pset5

Bhargavi Ganesh

11/26/2019

Question 1: Load Data

```
texts <- file.path("~", "Documents/Problem-Set-5/Party Platforms Data", "texts")
docs <- VCorpus(DirSource(texts))
```

Question 2: Preprocessing

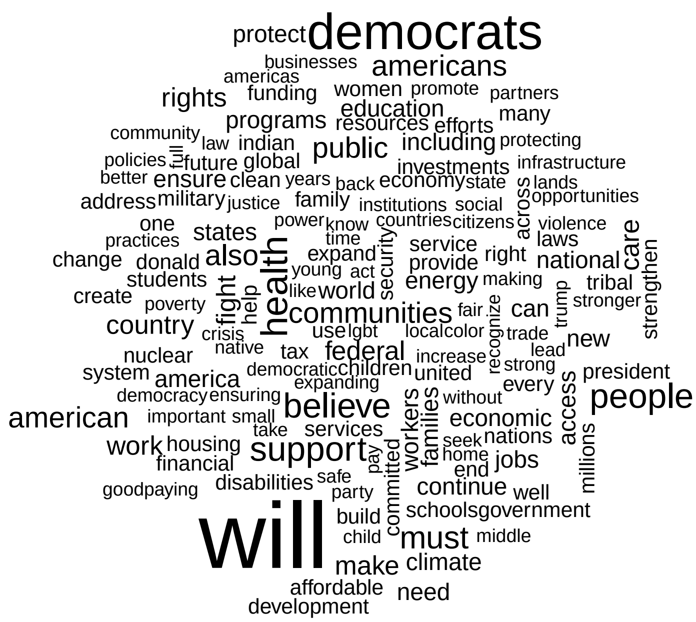
```
#remove punctuation
docs <-  docs %>%
  tm_map(stripWhitespace) %>%
  tm_map(removeNumbers) %>%
  tm_map(removePunctuation) %>%
  tm_map(content_transformer(tolower)) %>%
  tm_map(removeWords, stopwords("english"))
```

Question 3: Creating a document=term matrix for each party

Democrats

Below, I create a document-term matrix and wordcloud for the democratic party platform.

```
dtm_dem <- DocumentTermMatrix(docs[1])

frequency_dem <- sort(colSums(as.matrix(dtm_dem)),
                  decreasing=TRUE)
set.seed(12345)

wordcloud(names(frequency_dem), frequency_dem, min.freq = 2, max.words=150, scale=c(2.9, 0.5))
```



From the wordcloud, it is possible to see certain themes from the Democratic party platform. Some notable interesting words that I see are the words communities, support, believe, americans, health, climate and rights. This supports my general perception of the Democratic party focusing on supporting communities and providing services such as healthcare. It is also interesting that the word "Donald" shows up, because it shows how central Trump was to the democratic party's platform.

Republicans

```
dtm_repub <- DocumentTermMatrix(docs[2])
frequency_repub <- sort(colSums(as.matrix(dtm_repub)),
                   decreasing=TRUE)

set.seed(12345)
wordcloud(names(frequency_repub), frequency_repub, min.freq=2, max.words = 150, scale=c(2, 0.5))
```

By contrast, it is possible to see that the Republican party platform wordcloud includes the words states, rights, federal, national, freedom, and abortion. This supports my priors on what I believed to be central to the Republican Party: states rights, abortion, and freedom. It was interesting that certain words were in common between the Republican and Democratic wordclouds. For example, the word "rights" shows up in both, yet it appears that Republicans are more focused on states' rights whereas Democrats are more focused on the rights of communities and individuals.

Question 4: Calculating sentiment of each party platform.

```
t_corpus <- docs %>%
  tidy()

tidy_df<- t_corpus %>%
  unnest_tokens(word, text)
```

Below, I wanted to see if the number of words included in each platform was similar. I did this check because I knew that with an inner join of the sentiment dictionary, certain words would be dropped, and before comparing sentiments in terms of number of words, I wanted to make sure that each party had a close enough number of words. Below, it appears that the republican text has approximately 700 more words than the democratic text when using the Bing dictionary.

```
tidy_df %>%
  inner_join(get_sentiments("bing")) %>%
  group_by(id) %>%
  summarize(count = n())
```

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

```
## # A tibble: 2 x 2
##   id      count
##   <chr>   <int>
## 1 d16.txt  2183
## 2 r16.txt  2822
```

For the Afinn dictionary, I see that the Republican text has approximately 300 more words than the Democratic text.

```
tidy_df %>%
  inner_join(get_sentiments("afinn")) %>%
  group_by(id) %>%
  summarize(count = n())
```

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

```
## # A tibble: 2 x 2
##   id      count
##   <chr>   <int>
## 1 d16.txt  2315
## 2 r16.txt  2652
```

Question 5: Comparing sentiments across platforms

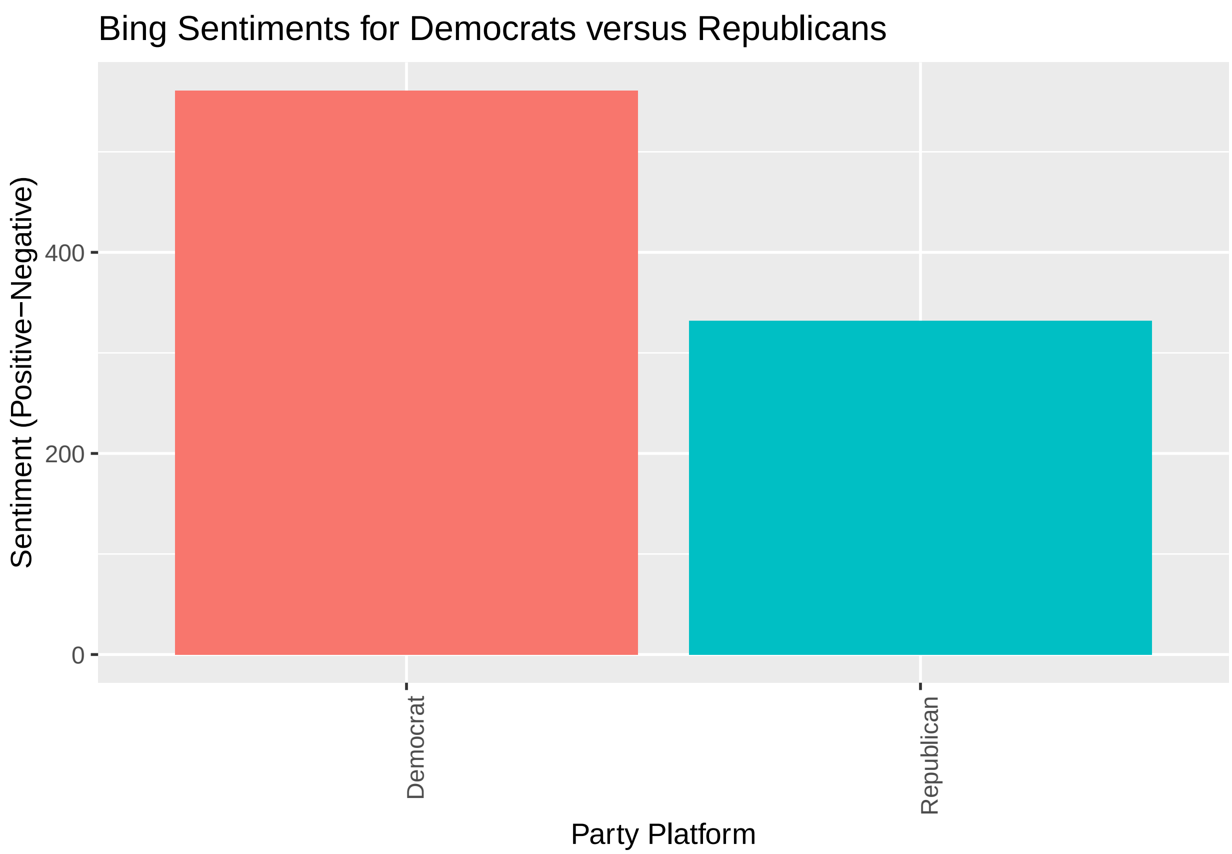
Bing Dictionary

Below, I present a chart of the sentiments for each party, using the Bing dictionary. The y-axis "sentiment" variable represents the difference between the number of words with positive and negative sentiments. It appears that Democrats overall have more words with positive sentiment than Republicans do.

```
tidy_df_sent_bing <- tidy_df %>%
  inner_join(get_sentiments("bing")) %>%
  count(id, sentiment) %>%
  mutate(id = case_when(id == "d16.txt" ~ "Democrat",
                  id == "r16.txt" ~ "Republican")) %>%
  spread(sentiment, n, fill = 0) %>%
  mutate(sentiment = positive - negative)
```

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

```
ggplot(tidy_df_sent_bing, aes(id, sentiment, fill = id)) +
  geom_col(show.legend = FALSE) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ggtitle("Bing Sentiments for Democrats versus Republicans") +
  labs(y="Sentiment (Positive-Negative)", x = "Party Platform")
```



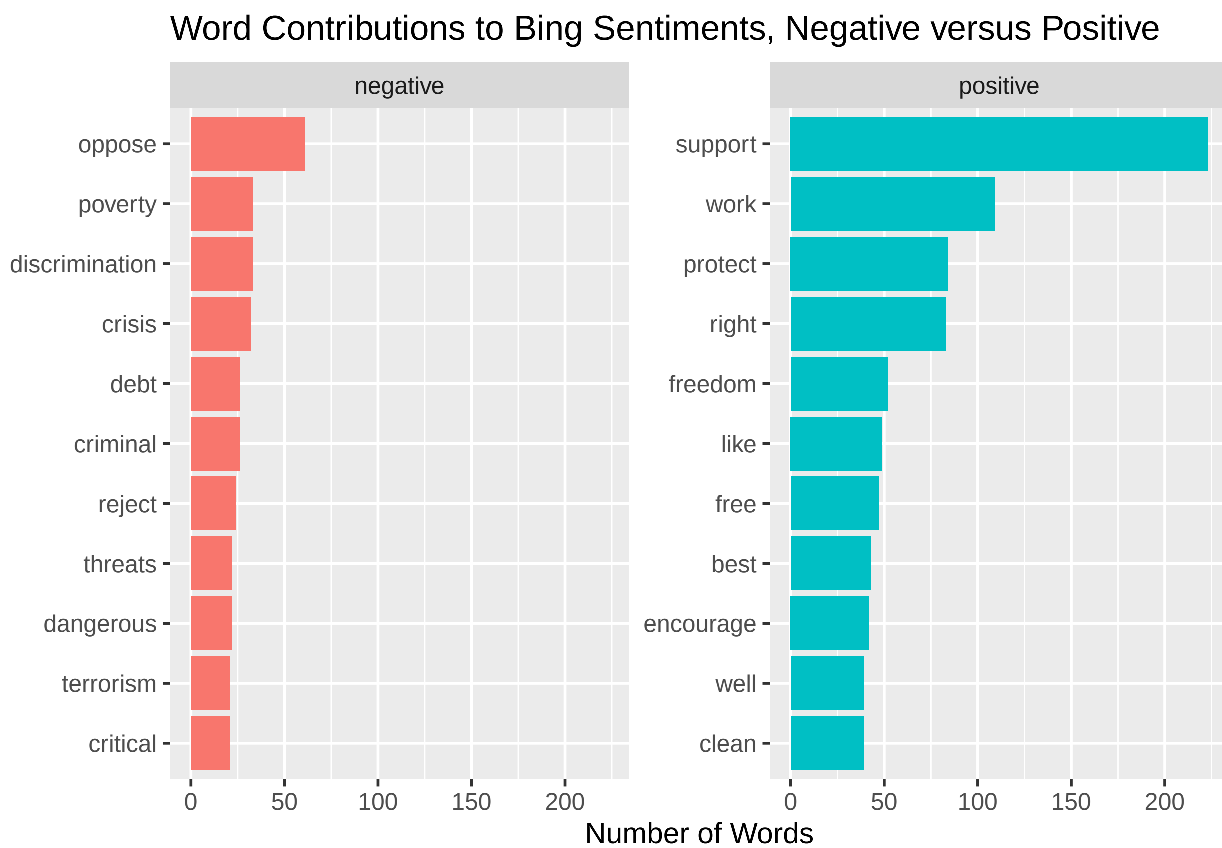
Drilling down further, I chose to see which words were contributing most to the positive and negative sentiments.

```
bing_word_counts <- tidy_df %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort = TRUE) %>%
  ungroup()
```

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

```
bing_word_counts %>%
  group_by(sentiment) %>%
  top_n(10) %>%
  ungroup() %>%
  mutate(word = reorder(word,n)) %>%
  ggplot(aes(word, n, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(y = "Number of Words",
       x = NULL) +
  ggtitle("Word Contributions to Bing Sentiments, Negative versus Positive") +
  coord_flip()
```

```
## Selecting by n
```



The chart above shows that the words oppose and support were the biggest contributors to negative and positive sentiment, respectively.

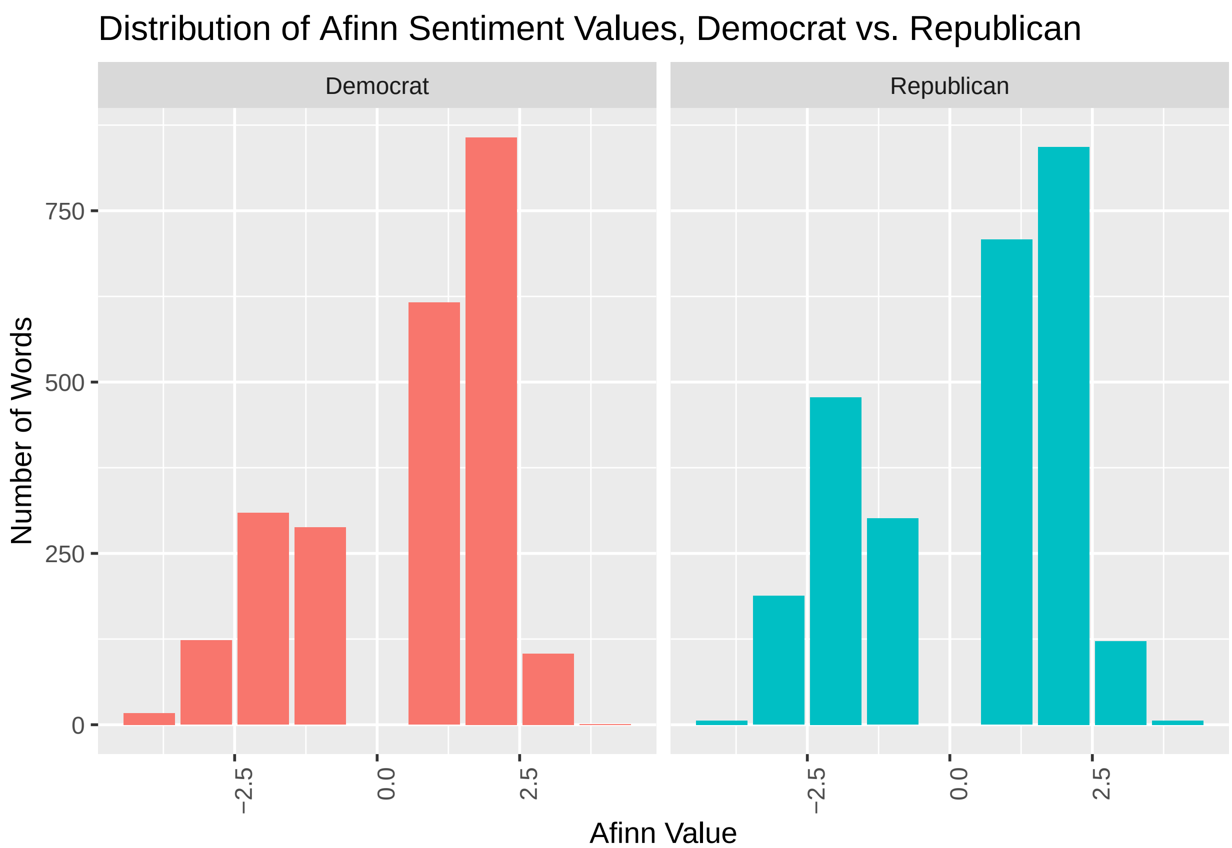
Afinn Dictionary

Below, I plot the number of words associated with each of the different values in the Afinn dictionary.

```
tidy_df_sent_afinn <- tidy_df %>%
  inner_join(get_sentiments("afinn")) %>%
  count(id, value) %>%
  mutate(id = case_when(id == "d16.txt" ~ "Democrat",
                   id == "r16.txt" ~ "Republican"))
```

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

```
ggplot(tidy_df_sent_afinn, aes(value, n, fill = id)) +
  geom_col(show.legend = FALSE) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  facet_wrap(~id) +
  labs(y = "Number of Words",
       x = "Afinn Value") +
  ggtitle("Distribution of Afinn Sentiment Values, Democrat vs. Republican")
```



It is difficult to perceive the differences across values with the chart alone, so I presented the results numerically below.

```
kable(tidy_df_sent_afinn)
```

id	value	n
Democrat	-4	17
Democrat	-3	123
Democrat	-2	309
Democrat	-1	288
Democrat	1	616
Democrat	2	857
Democrat	3	104
Democrat	4	1
Republican	-4	6
Republican	-3	188
Republican	-2	478
Republican	-1	301
Republican	1	708
Republican	2	843
Republican	3	122
Republican	4	6

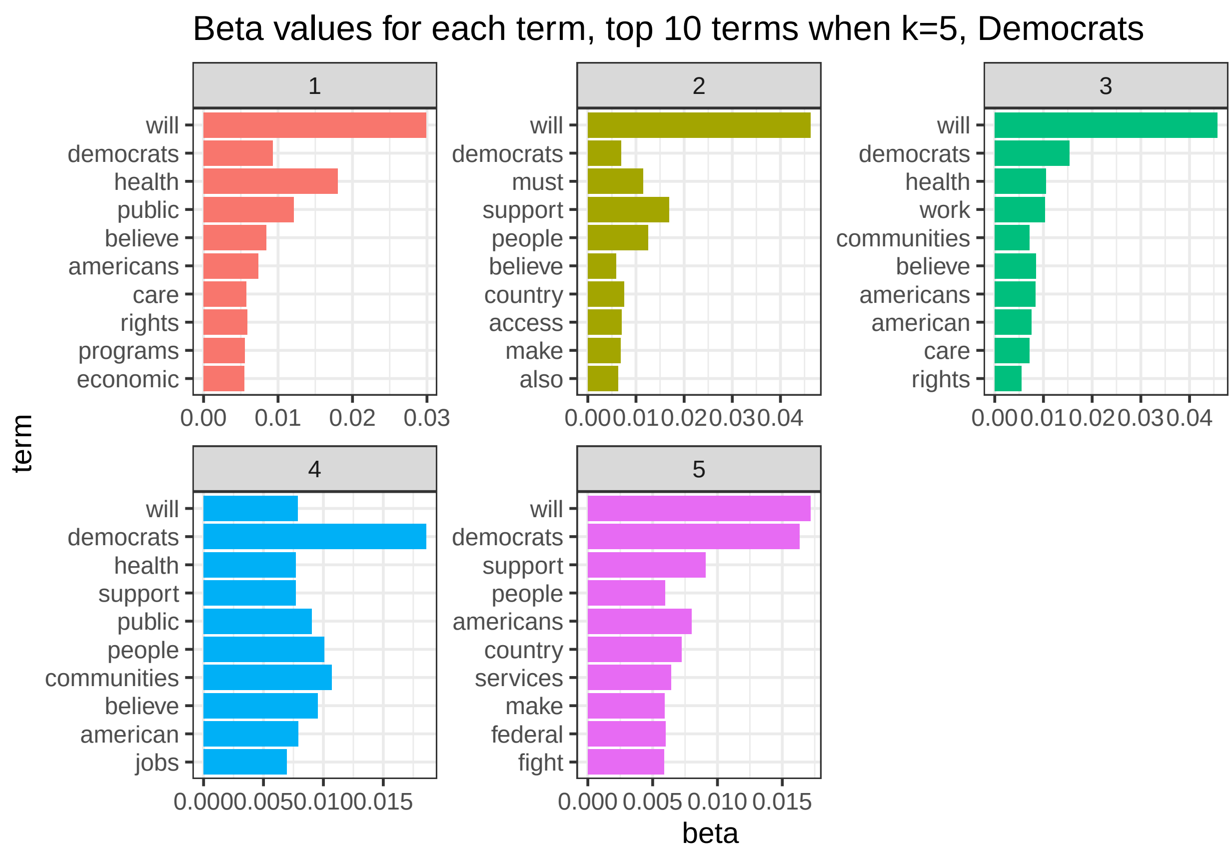
It's possible to see from the numerical representation above that there are 737 negative values for Democrats, compared to 974 negative values for Republicans. This further supports my previous observation that the Republican platform appears to have a generally more negative sentiment.

It is important to note that perhaps the differences in number of words with positive and negative sentiments could be due to the fact that there are some discrepancies in the number of words being compared for the Democratic platform versus the Republican platform. This would be an interesting point to expand upon and study further in the future.

Question 6: Initializing Topic Models with k=5

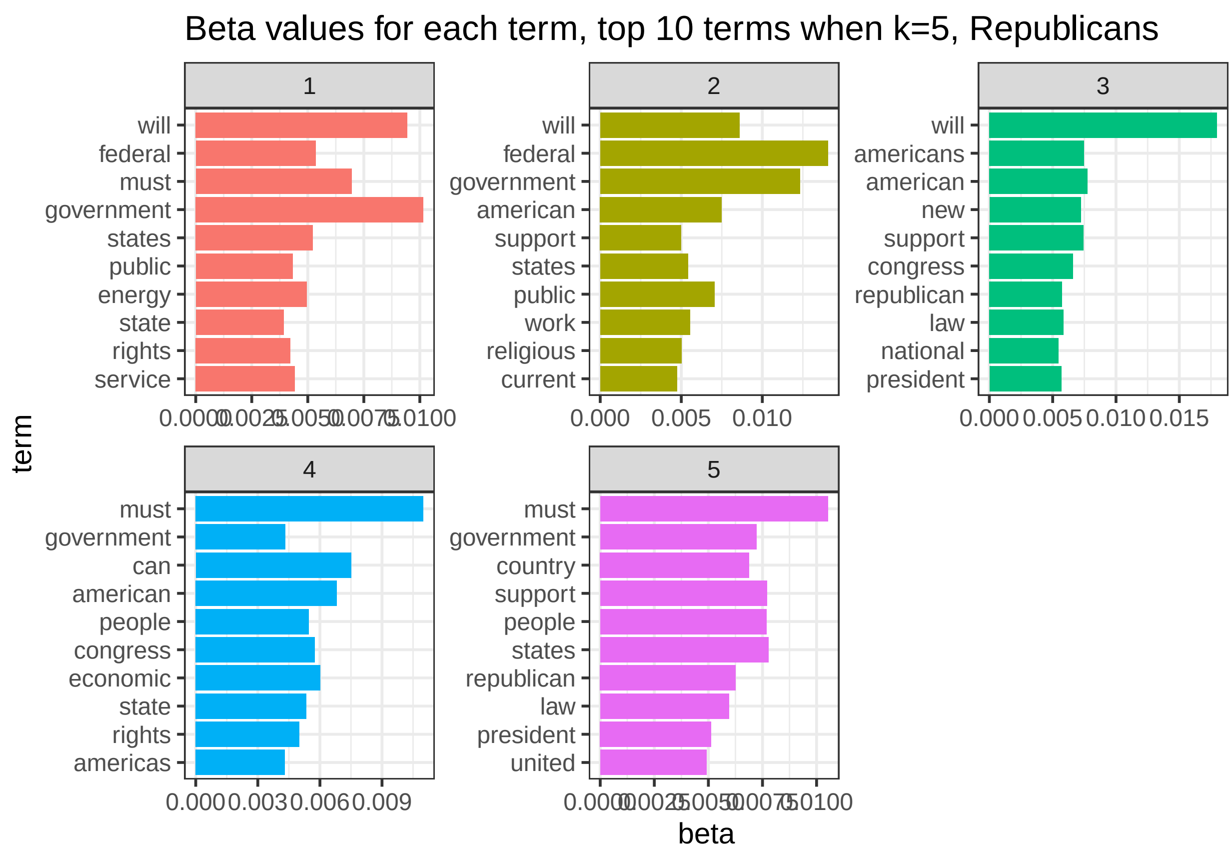
Below, I present a graphic of the top 10 words that appear in each topic model for k=5 in the Democratic party platform.

```
dem_lda5 <- LDA(dtm_dem, k=5, method = "vem", control = list(seed=72458), verbose=1)
topics_dem5 <- tidy(dem_lda5, matrix="beta")
terms_dem5 <- topics_dem5 %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

terms_dem5 %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill= factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  ggtitle("Beta values for each term, top 10 terms when k=5, Democrats") +
  theme_bw()
```



Below, I present a graphic of the top 10 words that appear in each topic model for k=5 in the Republican party platform.

```
repub_lda5 <- LDA(dtm_repub, k=5, method = "vem", control = list(seed=72458), verbose=1)
topics_repub5 <- tidy(repub_lda5, matrix="beta")
terms_repub5 <- topics_repub5 %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

terms_repub5 %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill= factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  ggtitle("Beta values for each term, top 10 terms when k=5, Republicans") +
  theme_bw()
```



Question 7: Interpreting topic modeling results for k=5

The topics that appear to emerge from the Democratic topic models are words like health, Democrats, believe, support, communities, and people. For Republicans, the words that emerge are government, federal, American, and states. It appears that the differences I have presented are consistent with the differences outlined in the wordclouds earlier.

Question 8: Fitting 6 more topic models

```
dem_lda10 <- LDA(dtm_dem, k=10, method = "vem", control = list(seed=72458), verbose=1)
dem_lda25 <- LDA(dtm_dem, k=25, method = "vem", control = list(seed=72458), verbose=1)
repub_lda10 <- LDA(dtm_repub, k=10, method = "vem", control = list(seed=72458), verbose=1)
repub_lda25 <- LDA(dtm_repub, k=25, method = "vem", control = list(seed=72458), verbose=1)
```

```
topics_dem10 <- tidy(dem_lda10, matrix="beta")
topics_dem25 <- tidy(dem_lda25, matrix="beta")
topics_repub10 <- tidy(repub_lda10, matrix="beta")
topics_repub25 <- tidy(repub_lda25, matrix="beta")
```

Below, I present the gamma values for k=5, k=10, and k=25 for the Democratic party.

Table 1: Gamma values for k=5, Democrats

document	topic	gamma
d16.txt	1	0.2003248
d16.txt	2	0.2073856
d16.txt	3	0.1997541
d16.txt	4	0.1990617
d16.txt	5	0.1934739

Table 2: Gamma values for k=10, Democrats

document	topic	gamma
d16.txt	1	0.0960585
d16.txt	2	0.1056441
d16.txt	3	0.1035029
d16.txt	4	0.0884071
d16.txt	5	0.0778009
d16.txt	6	0.0868989
d16.txt	7	0.2024502
d16.txt	8	0.0780557
d16.txt	9	0.0719824
d16.txt	10	0.0891993

```
gamma1 <- tidy(dem_lda5, matrix = "gamma")
gamma2 <- tidy(dem_lda10, matrix = "gamma")
gamma3 <- tidy(dem_lda25, matrix = "gamma")
kable(gamma1, caption= "Gamma values for k=5, Democrats")
```

```
kable(gamma2, caption= "Gamma values for k=10, Democrats")
```

```
kable(gamma3, caption = "Gamma values for k=25, Democrats")
```

Below, I present the gamma values for k=5, k=10, and k=25 for the Republican party.

```
gamma4 <- tidy(repub_lda5, matrix = "gamma")
gamma5 <- tidy(repub_lda10, matrix = "gamma")
gamma6 <- tidy(repub_lda25, matrix = "gamma")
kable(gamma4, caption = "Gamma values for k=5, Republicans")
```

```
kable(gamma5, caption = "Gamma values for k=10, Republicans")
```

```
kable(gamma6, caption = "Gamma values for k=25, Republicans")
```

Generally, it appears from the Gamma values that for both parties, there is an approximately similar percentage of the words that fall into each topic.

Below, I present the results for each model visually as well. The following plot presents the top 10 words for Democrats when k=10.

Table 3: Gamma values for k=25, Democrats

document	topic	gamma
d16.txt	1	0.0360842
d16.txt	2	0.0468802
d16.txt	3	0.0395367
d16.txt	4	0.0299835
d16.txt	5	0.0296698
d16.txt	6	0.0316156
d16.txt	7	0.1553082
d16.txt	8	0.0297881
d16.txt	9	0.0298984
d16.txt	10	0.0327348
d16.txt	11	0.0269589
d16.txt	12	0.0260359
d16.txt	13	0.0330589
d16.txt	14	0.0527368
d16.txt	15	0.0280324
d16.txt	16	0.0321299
d16.txt	17	0.0377628
d16.txt	18	0.0693940
d16.txt	19	0.0331709
d16.txt	20	0.0347224
d16.txt	21	0.0286903
d16.txt	22	0.0320423
d16.txt	23	0.0282230
d16.txt	24	0.0442831
d16.txt	25	0.0312591

Table 4: Gamma values for k=5, Republicans

document	topic	gamma
r16.txt	1	0.1560754
r16.txt	2	0.2130387
r16.txt	3	0.2208086
r16.txt	4	0.2175818
r16.txt	5	0.1924955

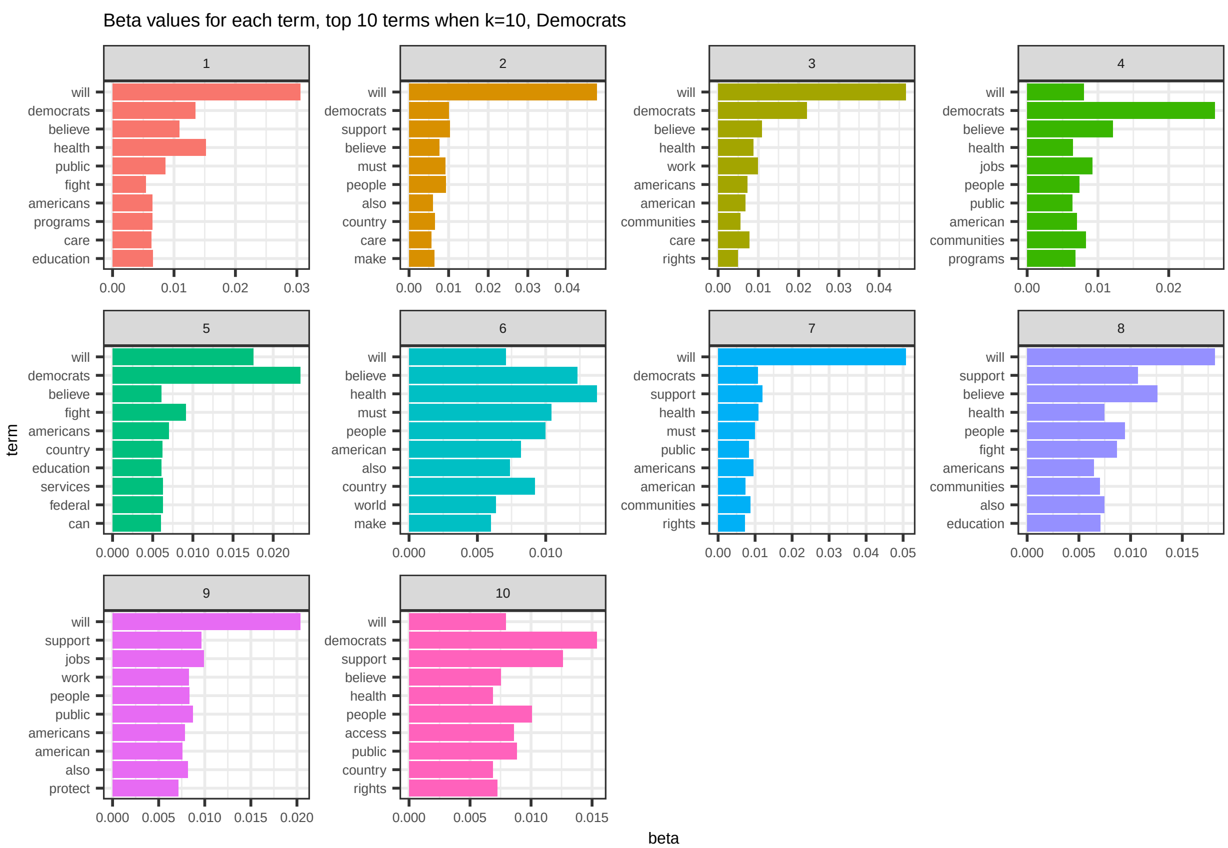
Table 5: Gamma values for k=10, Republicans

document	topic	gamma
r16.txt	1	0.0797756
r16.txt	2	0.1073378
r16.txt	3	0.1034078
r16.txt	4	0.1115084
r16.txt	5	0.0939569
r16.txt	6	0.0882199
r16.txt	7	0.0998455
r16.txt	8	0.1257610
r16.txt	9	0.0875681
r16.txt	10	0.1026189

Table 6: Gamma values for k=25, Republicans

document	topic	gamma
r16.txt	1	0.0350237
r16.txt	2	0.0393419
r16.txt	3	0.0369175
r16.txt	4	0.0415519
r16.txt	5	0.0356014
r16.txt	6	0.0345468
r16.txt	7	0.0368005
r16.txt	8	0.0454573
r16.txt	9	0.0347151
r16.txt	10	0.0376132
r16.txt	11	0.0360684
r16.txt	12	0.0415108
r16.txt	13	0.0648290
r16.txt	14	0.0445301
r16.txt	15	0.0451823
r16.txt	16	0.0353789
r16.txt	17	0.0386692
r16.txt	18	0.0343253
r16.txt	19	0.0363441
r16.txt	20	0.0468747
r16.txt	21	0.0500794
r16.txt	22	0.0373844
r16.txt	23	0.0373558
r16.txt	24	0.0354472
r16.txt	25	0.0384512

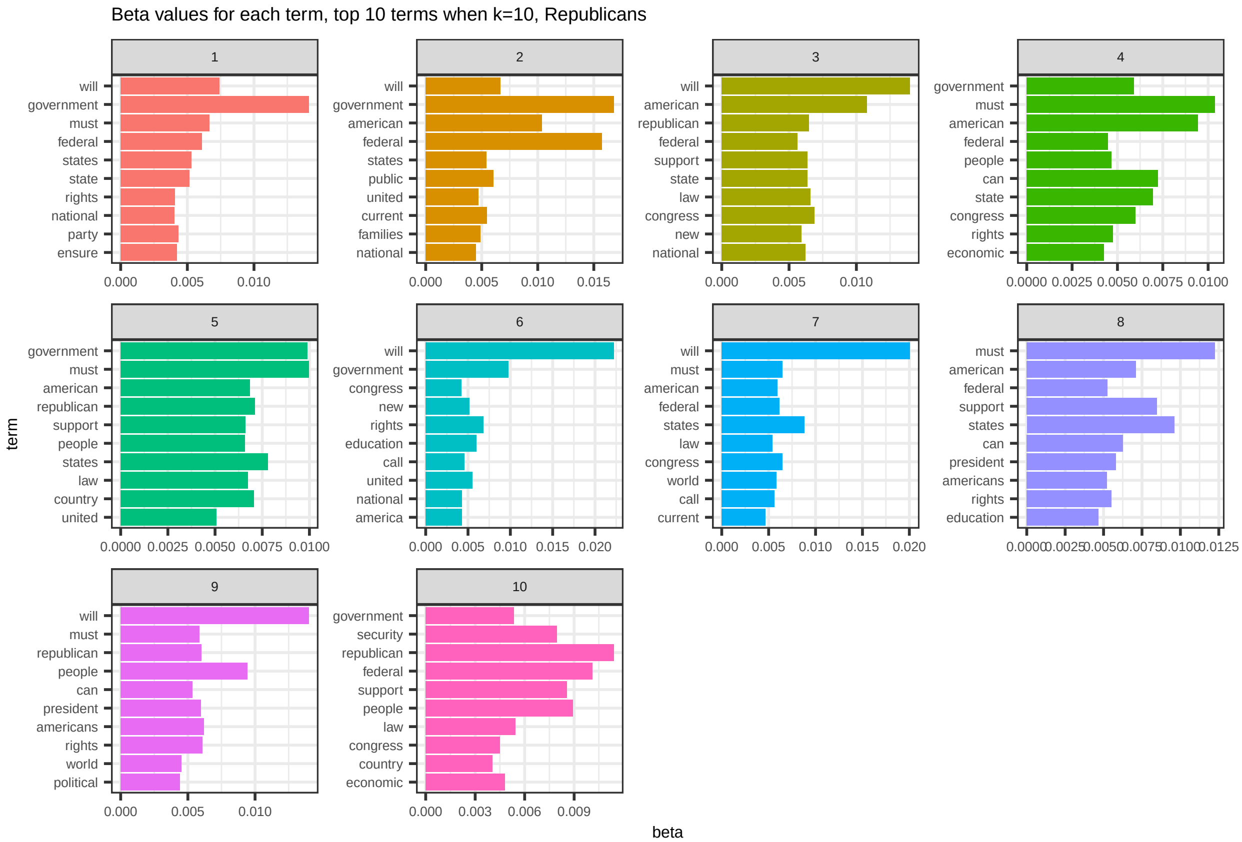
```r
terms_dem10 <- topics_dem10 %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

terms_dem10 %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill= factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  theme_bw() +
  ggtitle("Beta values for each term, top 10 terms when k=10, Democrats") +
  theme(text = element_text(size=6.0))
```



The following plot presents the top 10 words for Republicans when k=10.

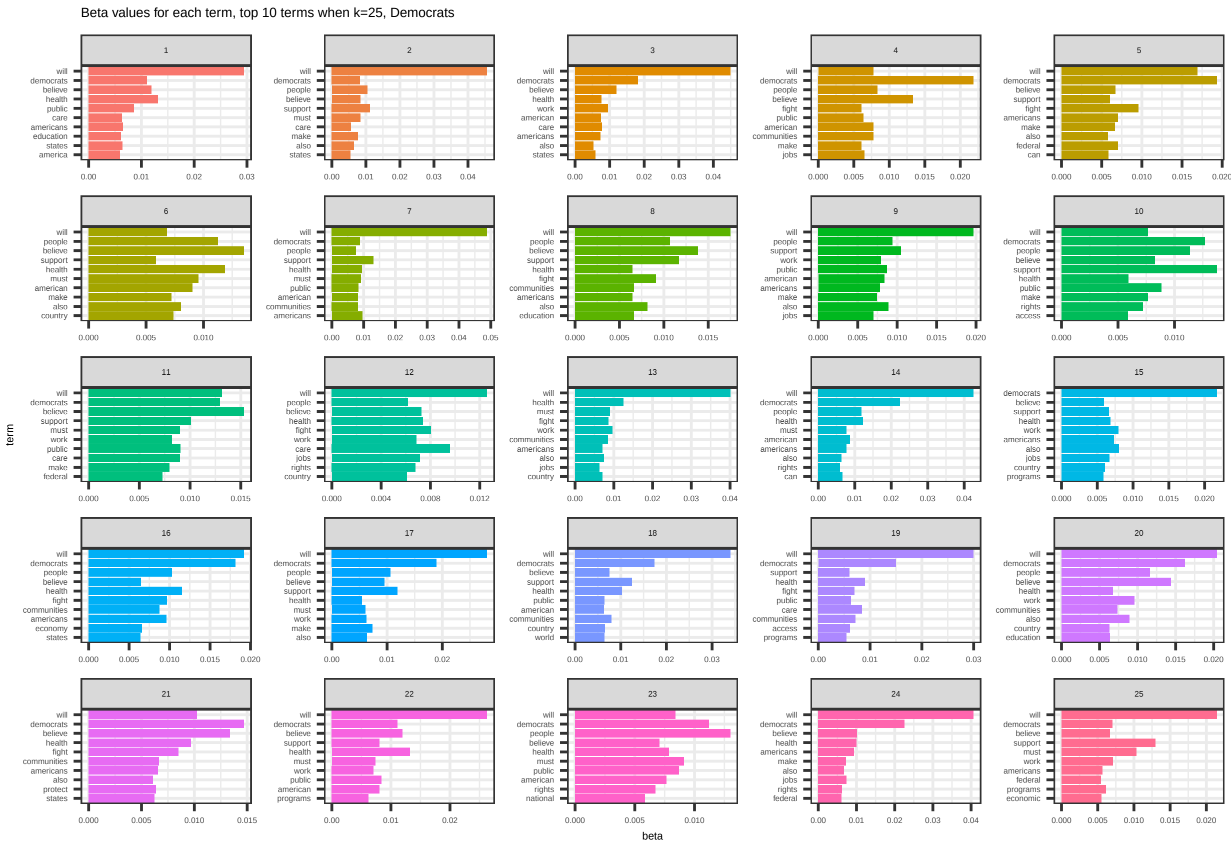
```r
topics_repub10 <- tidy(repub_lda10, matrix="beta")
terms_repub10 <- topics_repub10 %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

terms_repub10 %>%
```

```r
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill= factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  theme_bw() +
  ggtitle("Beta values for each term, top 10 terms when k=10, Republicans") +
  theme(text = element_text(size=6.0))
```



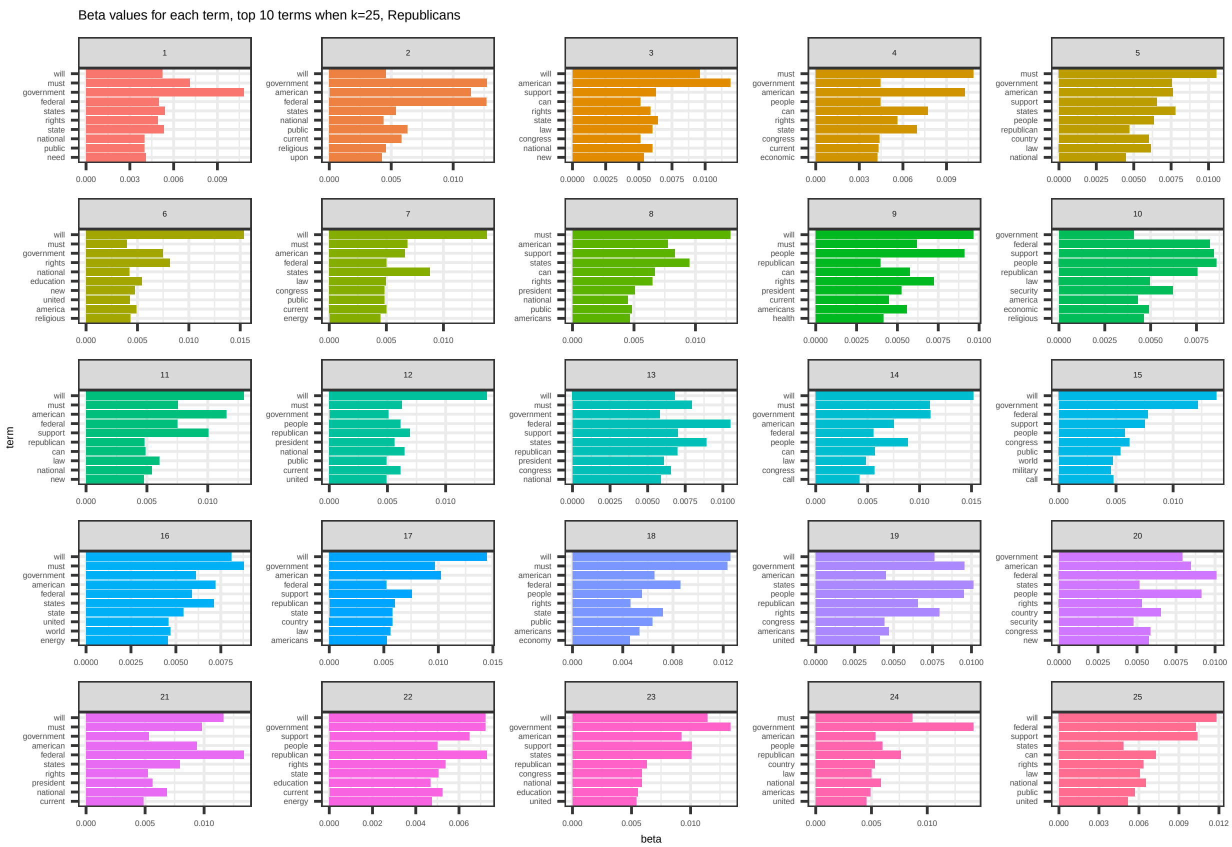
The following plot presents the top 10 words for Democrats when k=25.

```r
terms_dem25 <- topics_dem25 %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

terms_dem25 %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill= factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  theme_bw() +
  ggtitle("Beta values for each term, top 10 terms when k=25, Democrats") +
  theme(text = element_text(size=4.0))
```



The following plot presents the top 10 words for Republicans when k=25.

```
topics_repub25 <- tidy(repub_lda25, matrix="beta")
terms_repub25 <- topics_repub25 %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

terms_repub25 %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill= factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  theme_bw() +
  ggtitle("Beta values for each term, top 10 terms when k=25, Republicans") +
  theme(text = element_text(size=4.0))
```



Question 9: Calculating Perplexity

Below are the perplexity scores for Democrats at k=5, k=10, and k=25.

```r
perplexity(dem_lda5)
```

```
## [1] 1500.82
```

```r
perplexity(dem_lda10)
```

```
## [1] 1502.922
```

```r
perplexity(dem_lda25)
```

```
## [1] 1507.185
```

Below are the perplexity scores for Republicans at k=5, k=10, and k=25.

```r
perplexity(repub_lda5)
```

```
## [1] 2290.054
```

```r
perplexity(repub_lda10)
```

```
## [1] 2291.497
```

```r
perplexity(repub_lda25)
```

```
## [1] 2295.899
```

Above, I see that for Democrats, perplexity is around 1600, whereas for Republicans, perplexity is around 2300. I would expect that as the number of topics increase, the perplexity would increase. That holds true, but the perplexity overall appears to be more flat than I would have anticipated.

Question 10: Barplot for k=10

In question 8, I present a barplot for each party when k=10. The top words that emerge for Democrats and Republicans actually appear similar to what was seen for k=5, which leads me to believe that 5 topics were probably enough in this case. I see the same general trends as I explained in question 7.

Question 11: Conclusion

Based on my findings, I would support the Democratic party for two reasons. For one, the sentiment of their party platform appears to be generally more positive. Secondly, I appreciated the emphasis on fighting for and supporting communities, and the emphasis on healthcare.