# Unsupervised Learning: PSet 5

*Felipe Alamos*

*11/27/2019*

```r
#Setups
library(dplyr)
library(skimr)
library(seriation)
library(ggplot2)
library(dbscan)
library(mixtools)
library(plotGMM)
library(gridExtra)
library(tm)
library(wordcloud)
library(knitr)
library(tibble)
library(tidytext)
library(topicmodels)
library(tidyr)
```

## PREPROCESSING & (light) EDA

### 1

*Load the `platforms.csv` file containing the 2016 Democratic and Republican party platforms. Note the 2X2 format, where each row is a document, with the party recorded as a separate feature. Also, load the individual party `.txt` files as a corpus.*

```r
texts <-
  file.path(
    "/home/fhalamos/Unsupervised/Problem-Set-5/Party Platforms Data/texts"
    )

# Now we can create our raw corpus
docs <- VCorpus(DirSource(texts))
summary(docs)
```

```
##         Length Class             Mode
## d16.txt 2      PlainTextDocument list
## r16.txt 2      PlainTextDocument list
```

### 2

*Create a document-term matrix and preprocess the platforms by the following criteria (at a minimum): * Convert to lowercase * Remove the stopwords * Remove the numbers * Remove all punctuation * Remove the whitespace*

```r
# PREPROCESSING
docs <- docs %>%
  tm_map(content_transformer(tolower)) %>%
  tm_map(tolower) %>% # Remove captialization
  tm_map(removeWords, stopwords("english")) %>% #Remove stopwords
  tm_map(removeNumbers) %>% # Remove numbers
  tm_map(removePunctuation) %>%  #Remove Punctuation
  tm_map(stripWhitespace)  %>% #Remove white spaces
  tm_map(PlainTextDocument)

# A bit more cleaning of unique characters
for (j in seq(docs)) {
  docs[[j]] <- gsub("/", " ", docs[[j]])
  docs[[j]] <- gsub("'", " ", docs[[j]])
  docs[[j]] <- gsub("-", " ", docs[[j]])
  docs[[j]] <- gsub("\\|", " ", docs[[j]])
  docs[[j]] <- gsub("@", " ", docs[[j]])
  docs[[j]] <- gsub("\u2028", " ", docs[[j]])  # an ascii character that does not translate
}

docs <- docs %>%
  tm_map(PlainTextDocument)

#Manually removing some words that appear a lot and do not add a lot of value
docs <-
  tm_map(docs,
         removeWords,
         c("democrats", "republicans", "will", "must", "believe", "also"))

#Put together some specific words
for (j in seq(docs)) {
  docs[[j]] <- gsub("united states", "united_states", docs[[j]])
  docs[[j]] <- gsub("federal government", "federal_government", docs[[j]])
}

docs <- tm_map(docs, PlainTextDocument)
```

Now that we have cleaned the data, we create the document term matrix.

```r
dtm <- DocumentTermMatrix(docs)
```

# 3

*Visually inspect your cleaned documents by creating a wordcloud for each major party's platform. Based on this naive visualization, offer a few-sentence-description of general patterns you see (e.g., What are commonly used words? What are less commonly used words? Can you get a sense of differences between the parties at this early stage?*

```r
#Array of frequencies of word in each document
d_frequency <- sort(as.matrix(dtm)[1,],decreasing=TRUE)
r_frequency <- sort(as.matrix(dtm)[2,],decreasing=TRUE)
```

Democrats Term Frequency Wordcloud:

```r
wordcloud(names(d_frequency), d_frequency,
  min.freq = 1, # terms used at least once
  max.words = 150, # 300 most frequently used terms
  random.order = FALSE, # centers cloud by frequency, > = center
  rot.per = 0.30, # sets proportion of words oriented horizontally
  main = "Title",
  colors = brewer.pal(8, "Dark2")
  )
```


The most common words in the democrats platform are:

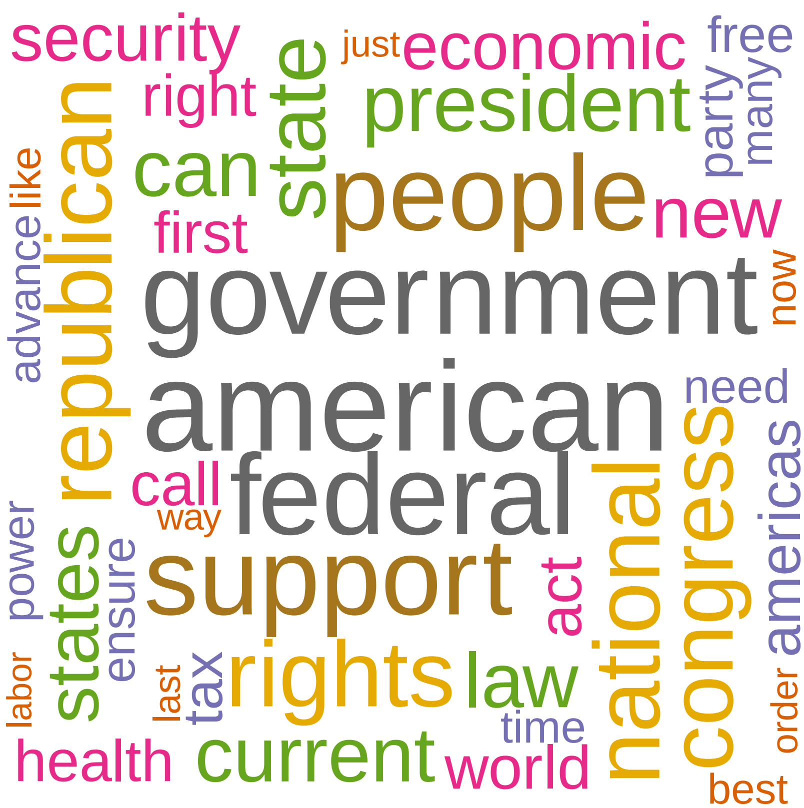
```r
kable(head(d_frequency, 10), caption="Most common words democrat platform")
```

Table 1: Most common words democrat platform

| | x |
|---|---|
| health | 130 |
| support | 123 |
| people | 111 |
| americans | 94 |
| american | 86 |
| communities | 81 |
| public | 79 |
| work | 72 |
| rights | 71 |
| care | 66 |

Republicans Term Frequency Wordcloud:

```r
#plot.new()
#text(x=0.5, y=0.5, "Republicans Term Frequency Wordcloud")
wordcloud(names(r_frequency), r_frequency,
  min.freq = 1, # terms used at least once
  max.words = 150,
  random.order = FALSE, # centers cloud by frequency, > = center
  rot.per = 0.30, # sets proportion of words oriented horizontally
  main = "Title",
  colors = brewer.pal(8, "Dark2")
  )
```




The most common used words in the republicans platforms are:

```r
kable(head(r_frequency, 10), caption="Most common words republican platform")
```

Table 2: Most common words republican platform

| | x |
|---|---|
| american | 121 |
| government | 110 |
| federal | 106 |
| support | 100 |
| people | 98 |
| national | 83 |
| republican | 83 |
| rights | 83 |
| congress | 81 |

| | x |
|---|---|
| state | 74 |

Most popular words from republican platform tend to be words very related to the state arena: government, federal, people, national, republican, congress, state.

In the other hand, popular democrats words are more related to words related to peoples needs: health, support, communities, public, work, care.

This is a first glace of the priorities and worries the platforms express. While the republican focus their speech in the state, democrats seem to talk more about peoples needs. As expected, both platforms use the word american a lot of times, but republicans do it more intensively. This also speaks about the common rhetoric in US politics, which is usually very US centered.

# SENTIMENT ANALYSIS

Reference used for this section: https://www.tidytextmining.com/sentiment.html

## 4.

*Use the "Bing" and "AFINN" dictionaries to calculate the sentiment of each cleaned party platform. Present the results however you'd like (e.g., visually and/or numerically).*

```
dem_freq_df <- as.data.frame(d_frequency) %>%
  rownames_to_column("word") %>%
  rename(freq = d_frequency)

rep_freq_df <- as.data.frame(r_frequency) %>%
  rownames_to_column("word") %>%
  rename(freq = r_frequency)
```

Lets first make a sentiment analysis study using bing dictionary, which classifies words as positive or negative.

We will first acknowledge which are the sentiments of the most frequent words used in each platform.

```
bing <- get_sentiments("bing")

dem_bing <- dem_freq_df %>%
  inner_join(bing)
kable(head(dem_bing,5), caption = "Democrats: Top 5 words")
```

Table 3: Democrats: Top 5 words

| word | freq | sentiment |
|---|---|---|
| support | 123 | positive |
| work | 72 | positive |
| protect | 46 | positive |
| right | 37 | positive |
| clean | 33 | positive |

```r
rep_bing <- rep_freq_df %>%
  inner_join(bing)
```

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

```r
kable(head(rep_bing,5), caption = "Democrats: Top 5 words")
```

Table 4: Democrats: Top 5 words

| word | freq | sentiment |
|---|---:|---|
| support | 100 | positive |
| right | 46 | positive |
| oppose | 43 | negative |
| freedom | 42 | positive |
| protect | 38 | positive |

A first difference observed is that the top 5 words most used by democrats are all positive, while republicans have one negative word as one of their top 5 used.

If we study the total amount of positive and negative words used, we get the following results:

```r
#Lets observe total amount of positive and negative words used
dem_bing_summary_df <-
  aggregate(dem_bing$freq, by=list(sentiment=dem_bing$sentiment), FUN=sum)
colnames(dem_bing_summary_df) <- c("sentiment", "freq")
kable(dem_bing_summary_df,
      caption="Democrats: Total positive/negative words used")
```

Table 5: Democrats: Total positive/negative words used

| sentiment | freq |
|---|---:|
| negative | 811 |
| positive | 1372 |

```r
rep_bing_summary_df <-
  aggregate(rep_bing$freq, by=list(sentiment=rep_bing$sentiment), FUN=sum)
colnames(rep_bing_summary_df) <- c("sentiment", "freq")

kable(rep_bing_summary_df,
      caption="Republicans: Total positive/negative words used")
```

Table 6: Republicans: Total positive/negative words used

| sentiment | freq |
|---|---:|
| negative | 1245 |
| positive | 1578 |

We observe that both parties use more positive than negative words in their platform. Nevertheless, the proportion of positive words respect to negatives is significantly bigger in the democratic one.

We can also visualize the previous statement through computation of the "tone" of the platforms, which is calculated as the difference between positive words and negative words over the total number of words:

```r
dem_tone <- (dem_bing_summary_df[2,2] - dem_bing_summary_df[1,2]) /
  (dem_bing_summary_df[2,2] + dem_bing_summary_df[1,2])

rep_tone <- (rep_bing_summary_df[2,2] - rep_bing_summary_df[1,2]) /
  (rep_bing_summary_df[2,2] + rep_bing_summary_df[1,2])

df <- data.frame(dem_tone, rep_tone)
colnames(df) <- c("Democratic tone", "Republican tone")
kable(df,caption="Tones in platforms")
```

Table 7: Tones in platforms

| Democratic tone | Republican tone |
|---:|---:|
| 0.2569858 | 0.1179596 |

Lets now use the afinn dictionary, which associates words with values in the range [-5,5].

```r
afinn <- get_sentiments("afinn") #Values -5...5

dem_afinn <- dem_freq_df %>%
  inner_join(afinn)
```

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

```r
rep_afinn <- rep_freq_df %>%
  inner_join(afinn)
```

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

```r
#Lets compute the total average score:
dem_afinn_score <- sum(dem_afinn$freq * dem_afinn$value)
rep_afinn_score <- sum(rep_afinn$freq * rep_afinn$value)

df <- data.frame(dem_afinn_score, rep_afinn_score)
colnames(df) <- c("Democratic score", "Republican score")
kable(df,caption="Average affin score by platforms")
```

Table 8: Average affin score by platforms

| Democratic score | Republican score |
|---:|---:|
| 1271 | 896 |

We again observe a higher score associated to the democratic platform when using the afinn dictionary.

## 5.

*Compare and discuss the sentiments of these platforms: which party tends to be more optimistic about the future? Does this comport with your perceptions of the parties?*

From the results exposed in the previous question, we can observe that the democratic platform tends to have a more optimistic view of the future that the republican one. As a matter of fact, when using the afinn dictionary, the sentiment score was 41.8526786% bigger, and when using the bing dictionary, their tone had more double score.

This did comport with my perception of the parties. I have usually associated the republican party with postures associated to danger/skepticism feelings, in topics for example related to war, immigration or climate change. Although Democrats also use negative language, particularly to refer to inequality, I do feel that their language is usually associated to positive feelings, especially around the rhetoric of building a U.S for all.

# TOPIC MODELS

Reference used for this section: https://cfss.uchicago.edu/notes/topic-modeling/

## 6.

*Now explore the topics they are highlighting in their platforms. This will give a sense of the key policy areas they're most interested in. Fit a topic model for each of the major parties (i.e. two topic models) using the latent Dirichlet allocation algorithm, initialized at `k = 5` topics as a start. Present the results however you'd like (e.g., visually and/or numerically).*

```
#Create the dtm matrixes
dem_freq_df$doc <- 1
dem_dtm <- dem_freq_df %>% cast_dtm(doc, word, freq)

rep_freq_df$doc <- 2
rep_dtm <- rep_freq_df %>% cast_dtm(doc, word, freq)

#We fit the models
dem_lda_5 <- LDA(dem_dtm, k = 5)
rep_lda_5 <- LDA(rep_dtm, k = 5)

#Transform them to tidy for analysis
dem_lda_td <- tidy(dem_lda_5)
rep_lda_td <- tidy(rep_lda_5)
```

We present the top 5 terms of each topic.

```
dem_top_terms <- dem_lda_td %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

df<-as.data.frame(terms(dem_lda_5,5))
kable(df, caption = "Democrats: Top 5 words in k=5 model")
```

Table 9: Democrats: Top 5 words in k=5 model

| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
|---|---|---|---|---|
| health | health | health | work | workers |
| support | communities | public | people | americans |
| people | rights | people | support | american |
| americans | support | including | families | public |

| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
|---|---|---|---|---|
| communities | americans | american | communities | support |

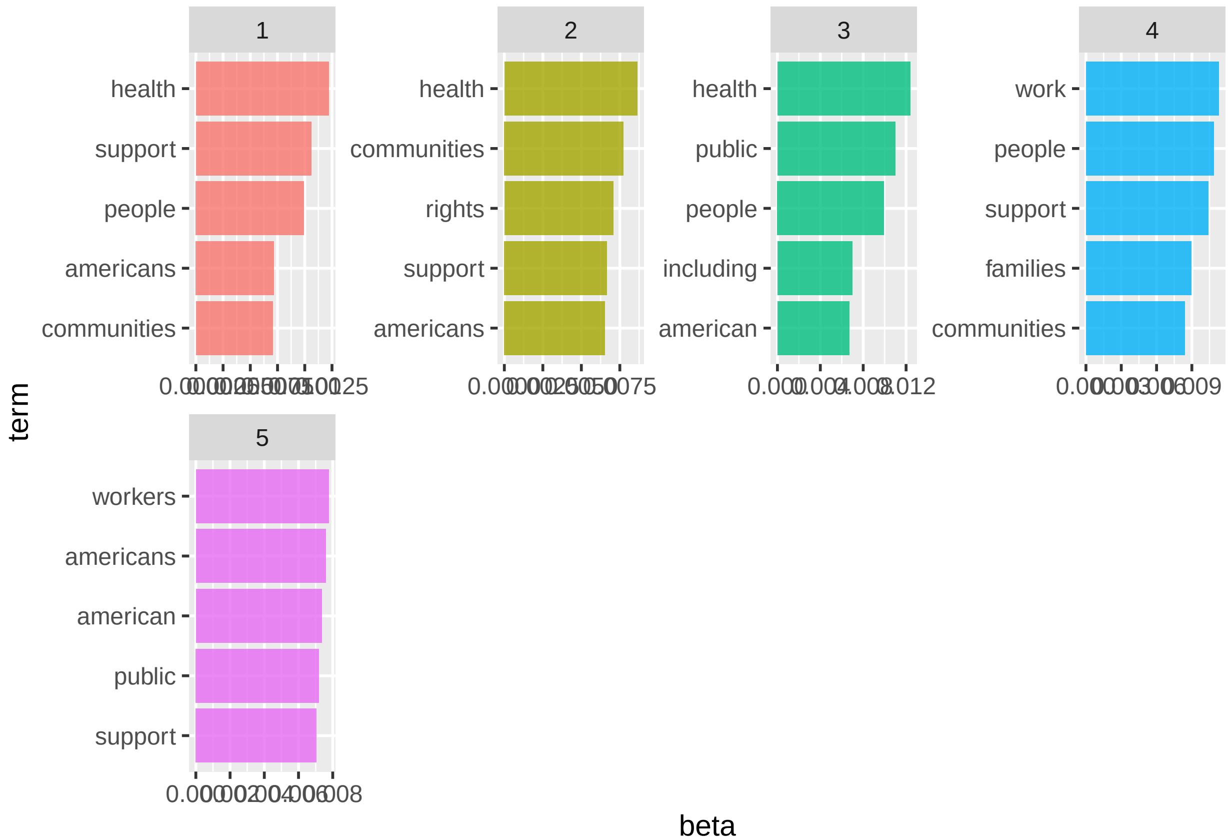
```
rep_top_terms <- rep_lda_td %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

df<-as.data.frame(terms(rep_lda_5,5))
kable(df, caption = "Republicans: Top 5 words in k=5 model")
```

Table 10: Republicans: Top 5 words in k=5 model

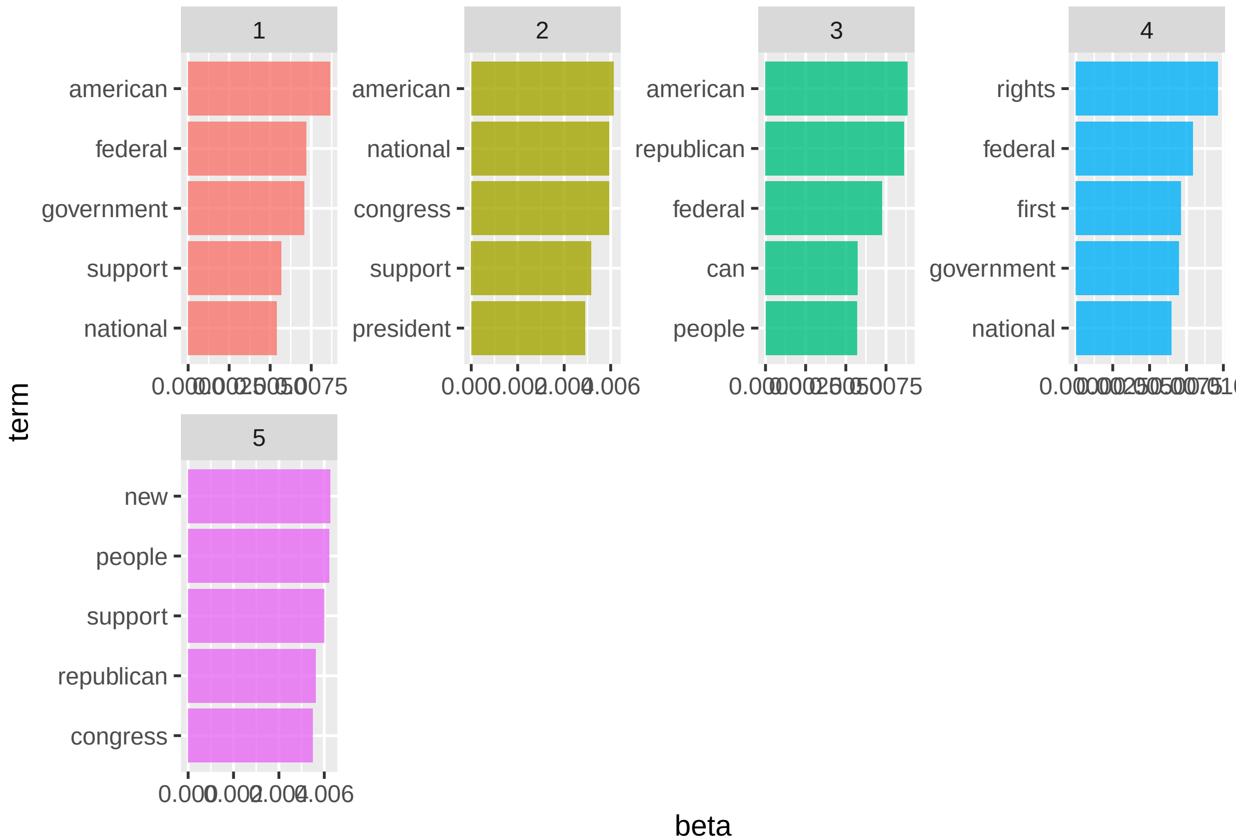
| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
|---|---|---|---|---|
| american | american | american | rights | new |
| federal | national | republican | federal | people |
| government | congress | federal | first | support |
| support | support | can | government | republican |
| national | president | people | national | congress |

Now we generate some visualizations of the terms to improve readability. We also include the terms respective beta values (probabiliy of term being generated by topic):

```
dem_top_terms %>%
  mutate(topic = factor(topic),
         term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = topic)) +
  geom_bar(alpha = 0.8, stat = "identity", show.legend = FALSE) +
  scale_x_reordered() +
  facet_wrap(~ topic, scales = "free", ncol = 4) +
  coord_flip()
```




```
rep_top_terms %>%
  mutate(topic = factor(topic),
         term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = topic)) +
  geom_bar(alpha = 0.8, stat = "identity", show.legend = FALSE) +
  scale_x_reordered() +
  facet_wrap(~ topic, scales = "free", ncol = 4) +
  coord_flip()
```




## 7

*Describe the general trends in topics that emerge from this stage. Are the parties focusing on similar or different topics, generally?*

[UPDATE THIS!!!]

We can observe that parties are focusing in quite different topics.

If we study the case of the democratic topics, we first of all observe some overlap between the topics. Example, the word health is in all of them, which reflects this term if quite consistent in their platform. Thtat said, we can identify the following topics: - Offering health, jobs and support in general. Terms: people, health, support, jobs - Ensuring people are supported/protected: Terms: public, ensure, support

On the other hand, republicans are focusing in the following main topics: - US state. Terms: state, public, federal, american - Institution and order focus: congress, law, rights. The word support also is present in most of the topics.

These differences were expected. Republicans have a very country/state centered speech, in contrast with Democrats, which seem to focus around helping the people and their needs.

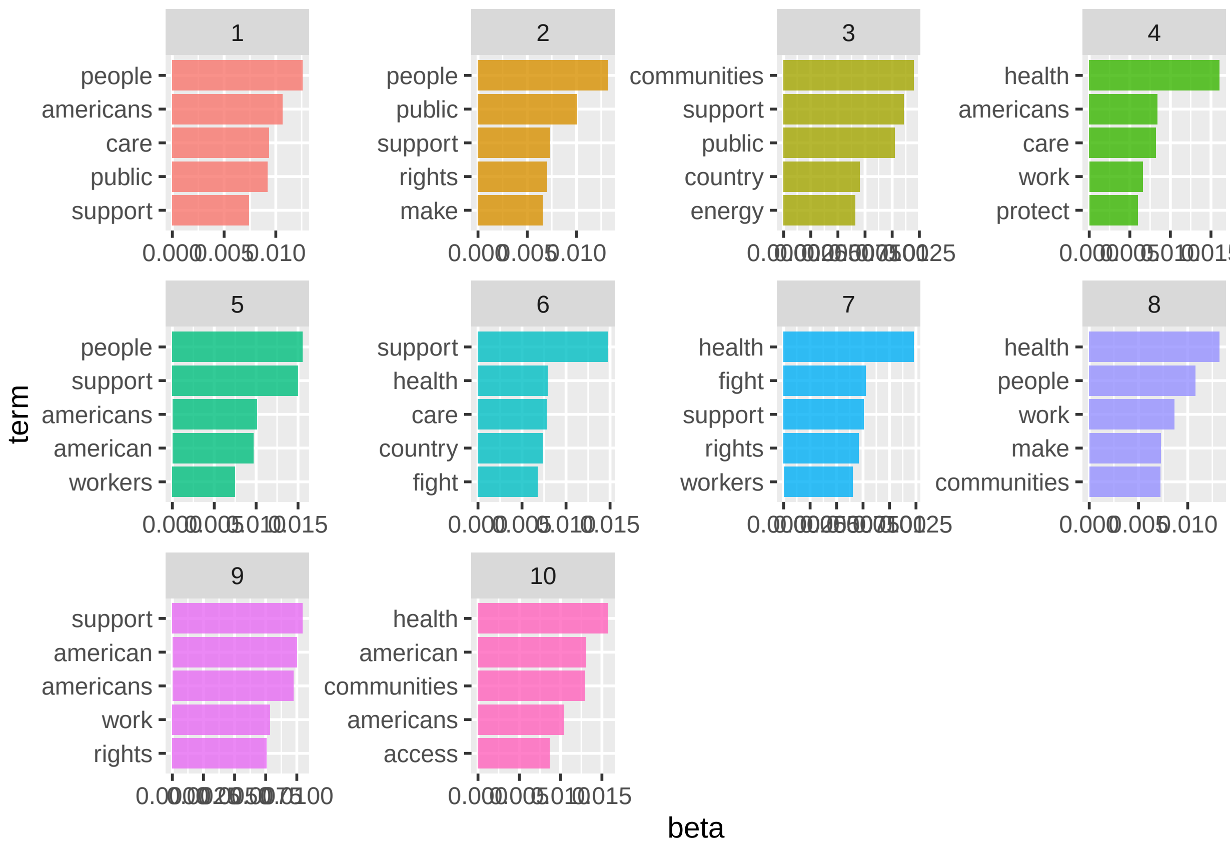
## 8

*Fit 6 more topic models at the follow levels of `k` for each party: `5, 10, 25`. Present the results however you'd like (e.g., visually and/or numerically).*

We have already fit models for k=5 in question 5, so we will focus in k=10 and k=25. We repeat the exercise done previously, we will get top 5 words of each topic and make visualizations of their beta values.

LDA with k=10 for Democrat platform:

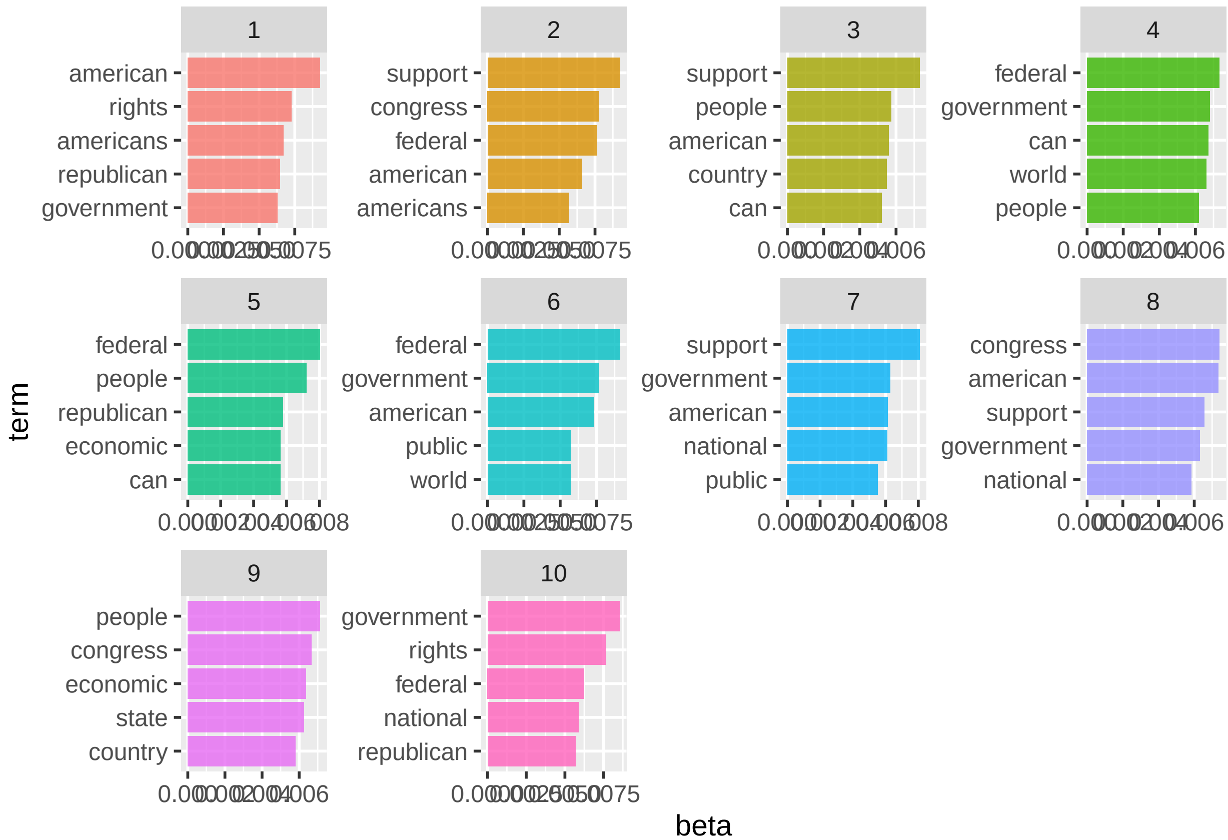
```
dem_lda_10 <- LDA(dem_dtm, k = 10)
dem_lda_td <- tidy(dem_lda_10)
dem_top_terms <- dem_lda_td %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
dem_top_terms_plot_10 <-
  dem_top_terms%>%
  mutate(topic = factor(topic),
         term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = topic)) +
  geom_bar(alpha = 0.8, stat = "identity", show.legend = FALSE) +
  scale_x_reordered() +
  facet_wrap(~ topic, scales = "free", ncol = 4) +
  coord_flip()
dem_top_terms_plot_10
```




LDA with k=10 for Republican platform:

```
rep_lda_10 <- LDA(rep_dtm, k = 10)
rep_lda_td <- tidy(rep_lda_10)
rep_top_terms <- rep_lda_td %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
```

```
  arrange(topic, -beta)
rep_top_terms_plot_10 <-
  rep_top_terms %>%
  mutate(topic = factor(topic),
         term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = topic)) +
  geom_bar(alpha = 0.8, stat = "identity", show.legend = FALSE) +
  scale_x_reordered() +
  facet_wrap(~ topic, scales = "free", ncol = 4) +
  coord_flip()
rep_top_terms_plot_10
```




Visualizations gets messy for k = 25 so we present tables.

LDA with k=25 for Democrat platform.

```
dem_lda_25 <- LDA(dem_dtm, k = 25)

kable(as.data.frame(terms(dem_lda_25,5))[,1:5])
```

| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
|---|---|---|---|---|
| health | people | americans | people | people |
| support | americans | american | support | american |
| public | rights | health | american | ensure |
| american | health | people | americans | support |
| fight | support | jobs | public | public |

```r
kable(as.data.frame(terms(dem_lda_25,5))[,6:10])
```

| Topic 6 | Topic 7 | Topic 8 | Topic 9 | Topic 10 |
|---|---|---|---|---|
| people | support | support | american | support |
| support | americans | health | ensure | people |
| health | communities | rights | care | communities |
| communities | health | people | rights | public |
| protect | care | americans | fight | work |

```r
kable(as.data.frame(terms(dem_lda_25,5))[,11:15])
```

| Topic 11 | Topic 12 | Topic 13 | Topic 14 | Topic 15 |
|---|---|---|---|---|
| people | health | health | support | health |
| support | americans | people | americans | americans |
| work | public | make | health | work |
| american | country | american | communities | world |
| fight | fight | support | work | support |

```r
kable(as.data.frame(terms(dem_lda_25,5))[,16:20])
```

| Topic 16 | Topic 17 | Topic 18 | Topic 19 | Topic 20 |
|---|---|---|---|---|
| communities | health | health | communities | americans |
| americans | people | american | people | including |
| country | american | public | health | public |
| care | support | make | rights | communities |
| work | communities | country | make | fight |

```r
kable(as.data.frame(terms(dem_lda_25,5))[,21:25])
```

| Topic 21 | Topic 22 | Topic 23 | Topic 24 | Topic 25 |
|---|---|---|---|---|
| communities | support | people | health | support |
| americans | work | ensure | americans | health |
| public | public | americans | support | american |
| people | world | communities | american | work |
| rights | jobs | public | people | americans |

LDA with k=25 for Republican platform:

```r
rep_lda_25 <- LDA(rep_dtm, k = 25)

kable(as.data.frame(terms(rep_lda_25,5))[,1:5])
```

| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
|---|---|---|---|---|
| american | american | federal | federal | american |
| religious | federal | people | government | congress |
| republican | government | united_states | rights | republican |
| americans | national | president | president | people |

| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
|---|---|---|---|---|
| call | support | americas | current | economic |

```
kable(as.data.frame(terms(rep_lda_25,5))[,6:10])
```

| Topic 6 | Topic 7 | Topic 8 | Topic 9 | Topic 10 |
|---|---|---|---|---|
| national | american | government | government | american |
| government | government | american | federal | government |
| federal | national | rights | support | congress |
| american | support | current | american | rights |
| security | republican | people | economic | people |

```
kable(as.data.frame(terms(rep_lda_25,5))[,11:15])
```

| Topic 11 | Topic 12 | Topic 13 | Topic 14 | Topic 15 |
|---|---|---|---|---|
| national | republican | government | american | congress |
| federal | federal | american | rights | american |
| people | people | support | congress | law |
| can | support | republican | people | states |
| congress | can | congress | president | rights |

```
kable(as.data.frame(terms(rep_lda_25,5))[,16:20])
```

| Topic 16 | Topic 17 | Topic 18 | Topic 19 | Topic 20 |
|---|---|---|---|---|
| american | federal | government | people | government |
| government | congress | federal | congress | national |
| people | states | american | government | support |
| state | state | national | families | people |
| republican | rights | support | security | energy |

```
kable(as.data.frame(terms(rep_lda_25,5))[,21:25])
```

| Topic 21 | Topic 22 | Topic 23 | Topic 24 | Topic 25 |
|---|---|---|---|---|
| american | federal | government | national | national |
| support | state | support | support | american |
| federal | republican | people | government | republican |
| rights | support | state | republican | people |
| president | public | rights | american | support |

## 9.

*Calculate the perplexity of each model iteration and describe which technically fits best.*

```
dem_perplexity <- c(perplexity(dem_lda_5),
                    perplexity(dem_lda_10),
```

```r
                    perplexity(dem_lda_25))

rep_perplexity <- c(perplexity(rep_lda_5),
                    perplexity(rep_lda_10),
                    perplexity(rep_lda_25))


df <- data.frame(dem_perplexity, rep_perplexity)
row.names(df) <- c("k=5","k=10","k=25")

colnames(df) <- c("Democratic platform", "Republican platform")
kable(df,caption="Perplexity on models with different number of topics")
```

Table 21: Perplexity on models with different number of topics

| | Democratic platform | Republican platform |
|---|---|---|
| k=5 | 1708.813 | 2382.272 |
| k=10 | 1709.620 | 2383.818 |
| k=25 | 1714.724 | 2389.753 |

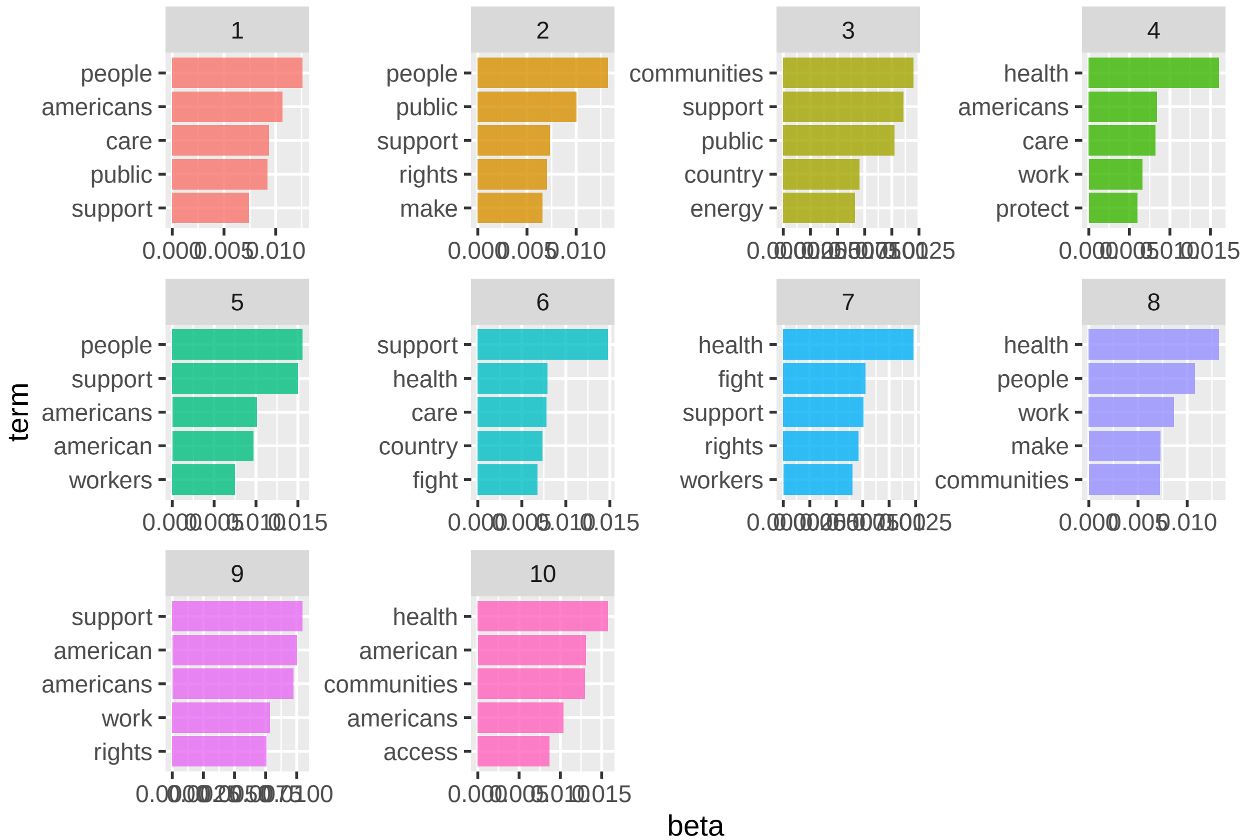
Based in the perplexity value, we can observe that for both democratic and republican platforms, the model that fits the best is the one with k=5.

## 10.

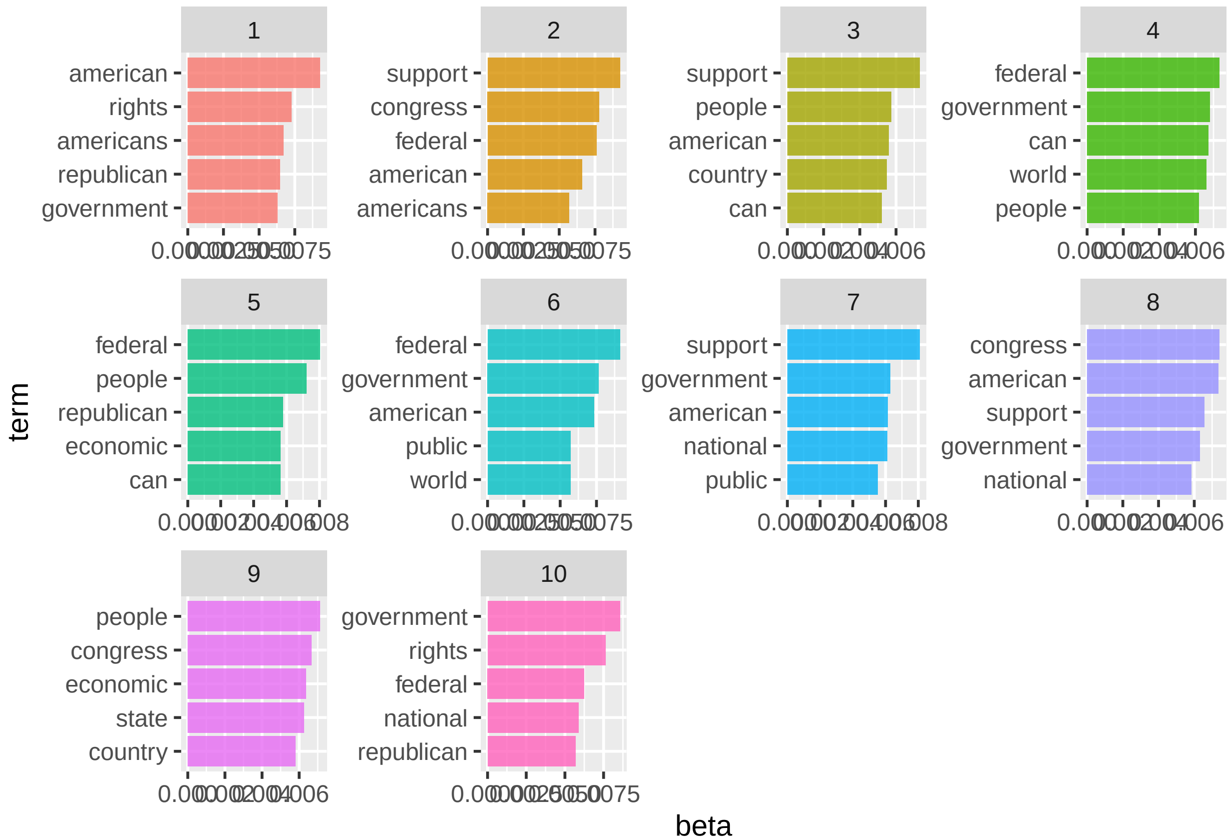
*Building on the previous question, display a barplot of the* `k = 10` *model for each party, and offer some general inferences as to the main trends that emerge. Are there similar themes between the parties? Do you think* `k = 10` *likely picks up differences more efficiently? Why or why not?*

The barplot of the `k = 10` model for each party were computed in question 8:

```r
dem_top_terms_plot_10
```




```
rep_top_terms_plot_10
```




It is important to notice, first of all, the overlapping of terms between different topics. For example, in the case of democrats, terms like health, support, people, communities, americans, show repetitively between topics. The same happens for the republican platform, topics like government, people, american show consistently across almost all topics.

In addition, some of these topics are similar between the two parties. For example, the terms government, people, american.

A difference that we can indeed perceive is that democrats topics are more related to services to the people (terms like health, people, public, rights, jobs are consistent), whereas the republican rhetoric seems to be more related to institutions or concepts related to the state (congress, government, federal, states, state).

In conclusion, especially after contrasting with the results obtained with k=5, we can establish that k=10 is not an adequate amount of topics to model these platforms. It seems that there are not 10 clearly distinguishable topics in each platform, and hence the overlap of terms. In addition, because we cannot capture purer topics, we are more prone to identify similarities between the two parties. As the perplexity analysis showed, k=5 should be preferred over k=10.

# CONCLUSION

## 11.

*Per the opening question, based on your analyses (including exploring party brands, general tones/sentiments, political outlook, and policy priorities), which party would you support in the 2020 election (again, this is hypothetical)?*

Based on this analysis, I will definitely support the democratic party.

First of all, I feel more attached to a positive look to the future. I believe that, if things are done properly, our current problems (such as inequality, climate change and public services in general) can be solved. I am enthusiastic about the future and do not have a pessimist thought. Hence, based on the sentiment analysis, the democratic platform feels more appealing to me.

Secondly, as a big picture, the topic analysis showed that the republican platform is structured around a US/state centered, enhancing values as the government, the state, the federal system. On the other hand, the topics observed in the democratic platform are more related to the people and its issues (such as health and jobs), without a too strong "US rhetoric". Personally, I feel more attached to the latter alternative. I do not have strong "patriotic" feelings, and feel that the republican topics usually are related to that. On the other hand, it does make sense to me to hear politicians dedicate their work and speech to peoples concrete problems such as health, employment and education. In addition, I prefere them to focuse on a united country that is opened to the world without fears and positive attitude.