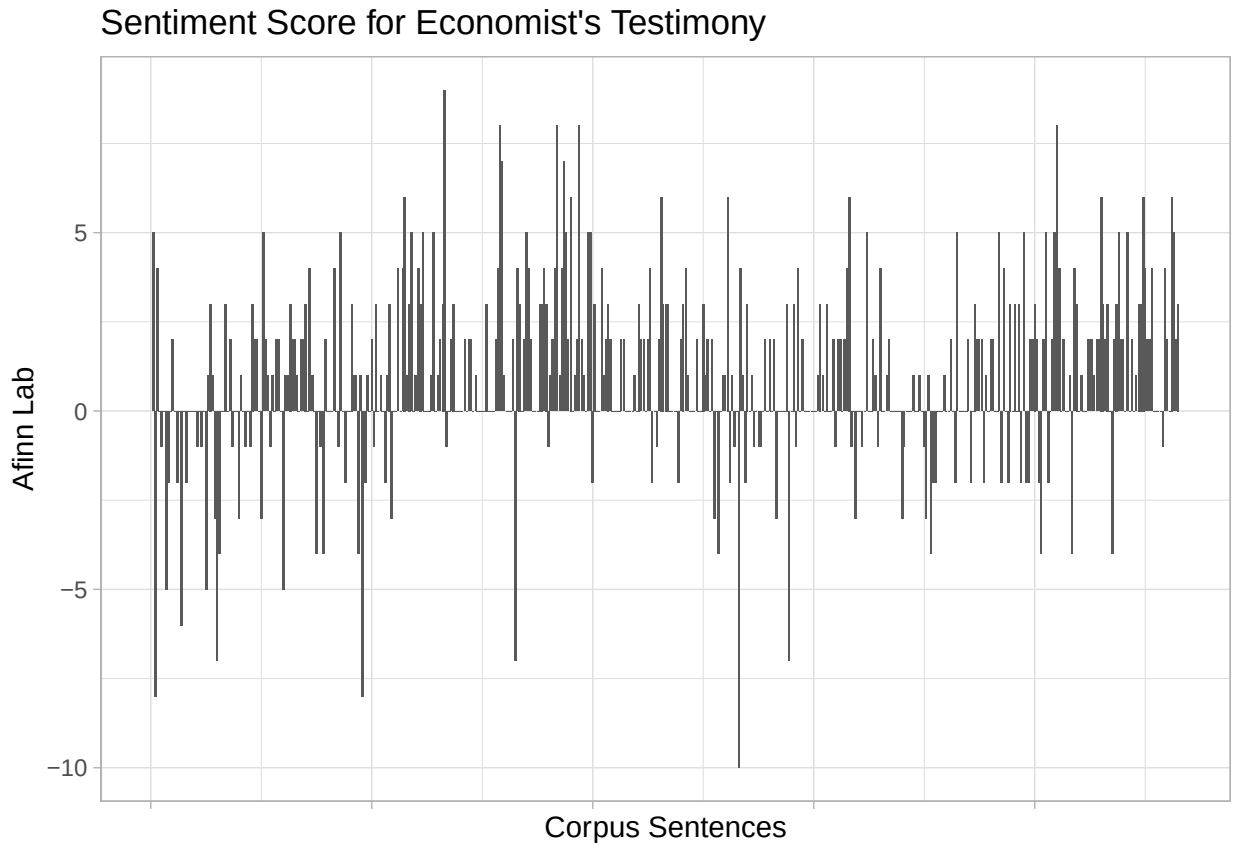
```
library(tokenizers)

### economists

TextDoc <- Corpus(VectorSource(econ$text))
toSpace <- content_transformer(function(x, pattern) gsub(pattern, " ", x))
TextDoc <- tm_map(TextDoc, toSpace, "/")
TextDoc <- tm_map(TextDoc, toSpace, "@")
TextDoc <- tm_map(TextDoc, toSpace, "\\|")
TextDoc <- tm_map(TextDoc, content_transformer(tolower))
TextDoc <- tm_map(TextDoc, removeNumbers)
TextDoc <- tm_map(TextDoc, removeWords, stopwords("english"))
TextDoc <- tm_map(TextDoc, removePunctuation)
TextDoc <- tm_map(TextDoc, stripWhitespace)

# term-document matrix
TextDoc_dtm <- TermDocumentMatrix(TextDoc)
dtm_m <- as.matrix(TextDoc_dtm)

# sort by descearing value of frequency
dtm_v <- sort(rowSums(dtm_m),decreasing=TRUE)
dtm_d <- data.frame(word = names(dtm_v),freq=dtm_v)
dtm_d <- dtm_d %>%
  rename("Word" = "word") %>%
  rename("Frequency" = "freq")

# top 5 most frequent words
datatable(head(dtm_d, 5))
```

```
sentences <- tokenize_sentences(econ$text)
sentence_words <- tokenize_words(sentences[[1]])

v <- c()
for (i in sapply(sentence_words, length)){
  v <- c(v, mean(i))
}

mean(v)
```

```
## [1] 22.00885
```

```
y <- c()
for (i in sentences){
  y <- c(y, get_sentiment(i, method="afinn"))
}

y_df <- data.frame(y) %>%
  mutate(id = row_number())

ggplot(y_df, aes(id, y)) +
  geom_col(show.legend = TRUE) +
  theme_light() +
```

```
  ggtitle("Sentiment Score for Economist's Testimony") +
  xlab("Corpus Sentences") +
  ylab("Afinn Lab") +
  theme(axis.text.x = element_blank())
```



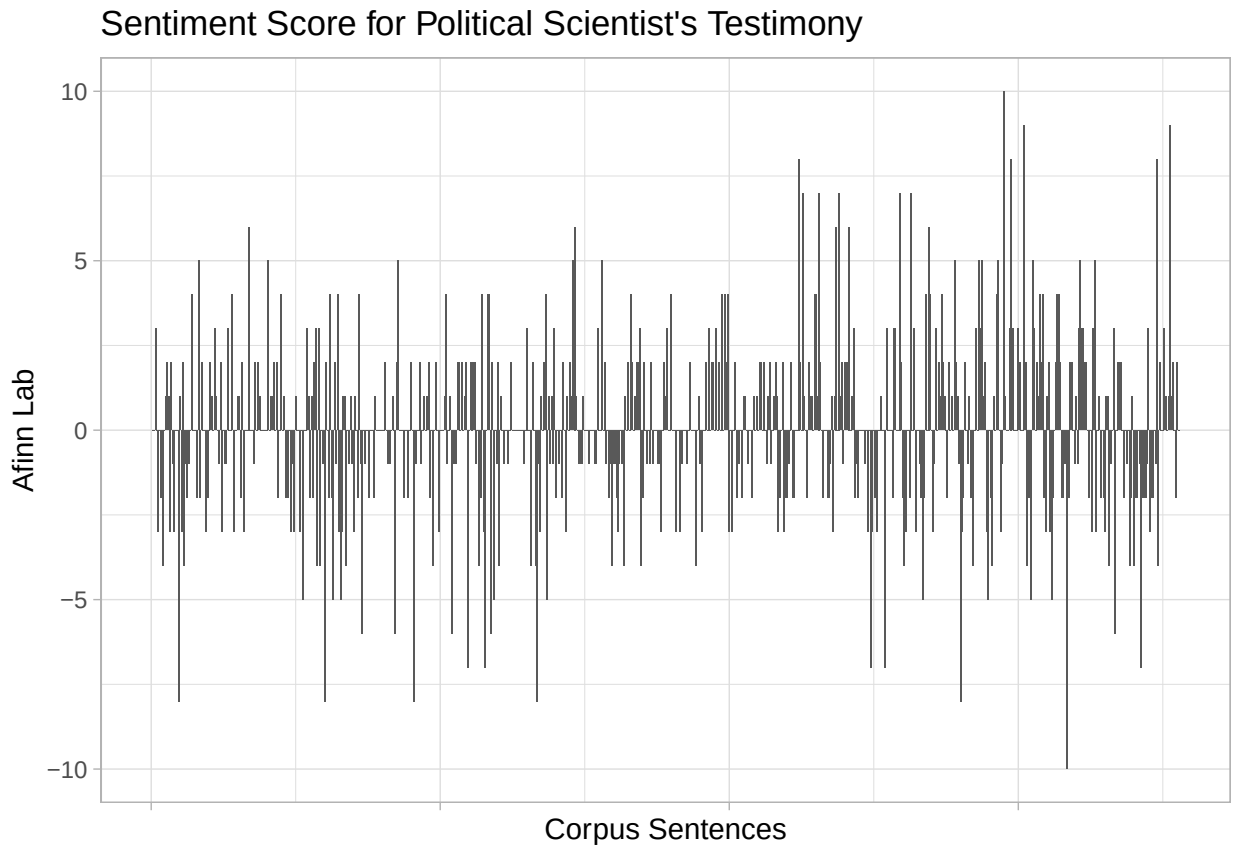

```
### political scientist

TextDoc <- Corpus(VectorSource(polisci$text))
toSpace <- content_transformer(function(x, pattern) gsub(pattern, " ", x))
TextDoc <- tm_map(TextDoc, toSpace, "/")
TextDoc <- tm_map(TextDoc, toSpace, "@")
TextDoc <- tm_map(TextDoc, toSpace, "\\|")
TextDoc <- tm_map(TextDoc, content_transformer(tolower))
TextDoc <- tm_map(TextDoc, removeNumbers)
TextDoc <- tm_map(TextDoc, removeWords, stopwords("english"))
TextDoc <- tm_map(TextDoc, removePunctuation)
TextDoc <- tm_map(TextDoc, stripWhitespace)

# term-document matrix
TextDoc_dtm <- TermDocumentMatrix(TextDoc)
dtm_m <- as.matrix(TextDoc_dtm)

# sort by descearing value of frequency
dtm_v <- sort(rowSums(dtm_m),decreasing=TRUE)
dtm_d <- data.frame(word = names(dtm_v),freq=dtm_v)
dtm_d <- dtm_d %>%
  rename("Word" = "word") %>%
```

```
  rename("Frequency" = "freq")

# top 5 most frequent words
datatable(head(dtm_d, 5))
```

```
sentences <- tokenize_sentences(polisci$text)
sentence_words <- tokenize_words(sentences[[1]])

v <- c()
for (i in sapply(sentence_words, length)){
  v <- c(v, mean(i))
}

mean(v)
```

```
## [1] 23.9
```

```
y <- c()
for (i in sentences){
  y <- c(y, get_sentiment(i, method="afinn"))
}

y_df <- data.frame(y) %>%
  mutate(id = row_number())

ggplot(y_df, aes(id, y)) +
  geom_col(show.legend = TRUE) +
  theme_light() +
  ggtitle("Sentiment Score for Political Scientist's Testimony") +
  xlab("Corpus Sentences") +
  ylab("Afinn Lab") +
  theme(axis.text.x = element_blank())
```

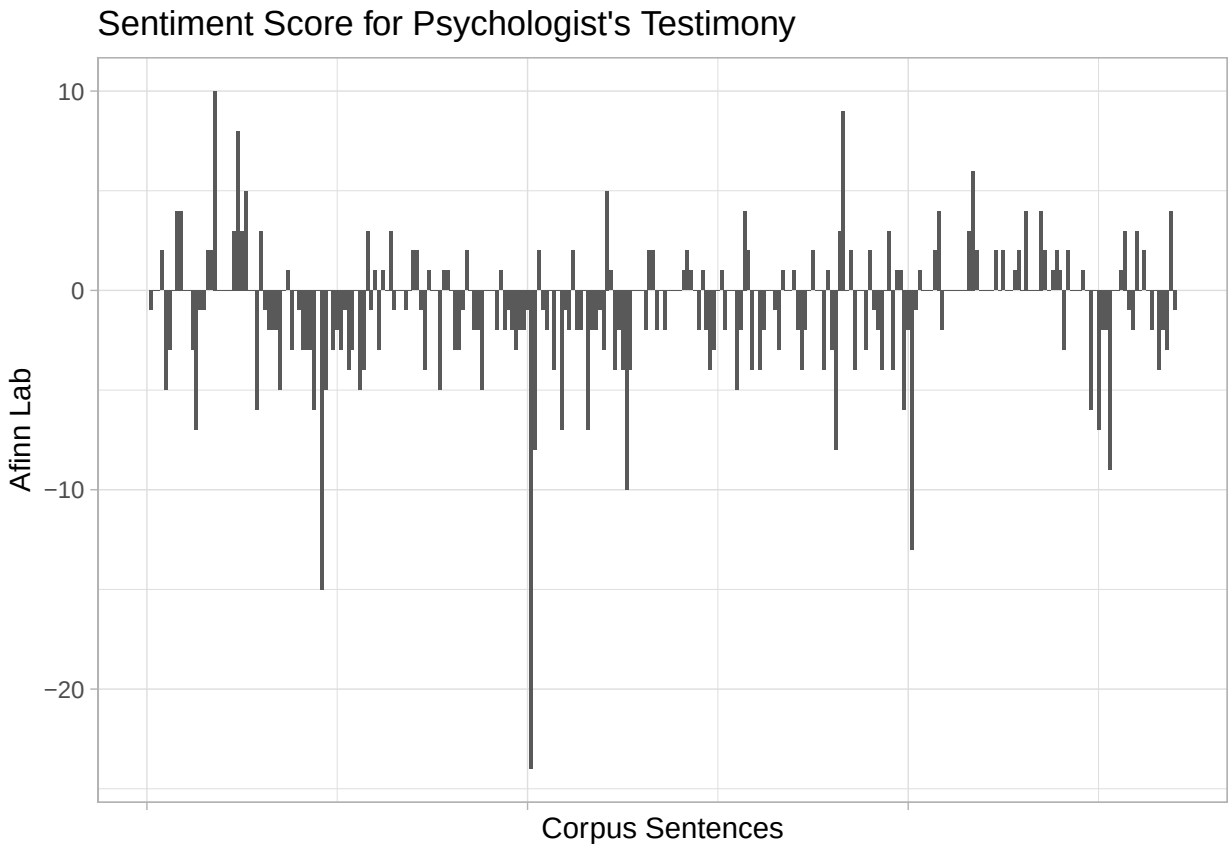



```
### psychologist

TextDoc <- Corpus(VectorSource(psych$text))
toSpace <- content_transformer(function(x, pattern) gsub(pattern, " ", x))
TextDoc <- tm_map(TextDoc, toSpace, "/")
TextDoc <- tm_map(TextDoc, toSpace, "@")
TextDoc <- tm_map(TextDoc, toSpace, "\\|")
TextDoc <- tm_map(TextDoc, content_transformer(tolower))
TextDoc <- tm_map(TextDoc, removeNumbers)
TextDoc <- tm_map(TextDoc, removeWords, stopwords("english"))
TextDoc <- tm_map(TextDoc, removePunctuation)
TextDoc <- tm_map(TextDoc, stripWhitespace)

# term-document matrix
TextDoc_dtm <- TermDocumentMatrix(TextDoc)
dtm_m <- as.matrix(TextDoc_dtm)

# sort by descearing value of frequency
dtm_v <- sort(rowSums(dtm_m),decreasing=TRUE)
dtm_d <- data.frame(word = names(dtm_v),freq=dtm_v)
dtm_d <- dtm_d %>%
  rename("Word" = "word") %>%
  rename("Frequency" = "freq")

# top 5 most frequent words
datatable(head(dtm_d, 5))
```

```
sentences <- tokenize_sentences(psych$text)
sentence_words <- tokenize_words(sentences[[1]])

v <- c()
for (i in sapply(sentence_words, length)){
  v <- c(v, mean(i))
}

mean(v)
```

```
## [1] 20.89286
```

```
y <- c()
for (i in sentences){
  y <- c(y, get_sentiment(i, method="afinn"))
}

y_df <- data.frame(y) %>%
  mutate(id = row_number())

ggplot(y_df, aes(id, y)) +
  geom_col(show.legend = TRUE) +
  theme_light() +
  ggtitle("Sentiment Score for Psychologist's Testimony") +
  xlab("Corpus Sentences") +
  ylab("Afinn Lab") +
  theme(axis.text.x = element_blank())
```

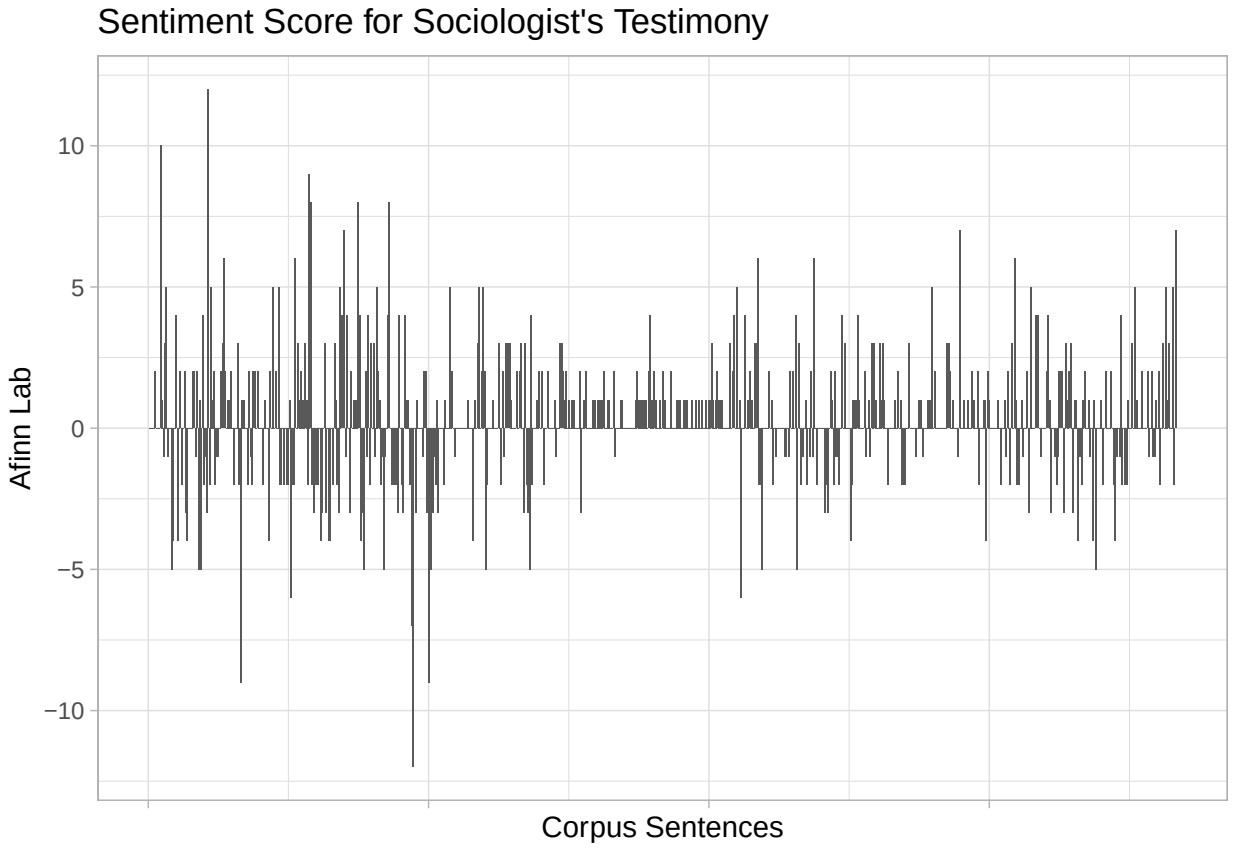


```
### sociologist

TextDoc <- Corpus(VectorSource(soc$text))
toSpace <- content_transformer(function(x, pattern) gsub(pattern, " ", x))
TextDoc <- tm_map(TextDoc, toSpace, "/")
TextDoc <- tm_map(TextDoc, toSpace, "@")
TextDoc <- tm_map(TextDoc, toSpace, "\\|")
TextDoc <- tm_map(TextDoc, content_transformer(tolower))
TextDoc <- tm_map(TextDoc, removeNumbers)
TextDoc <- tm_map(TextDoc, removeWords, stopwords("english"))
TextDoc <- tm_map(TextDoc, removePunctuation)
TextDoc <- tm_map(TextDoc, stripWhitespace)

# term-document matrix
TextDoc_dtm <- TermDocumentMatrix(TextDoc)
dtm_m <- as.matrix(TextDoc_dtm)

# sort by descearing value of frequency
dtm_v <- sort(rowSums(dtm_m),decreasing=TRUE)
dtm_d <- data.frame(word = names(dtm_v),freq=dtm_v)
dtm_d <- dtm_d %>%
  rename("Word" = "word") %>%
  rename("Frequency" = "freq")

# top 5 most frequent words
datatable(head(dtm_d, 5))
```

```
sentences <- tokenize_sentences(soc$text)
sentence_words <- tokenize_words(sentences[[1]])

v <- c()
for (i in sapply(sentence_words, length)){
  v <- c(v, mean(i))
}

mean(v)
```

```
## [1] 26.05063
```

```
y <- c()
for (i in sentences){
  y <- c(y, get_sentiment(i, method="afinn"))
}

y_df <- data.frame(y) %>%
  mutate(id = row_number())

ggplot(y_df, aes(id, y)) +
  geom_col(show.legend = TRUE) +
  theme_light() +
  ggtitle("Sentiment Score for Sociologist's Testimony") +
  xlab("Corpus Sentences") +
  ylab("Afinn Lab") +
  theme(axis.text.x = element_blank())
```



NEXT STEPS (02/25/21):

-Congressional testimony transcripts analysis?

> Example of transcript from a key witness: https://congressional-proquest-com.ezproxy.cul.columbia.edu/congressional/result/congressional/congdocumentview?accountid=10226&groupid=106481&parmId=177405BB234&rsId=177405B66AD#Witness
>
> What makes a great testimony? Scrape the testimonies from the most commonly invited witnesses and the least popular witnesses and compare their testimonies

-More robust models, what are my variables? What is my ultimate research question? What are the outcome variables? What can add to my proposal to continue working on this project?

> Who is the audience? Is it social scientists hoping to make testimonies?

-Ways to improve: Data collection is tricky, better scraping techniques for congressional testimonies

-Presentation: Powerpoint? Google Slides?

Citations:

https://www-tandfonline-com.ezproxy.cul.columbia.edu/doi/full/10.1080/09644016.2019.1565463

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7094826/

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0230104