pset3

Maria Neely

11/19/2019

Question 1

```
platforms <- read.csv("~/Desktop/macs-405/problem-set-3/platforms.csv", header = TRUE)
```

```
## Warning in read.table(file = file, header = header, sep = sep, quote =
## quote, : incomplete final line found by readTableHeader on '~/Desktop/
## macs-405/problem-set-3/platforms.csv'
```

```
library(readr)
library(tm)
```

```
## Loading required package: NLP
```

```
corpus <- file.path("~", "Desktop", "macs-405", "problem-set-3", "Party Platforms Data")
dir(corpus)
```

```
## [1] "d16.txt" "r16.txt"
```

```
platforms <- VCorpus(DirSource(corpus))
```

Question 2

```
library(tidyverse)
```

```
## Warning: As of rlang 0.4.0, dplyr must be at least version 0.8.0.
## x dplyr 0.7.8 is too old for rlang 0.4.1.
## i Please update dplyr with `install.packages("dplyr")`.

## -- Attaching packages --------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v ggplot2 3.1.0     v forcats 0.3.0

## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x ggplot2::annotate() masks NLP::annotate()
## x dplyr::filter()     masks stats::filter()
## x dplyr::lag()        masks stats::lag()
```

```
library(tidytext)
library(tm)
library(dplyr)

#remove punctuation
tidy_platforms <- tm_map(platforms, removePunctuation)

#remove special characters
for (j in seq(tidy_platforms)) {
  tidy_platforms[[j]] <- gsub("/", " ", tidy_platforms[[j]])
  #tidy_platforms[[j]] <- gsub(",", " ", tidy_platforms[[j]])
  #tidy_platforms[[j]] <- gsub('"', " ", tidy_platforms[[j]])
```

```r
  tidy_platforms[[j]] <- gsub("'", " ", tidy_platforms[[j]])
  tidy_platforms[[j]] <- gsub("-", " ", tidy_platforms[[j]])
  tidy_platforms[[j]] <- gsub("\\|", " ", tidy_platforms[[j]])
  tidy_platforms[[j]] <- gsub("@", " ", tidy_platforms[[j]])
  tidy_platforms[[j]] <- gsub("\u2028", " ", tidy_platforms[[j]])  # an ascii character that does not t
}

#remove numbers
tidy_platforms <- tm_map(tidy_platforms, removeNumbers)

#remove uppercase
tidy_platforms <- tm_map(tidy_platforms, tolower)
(tidy_platforms <- tm_map(tidy_platforms, PlainTextDocument)) #redefine
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 2
```

```r
#remove stopwords
tidy_platforms <- tm_map(tidy_platforms,
               removeWords,
               stopwords("english"))
tidy_platforms <- tm_map(tidy_platforms, PlainTextDocument) #redefine

#remove also
tidy_platforms <- tm_map(tidy_platforms, removeWords, c("also"))
tidy_platforms <- tm_map(tidy_platforms, PlainTextDocument) #redefine

#get rid of white space
tidy_platforms <- tm_map(tidy_platforms, stripWhitespace)
tidy_platforms <- tm_map(tidy_platforms, PlainTextDocument)

#check the corpus
#writeLines(as.character(tidy_platforms[1]))
```

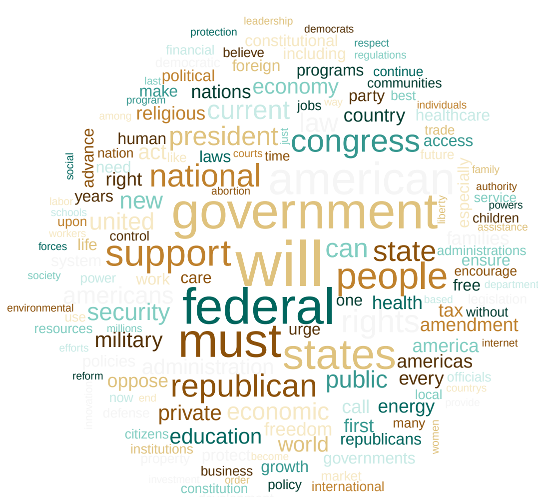
Question 3

```r
library("wordcloud")
```

```
## Loading required package: RColorBrewer
```

```r
#democrats
set.seed(123)
wordcloud(as.character(tidy_platforms[1]),
          max.words = 150,
          random.order=FALSE,
          colors = brewer.pal(11, "BrBG"),
          random.color = TRUE)
```


```
#republicans
set.seed(123)
wordcloud(as.character(tidy_platforms[2]),
          max.words = 150,
          random.order=FALSE,
          scale=c(2, .1),
          colors = brewer.pal(11, "BrBG"),
          random.color = TRUE)
```



From our wordclouds, we can see the most common words Democrats and Republicans. For democrats, most commonly used words include, "will", "democrats", "support" and least commonly used words include common words about our political system such as "states", "party", "institution", and "laws" and words about the economy such as "wealth" and "businesses". For republicans, most commonly used words include "will", "federal", "government", "american", "must", "support", "republican", "rights" and "people". Least commonly used words include "innovation", "protection", "reform", "department" and "women". When comparing the two word clouds, we can see preliminary platform differences. Republicans tend to talk about the federal government, with words such as “federal”, “american”, “government”, “president”, and “congress” being popular. For the Democrats, it is harder to get a sense of their platform based on the wordcloud. It seems that the Democrats’ platform may be more focused on issue areas, as they use frequent words such as “health”, “education”, " affordable“,”communities“,”jobs and “public”. Both parties tend to talk about themselves, with their party name being a frequently used word in each party’s respective word cloud. Both parties also had a common frequent word, “will”. Interestingly, in the Democrats’ wordcloud, there are many more smaller sized words than large sized words than in the Republicans’ wordcloud, in which many more words stand out in terms of largeness in size. This indicates that the Republicans have consistent talking points, causing those words to be more frequent and thus larger in the wordcloud, than the Democrats.

Question 4

```
# tokenize
#tokens <- data_frame(text = as.character(tidy_platforms)) %>% unnest_tokens(word, text)
tokens_dem <- data_frame(text = as.character(tidy_platforms[1])) %>% unnest_tokens(word, text)
```

```
## Warning: `list_len()` is deprecated as of rlang 0.2.0.
## Please use `new_list()` instead.
## This warning is displayed once per session.
```

```
tokens_rep <- data_frame(text = as.character(tidy_platforms[2])) %>% unnest_tokens(word, text)

afinn_dem <- tokens_dem %>%
  inner_join(get_sentiments("afinn"))
```

```
## Joining, by = "word"
```

```
afinn_rep <- tokens_rep %>%
  inner_join(get_sentiments("afinn"))
```

```
## Joining, by = "word"
```

```
bing_dem <- tokens_dem %>%
  inner_join(get_sentiments("bing"))
```

```
## Joining, by = "word"
```

```
bing_rep <- tokens_rep %>%
  inner_join(get_sentiments("bing"))
```

```
## Joining, by = "word"
```

```
#afinn <- tokens %>%
  #inner_join(get_sentiments("afinn"))

#bing <- tokens %>%
  #inner_join(get_sentiments("bing"))
```

```
#top words for democrats afinn
tokens_dem %>%
  count(word, sort = TRUE) %>%
  inner_join(get_sentiments("afinn"))
```

```
## Warning: The `printer` argument is deprecated as of rlang 0.3.0.
## This warning is displayed once per session.
```

```
## Joining, by = "word"
```

```
## # A tibble: 441 x 3
##    word          n score
```

```
##    <chr>     <int> <int>
##  1 support     123     2
##  2 care         66     2
##  3 fight        58    -1
##  4 ensure       50     1
##  5 protect      46     1
##  6 help         41     2
##  7 united       39     1
##  8 committed    36     1
##  9 clean        33     2
## 10 expand       32     1
## # ... with 431 more rows
```

```r
#so all but 1 of top ten words are positive

#top words for rep afinn
tokens_rep %>%
  count(word, sort = TRUE) %>%
  inner_join(get_sentiments("afinn"))
```

```
## Joining, by = "word"
```

```
## # A tibble: 630 x 3
##    word          n score
##    <chr>     <int> <int>
##  1 support     100     2
##  2 united       58     1
##  3 freedom      42     2
##  4 protect      38     1
##  5 care         37     2
##  6 free         37     1
##  7 growth       36     2
##  8 ensure       35     1
##  9 encourage    30     2
## 10 best         29     3
## # ... with 620 more rows
```

```r
#all top ten words are positive


#top words for democrats bing
tokens_dem %>%
  count(word, sort = TRUE) %>%
  inner_join(get_sentiments("bing"))
```

```
## Joining, by = "word"
```

```
## # A tibble: 607 x 3
##    word           n sentiment
##    <chr>      <int> <chr>
##  1 support      123 positive
##  2 work          72 positive
##  3 protect       46 positive
##  4 right         37 positive
##  5 clean         33 positive
##  6 affordable    27 positive
##  7 well          25 positive
```

```
##   8 strong        24 positive
##   9 trump         24 positive
##  10 better        21 positive
## # ... with 597 more rows
```

```r
#so all but 1 of top ten words are positive

#top words for rep bing
tokens_rep %>%
  count(word, sort = TRUE) %>%
  inner_join(get_sentiments("bing"))
```

```
## Joining, by = "word"
```

```
## # A tibble: 898 x 3
##    word          n sentiment
##    <chr>     <int> <chr>
##  1 support     100 positive
##  2 right        46 positive
##  3 oppose       43 negative
##  4 freedom      42 positive
##  5 protect      38 positive
##  6 free         37 positive
##  7 work         37 positive
##  8 encourage    30 positive
##  9 best         29 positive
## 10 like         28 positive
## # ... with 888 more rows
```

```r
#so all but one of the top ten words are positive

#afinn mean analysis
mean_afinn_dem <- mean(afinn_dem$score)
mean_afinn_dem
```

```
## [1] 0.562851
```

```r
mean_afinn_rep <- mean(afinn_rep$score)
mean_afinn_rep
```

```
## [1] 0.3540724
```

```r
#bing mean analysis, convert to binary 0 = negative, 1 = positive
bing_dem <- bing_dem %>%
  mutate(sentiment = recode(sentiment,
                    "negative" = "0",
                    "positive" = "1"))

bing_dem$sentiment <- as.numeric(as.character(bing_dem$sentiment))

mean_bing_dem <- mean(bing_dem$sentiment)
mean_bing_dem
```

```
## [1] 0.6284929
```

```r
bing_rep <- bing_rep %>%
  mutate(sentiment = recode(sentiment,
                    "negative" = "0",
```

```r
                      "positive" = "1"))
bing_rep$sentiment <- as.numeric(as.character(bing_rep$sentiment))

mean_bing_rep <- mean(bing_rep$sentiment)
mean_bing_rep
```

```
## [1] 0.5588235
```

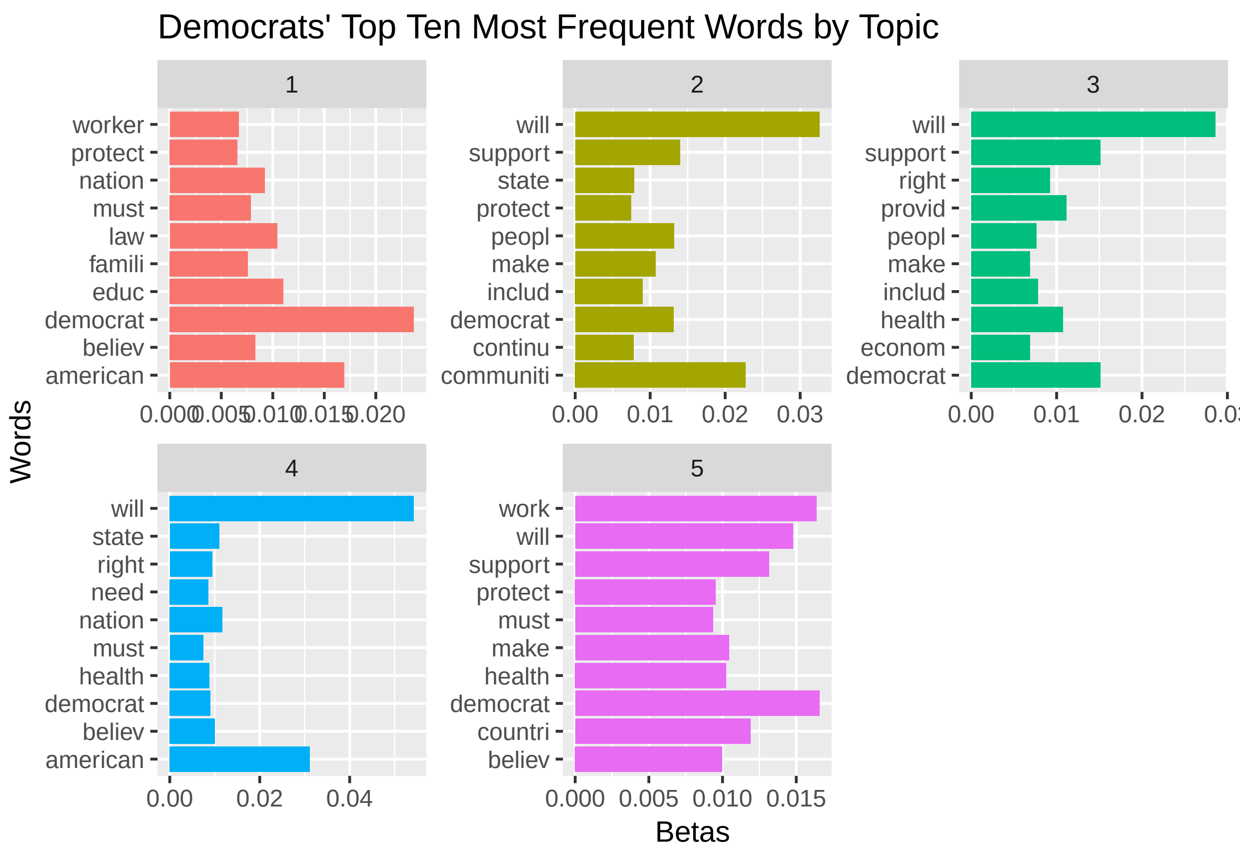
Question 5

After performing sentiment analysis using both the Bing and AFINN dictionaries, we see that the Democratic party is, on average, more positive than the republican party, based on mean calculations. Interestingly, for both parties, most (if not all) of the top ten most common words used are positive. When using the bing dictionary, all of the ten most frequent words for democrats and all but one of the ten most frequent words for republicans are of positive sentiments. When using the afinn dictionary, all but one of the ten most frequent words for democrats and all of the ten most frequent words for republicans are of positive sentiments. This indicates that the most frequent words used by both parties are positive, which makes sense given each parties' campaigning and re-election incentives. It generally is not a strong election strategy to use negative words frequently. However, the true difference is sentiment between the parties is observed when considering the average sentiment score of each party.

Question 6

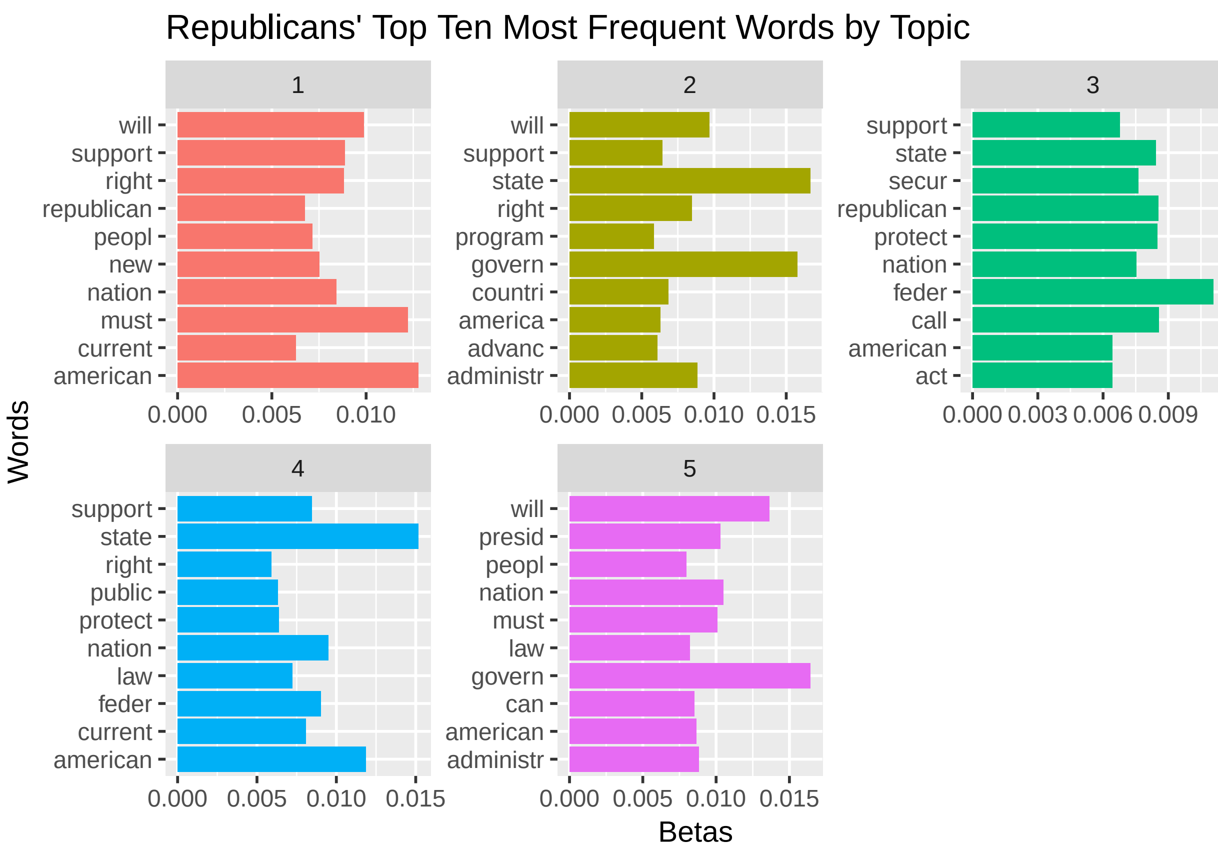
```r
library(topicmodels)
tidy_platforms_dem <- tidy_platforms[1]
tidy_platforms_rep <- tidy_platforms[2]

stem_democrat <- tm_map(tidy_platforms_dem,stemDocument)
stem_democrat<- tm_map(stem_democrat, PlainTextDocument)


stem_republican <- tm_map(tidy_platforms_rep, stemDocument)
stem_republican<- tm_map(stem_republican, PlainTextDocument)


dtm_democrat <- DocumentTermMatrix(stem_democrat)
dtm_republican <- DocumentTermMatrix(stem_republican)

#create topic model
dem_lda <- LDA(dtm_democrat, k = 5, control = list(seed = 1234))


library(tidytext)
tidy_dem_lda <- tidy(dem_lda, matrix = "beta")


library(ggplot2)
library(dplyr)
library(tidyr)

dem_lda_ten_top_terms <- tidy_dem_lda %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
```

```r
dem_lda_ten_top_terms %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  labs(title= "Democrats' Top Ten Most Frequent Words by Topic",
                      y="Betas", x = "Words") +
  coord_flip()
```




```r
#try this out... not sure what it means read about
#beta= topic per word probability
beta_spread <- dem_lda_ten_top_terms %>%
  mutate(topic = paste0("topic", topic)) %>%
  spread(topic, beta) %>%
  filter(topic1 > .001 | topic2 > .001) %>%
  mutate(log_ratio = log2(topic2 / topic1))

#beta_spread_dem

#create topic model
rep_lda <- LDA(dtm_republican, k = 5, control = list(seed = 1234))



tidy_rep_lda <- tidy(rep_lda, matrix = "beta")

rep_lda_ten_top_terms <- tidy_rep_lda %>%
  group_by(topic) %>%
```

```
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

rep_lda_ten_top_terms %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  labs(title= "Republicans' Top Ten Most Frequent Words by Topic",
                    y="Betas", x = "Words") +
  coord_flip()
```



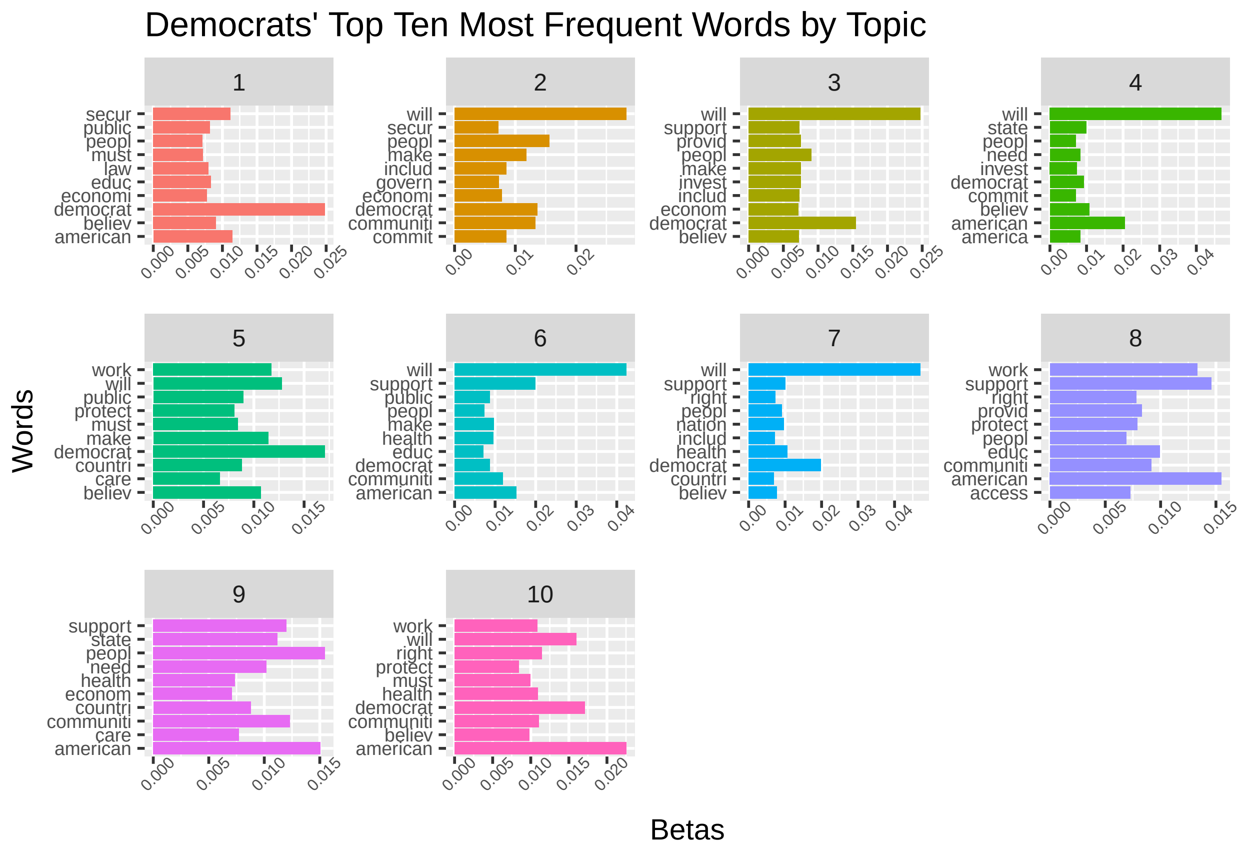

```
#try this out... not sure what it means read about
beta_spread <- rep_lda_ten_top_terms %>%
  mutate(topic = paste0("topic", topic)) %>%
  spread(topic, beta) %>%
  filter(topic1 > .001 | topic2 > .001) %>%
  mutate(log_ratio = log2(topic2 / topic1))

#beta_spread_rep
```

Question 7

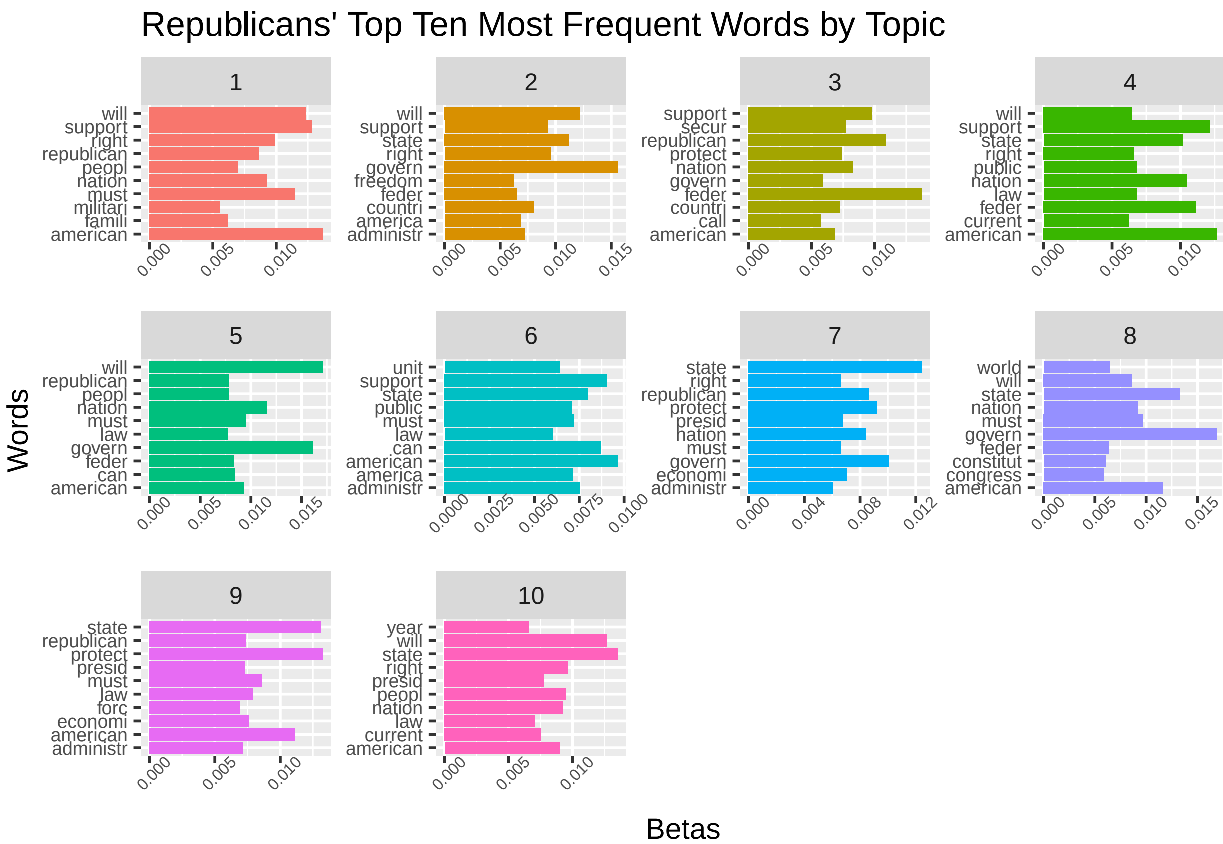
After graphing the topic models for each party, we see that the parties are focused on different topics generally. Democrats focus on the people in their topics, while Republicans focus on the government in their topics. Similar to trends seen when comparing the wordclouds of the two parties, we see that Democrats seem to have topic areas around platform issues, and the idea of protecting or supporting the people in this issue area. This is based on the inclusion of "protect" or "support" in each topic, as well as the inclusion of various hot topics, such as "health", "workers", "education", or "econom" in each topic. Republicans, on the other hand, have topics centered around federal or state functions, with common words such as "nation", "state", "adminstr", "feder" and "law" included in various topics.

Question 8

```
#create topic model for k = 10
dem_lda <- LDA(dtm_democrat, k = 10, control = list(seed = 1234))


library(tidytext)
tidy_dem_lda <- tidy(dem_lda, matrix = "beta")

library(ggplot2)
library(dplyr)
library(tidyr)

dem_lda_ten_top_terms <- tidy_dem_lda %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

dem_lda_ten_top_terms %>%
  #mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
   theme(axis.text.x =
                   element_text(size  = 6,
                                angle = 45
                                ),
        axis.text.y = element_text(size  = 7
                                )) +
   labs(title= "Democrats' Top Ten Most Frequent Words by Topic",
                      y="Betas", x = "Words") +
  coord_flip()
```

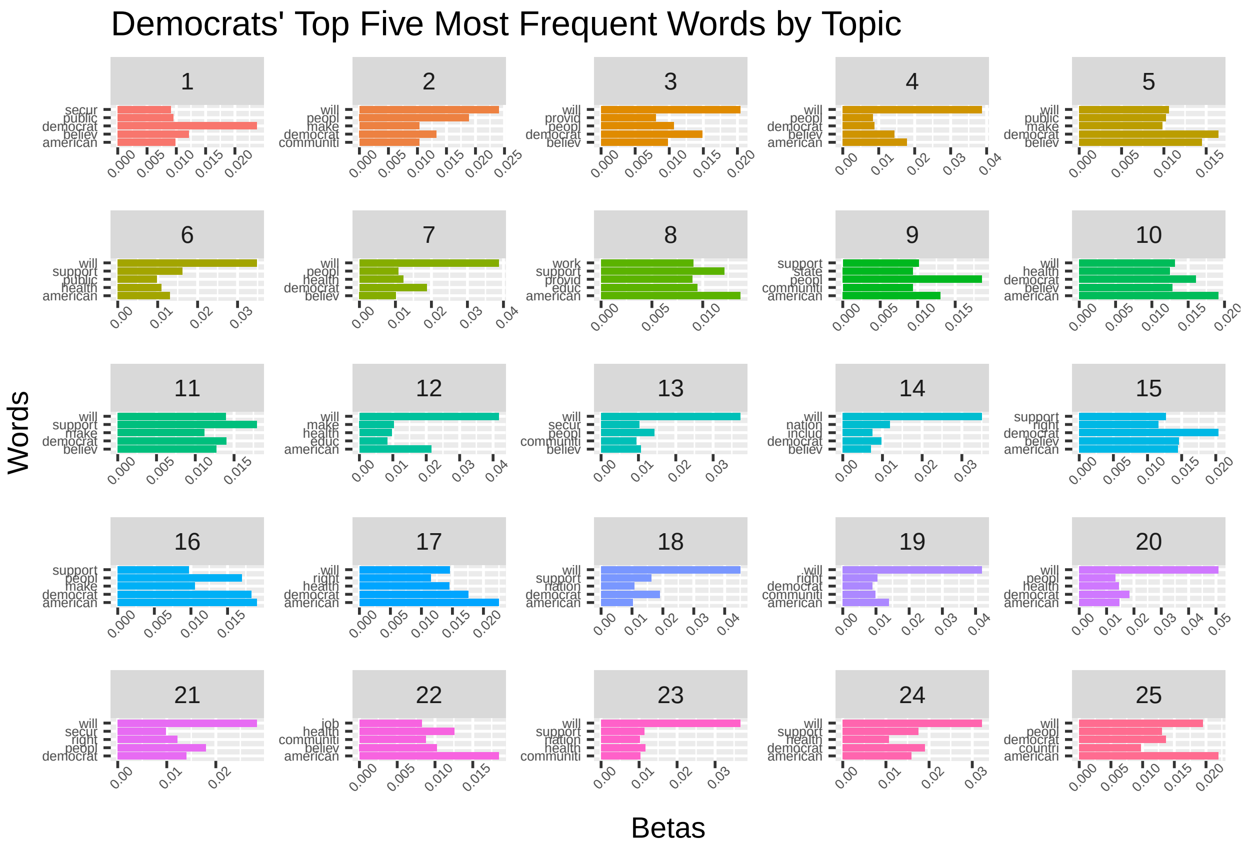



```r
#republicans
rep_lda <- LDA(dtm_republican, k = 10, control = list(seed = 1234))

tidy_rep_lda <- tidy(rep_lda, matrix = "beta")


rep_lda_ten_top_terms <- tidy_rep_lda %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

rep_lda_ten_top_terms %>%
  #mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  theme(axis.text.x =
                  element_text(size  = 6,
                          angle = 45
                          ),
        axis.text.y = element_text(size  = 7
                          )) +
   labs(title= "Republicans' Top Ten Most Frequent Words by Topic",
                 y="Betas", x = "Words") +
  coord_flip()
```



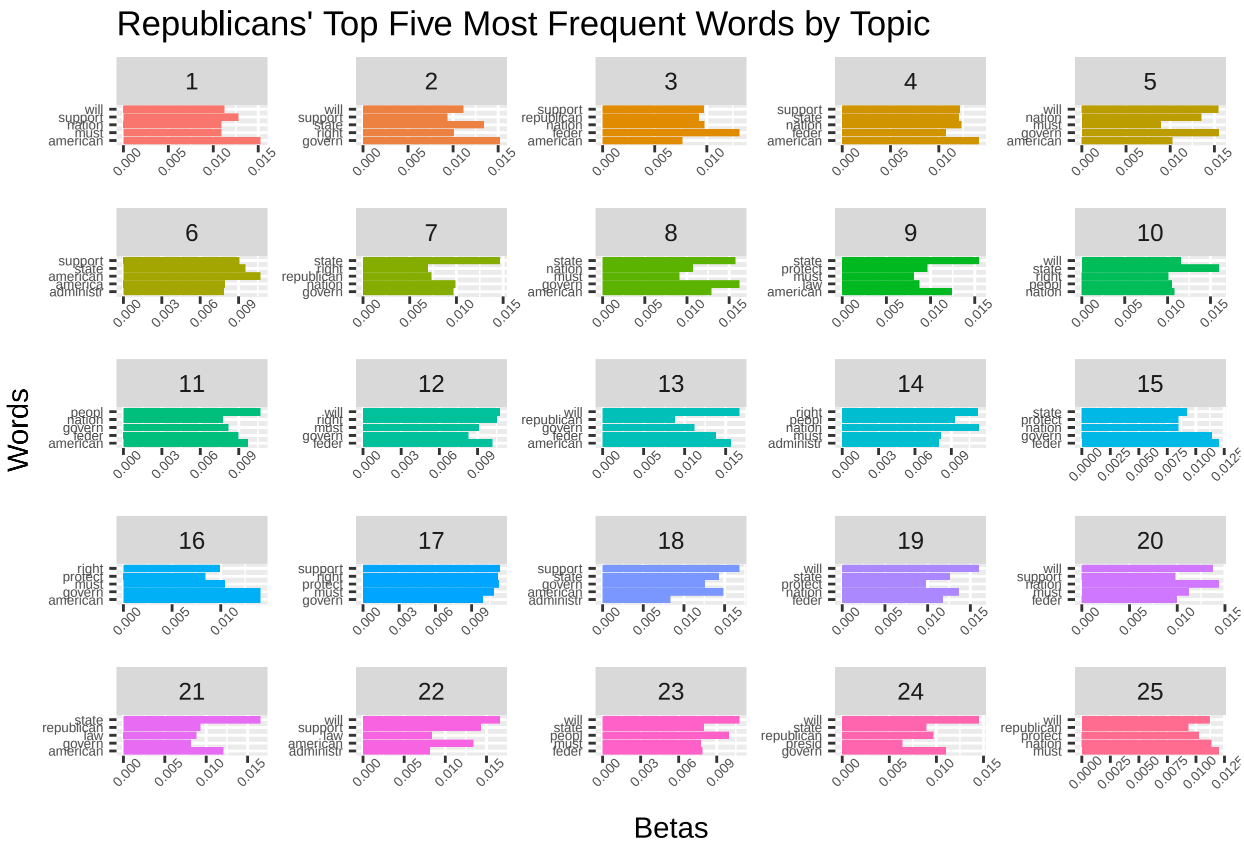

```
#create topic model for k = 25
dem_lda <- LDA(dtm_democrat, k = 25, control = list(seed = 1234))


library(tidytext)
tidy_dem_lda <- tidy(dem_lda, matrix = "beta")

library(ggplot2)
library(dplyr)
library(tidyr)

dem_lda_five_top_terms <- tidy_dem_lda %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

dem_lda_five_top_terms %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
   theme(axis.text.x =
                   element_text(size  = 5,
                                angle = 45
                                ),
        axis.text.y = element_text(size  = 5
                                )) +
```

```r
  labs(title= "Democrats' Top Five Most Frequent Words by Topic",
                     y="Betas", x = "Words") +
  coord_flip()
```




```r
#republicans
rep_lda <- LDA(dtm_republican, k = 25, control = list(seed = 1234))

tidy_rep_lda <- tidy(rep_lda, matrix = "beta")

#do top 5 words because many topics, want it to be readable on my graph
rep_lda_five_top_terms <- tidy_rep_lda %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

rep_lda_five_top_terms %>%
  #mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  theme(axis.text.x =
                  element_text(size  = 5,
                               angle = 45
                               ),
        axis.text.y = element_text(size  = 5
                               )) +
```

```
  labs(title= "Republicans' Top Five Most Frequent Words by Topic",
                  y="Betas", x = "Words") +
  coord_flip()
```



Question 9

```
library(topicmodels)
dem_lda_5 <- LDA(dtm_democrat, k = 5, control = list(seed = 1234))
dem_lda_10 <- LDA(dtm_democrat, k = 10, control = list(seed = 1234))
dem_lda_25 <- LDA(dtm_democrat, k = 25, control = list(seed = 1234))

perplexity(dem_lda_5)
```

```
## [1] 970.6281
```

```
perplexity(dem_lda_10)
```

```
## [1] 971.9451
```

```
perplexity(dem_lda_25)
```

```
## [1] 974.8237
```

```
#so 5 is best for dem

rep_lda_5 <- LDA(dtm_republican, k = 5, control = list(seed = 1234))
rep_lda_10 <- LDA(dtm_republican, k = 10, control = list(seed = 1234))
rep_lda_25 <- LDA(dtm_republican, k = 25, control = list(seed = 1234))

perplexity(rep_lda_5)
```

```
## [1] 1370.624
```

```r
perplexity(rep_lda_10)
```

```
## [1] 1371.664
```

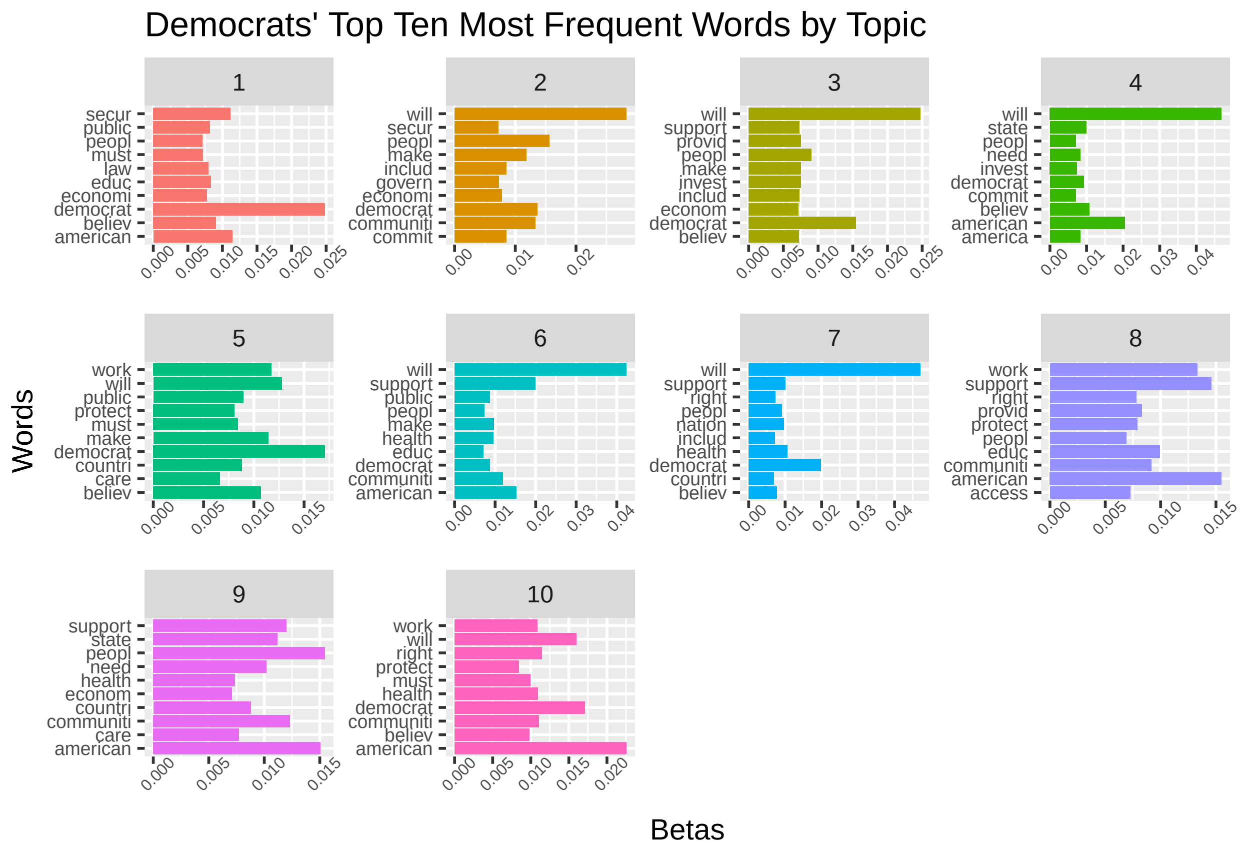
```r
perplexity(rep_lda_25)
```

```
## [1] 1374.173
```

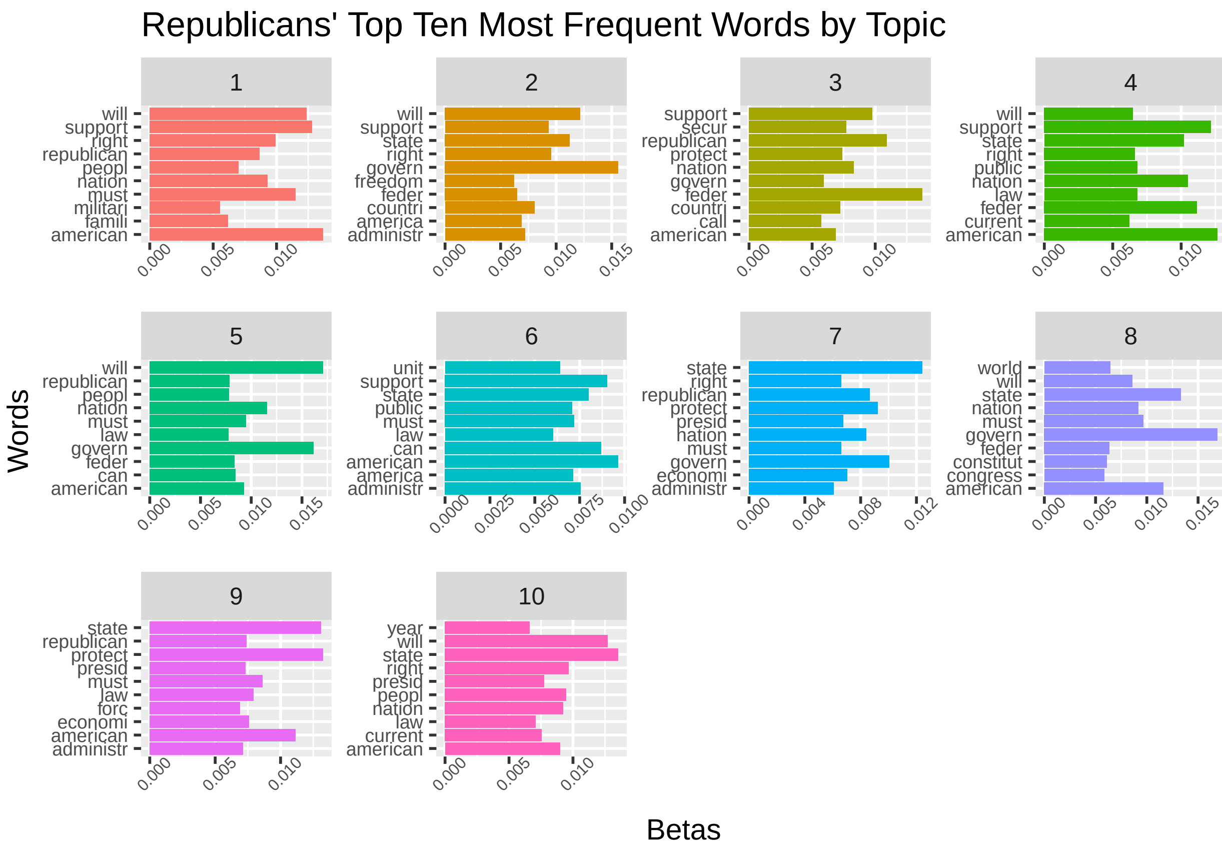
Based on our perplexity score numbers, our model with 5 topics technically fits best for both democrats and republicans, as it has the smallest perplexity score compared to that of the three models for each party.

Question 10

```r
#create topic model for k = 10
dem_lda <- LDA(dtm_democrat, k = 10, control = list(seed = 1234))


library(tidytext)
tidy_dem_lda <- tidy(dem_lda, matrix = "beta")

library(ggplot2)
library(dplyr)
library(tidyr)

dem_lda_ten_top_terms <- tidy_dem_lda %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

dem_lda_ten_top_terms %>%
  #mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
   theme(axis.text.x =
                  element_text(size  = 6,
                               angle = 45
                               ),
        axis.text.y = element_text(size  = 7
                               )) +
   labs(title= "Democrats' Top Ten Most Frequent Words by Topic",
                     y="Betas", x = "Words") +
  coord_flip()
```




```
#republicans
rep_lda <- LDA(dtm_republican, k = 10, control = list(seed = 1234))

tidy_rep_lda <- tidy(rep_lda, matrix = "beta")


rep_lda_ten_top_terms <- tidy_rep_lda %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

rep_lda_ten_top_terms %>%
  #mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  theme(axis.text.x =
                  element_text(size  = 6,
                               angle = 45
                               ),
        axis.text.y = element_text(size  = 7
                               )) +
   labs(title= "Republicans' Top Ten Most Frequent Words by Topic",
                     y="Betas", x = "Words") +
  coord_flip()
```



Examining the topic model with k = 10 for each party, we see similar themes in topic model within the parties. Once again, the Democrats have topics with themes around key issues areas for the people, while the Republicans have topics with themes around the government. I do not think k = 10 picks up the differences efficiently, particularly in the republican model, as there additional words included in both the Democrat and Republican topic models that do not match the apparent themes. For example, topic 10 in the Republicans' topic model includes words such as "year" and "current", which are irrelevant to the theme of government. For the democrats, topic 9 includes "state" and "countri", as well as topic 3 includes "make" and "invest", all of which are irrelevant to the theme of key issues of the people. These graphs reflect the results of our perplexity score analysis nicely.

Question 11

Based on my analysis, I would support the democrats in the 2020 election. This is because, the democratic platform is much more "American public-centered", commonly addressing issues that Americans face daily in the topics they discuss, whereas the Republicans discuss the federal/state government in the topics they discuss. I feel as a voter, I would be more persuaded by hearing about how a party addresses specific issues that concern me rather than by hearing a party discuss our government in a high-level manner, which I do not feel the effects of on my everyday life. In terms of sentiment, the democratic party also is attractive because it is on average, more positive than the republican party, when analyzing with the afinn dictionary and the bing dictionary. I would prefer to vote for a party that is positive about the future of our country than one that is negative about it. Thus, based on the sentiments and topics used by each party, I believe I would support the democratic party in the 2020 elections.