Problem Set 3

Claire Brockway

11/21/2019

General NLP/Pre-Processing

Problem 1

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

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

```
library(grid)
library(wordcloud)
```

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

```
library(wordcloud2)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------- ti
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v ggplot2 3.2.1     v forcats 0.4.0
```

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

```
platforms <- read_csv("problem-set-3-master/platforms.csv")
```

```
## Parsed with column specification:
## cols(
##   party = col_character(),
##   platform = col_character()
## )
```

```
#reading in txt file as a corpus
corpus <- file.path("problem-set-3-master", "Party Platforms Data")
dir(corpus)
```

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

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

Problem 2

```
#creating document term matrix
#a. convert to lowercase
plat <- tm_map(plat, tolower)

plat <- tm_map(plat, removeWords, c("will"))

```

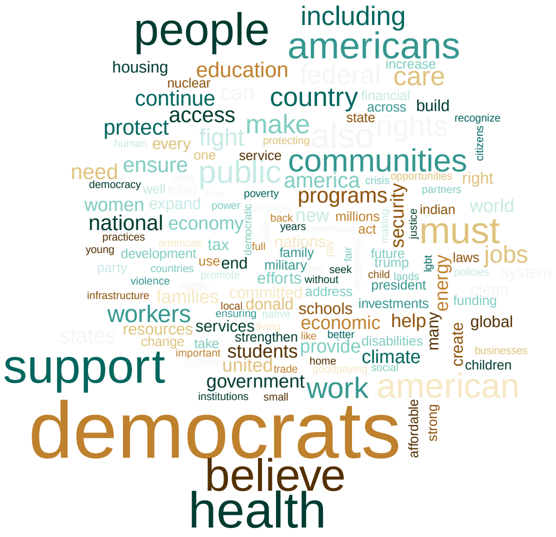
```r
#b. remove the stopwords
plat <- tm_map(plat,
               removeWords,
               stopwords("english"))

#c. remove the numbers
plat <- tm_map(plat, removeNumbers)

#d. remove the punctuation
plat <- tm_map(plat, removePunctuation)

#remove more punctuation
for (j in seq(plat)) {
  plat[[j]] <- gsub("/", " ", plat[[j]])
  plat[[j]] <- gsub("’", " ", plat[[j]])
  plat[[j]] <- gsub("-", " ", plat[[j]])
  plat[[j]] <- gsub("\\|", " ", plat[[j]])
  plat[[j]] <- gsub("@", " ", plat[[j]])
  plat[[j]] <- gsub("\u2028", " ", plat[[j]])  # an ascii character that does not translate
  plat[[j]] <- gsub("," , " ", plat[[j]])
}
```

Problem 3:

```r
set.seed(1122)

wordcloud(as.character(plat[1]),
          max.words=150, scale=c(2.5,.1),
          colors=brewer.pal(11, "BrBG"),
          random.color=TRUE)
```

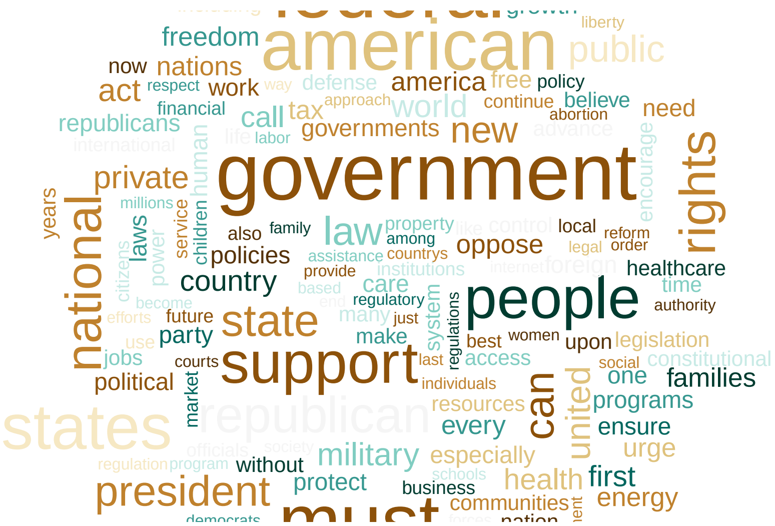
```
## Warning in tm_map.SimpleCorpus(corpus, tm::removePunctuation):
## transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(corpus, function(x) tm::removeWords(x,
## tm::stopwords())): transformation drops documents
```



```
set.seed(1123)

wordcloud(as.character(plat[2]),scale=c(2.5,.1),
          max.words=150,
          colors=brewer.pal(11, "BrBG"),
          random.color=TRUE)
```

```
## Warning in tm_map.SimpleCorpus(corpus, tm::removePunctuation):
## transformation drops documents

## Warning in tm_map.SimpleCorpus(corpus, tm::removePunctuation):
## transformation drops documents
```



After pre-processing the data it does seem like there are a lot of similarities between the two parties' platforms. Both the Demoratic and Republican platform appear to have a lot of mentions of "Americans" and "people", likely referencing their respective coalitions. The Democratic platform uses the term "democrats" alot while Republicans mention "Republican", as would be expected. A clear difference is Democrats seem to mention "health" alot while Republicans seem more focused on "security" and "rights". This might give us a sense of the parties' policy priorities. Republicans don't talk much about "women" and "child" and "health". Democrats don't talk much about "trade".

Sentiment Analysis:

Problem 4:

```
library(tidytext)
library(textdata)
library(dplyr)
library(ggplot2)
#using AFIN and BING dictionaries


# remove any dollar signs (they're special characters in R)
text <- gsub("\\$", "", plat)
text_dem <- gsub("\\$", "", plat[1])
text_rep <- gsub("\\$", "", plat[2])

# tokenize
tokens <- data_frame(text = text) %>% unnest_tokens(word, text)
```

```
## Warning: `data_frame()` is deprecated, use `tibble()`.
## This warning is displayed once per session.
```

```
tokens_dem <- data_frame(text = text_dem) %>% unnest_tokens(word, text)
tokens_rep <- data_frame(text = text_rep) %>% unnest_tokens(word, text)




afinn_dem <- tokens_dem %>%
  inner_join(get_sentiments("afinn")) %>%
  summarise(sentiment=sum(value)) %>%
  mutate(method="AFINN")
```

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

```
afinn_rep <- tokens_rep %>%
  inner_join(get_sentiments("afinn")) %>%
  summarise(sentiment=sum(value)) %>%
  mutate(method="AFINN")
```

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

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

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

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

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

```
#visually assess the output for AFINN


##assess frequencies of the most common joy words in afinn

head(tokens_dem %>%
  count(word, sort = TRUE) %>%
  inner_join(get_sentiments("afinn")))
```

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

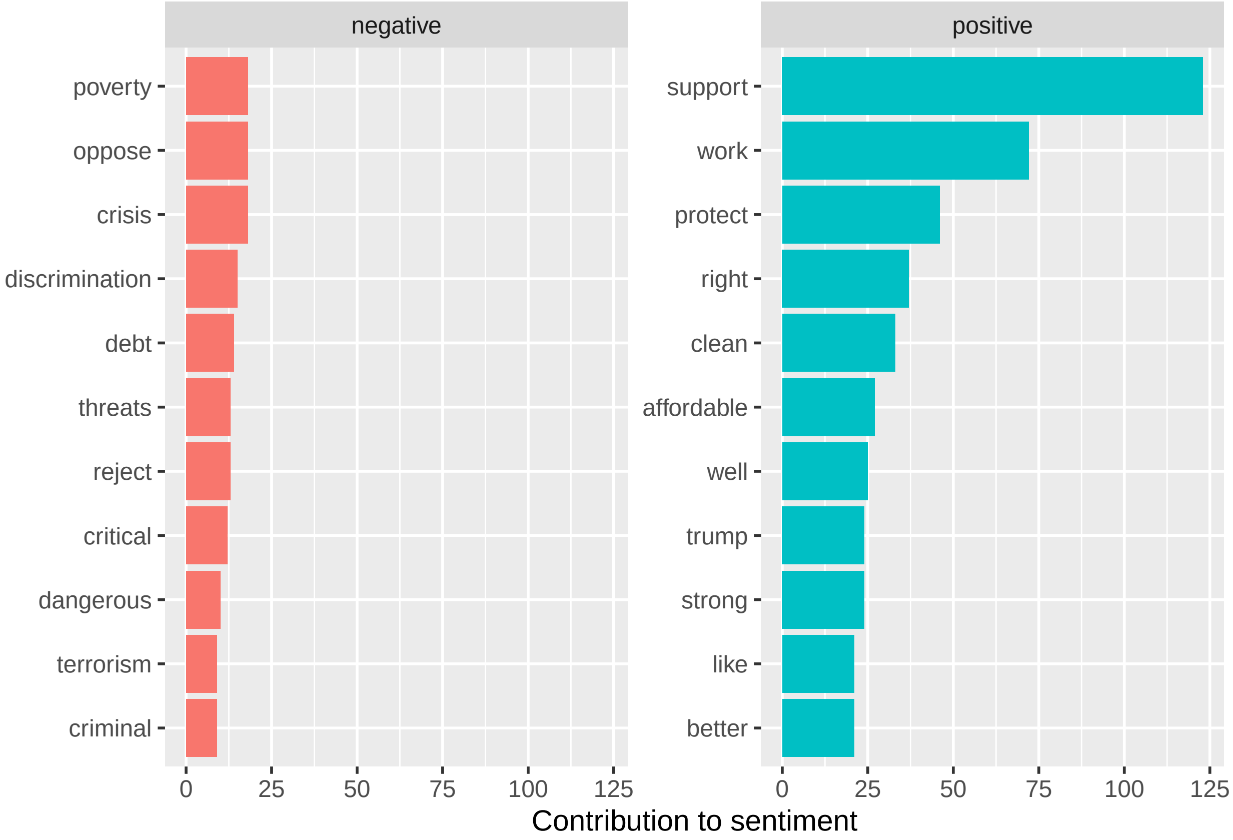
```
## # A tibble: 6 x 3
##   word        n value
##   <chr>   <int> <dbl>
## 1 support   123     2
## 2 care       66     2
## 3 fight      58    -1
## 4 ensure     50     1
## 5 protect    46     1
## 6 help       41     2
```

```
head(tokens_rep %>%
  count(word, sort = TRUE) %>%
  inner_join(get_sentiments("afinn")))
```

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

## # A tibble: 6 x 3
##   word        n value
##   <chr>   <int> <dbl>
## 1 support   100     2
## 2 united     58     1
## 3 freedom    42     2
## 4 growth     38     2
## 5 protect    38     1
## 6 care       37     2
```

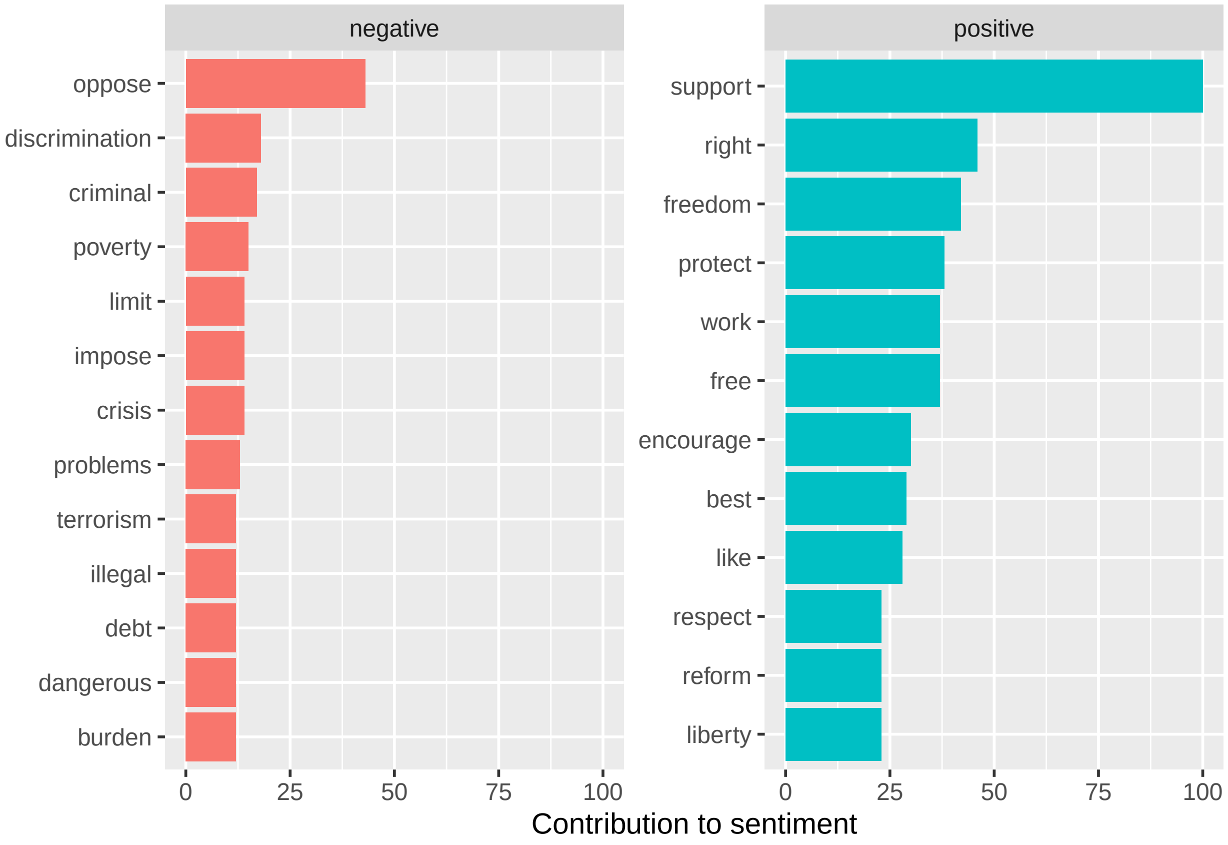
```
#visually assess the output for BING
bing_dem %>%
  group_by(sentiment) %>%
  top_n(10) %>%
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(y = "Contribution to sentiment",
       x = NULL) +
  coord_flip()
```

```
## Selecting by n
```



```
bing_rep %>%
  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 = "Contribution to sentiment",
       x = NULL) +
  coord_flip()
```

```
## Selecting by n
```



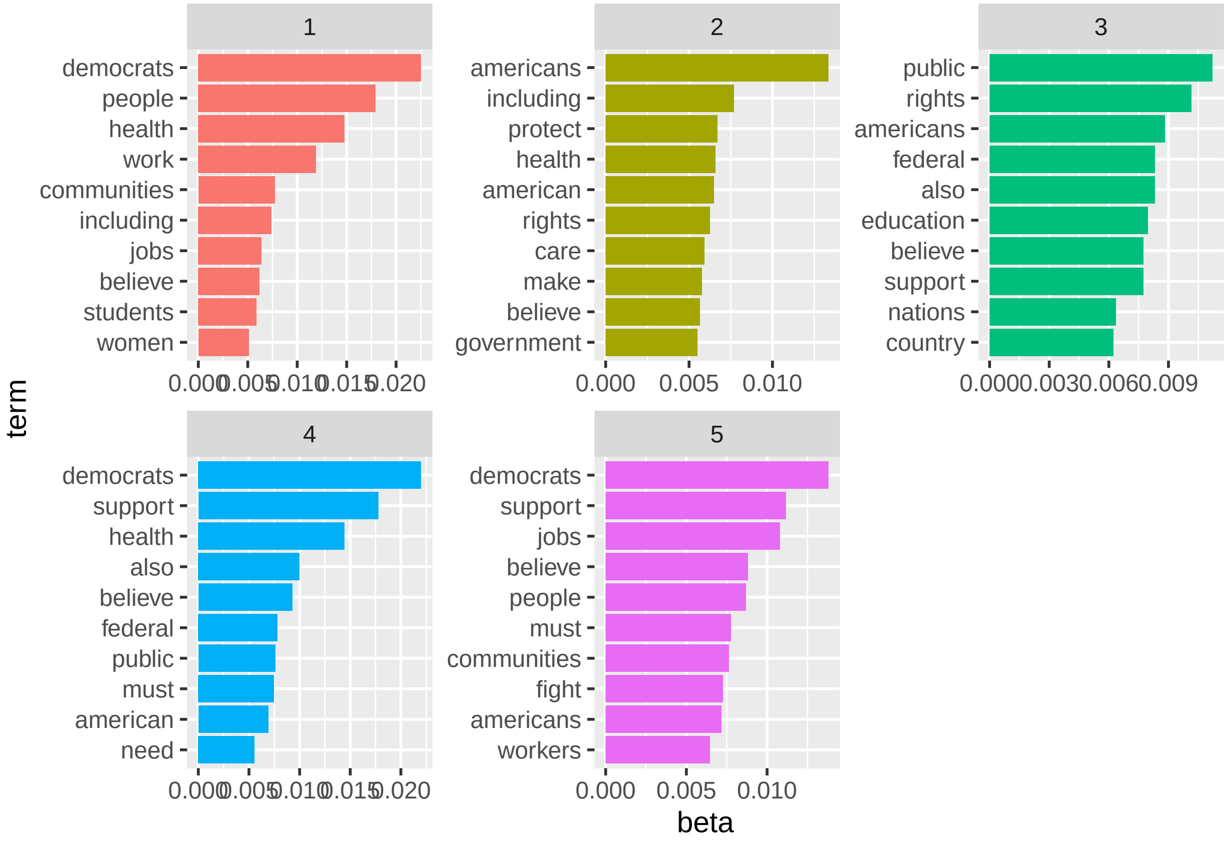
Problem 5:

After looking at the output for the BING and the AFINN dictionaries, it seems that the Democratic platform is slightly more positive than the Republican platform. Based on the BING output, the Republican platform seems to have higher instances of saying they are "opposing" policy than the Democratic party. Furthermore, the Democratic platform seems to use "support" more than the Republican platform. The AFINN output tells us that both parties use positive words like "support" quite frequently (the Democratic party more frequently). The Democratic party also uses "care" alot while the Republican party uses "united". In their top 6 most used words, the Democratic party does have one negative word that they use 58 times which is "fight", but in a policy context this might be a good thing if it is referring to fighting for their goals or their constituents.

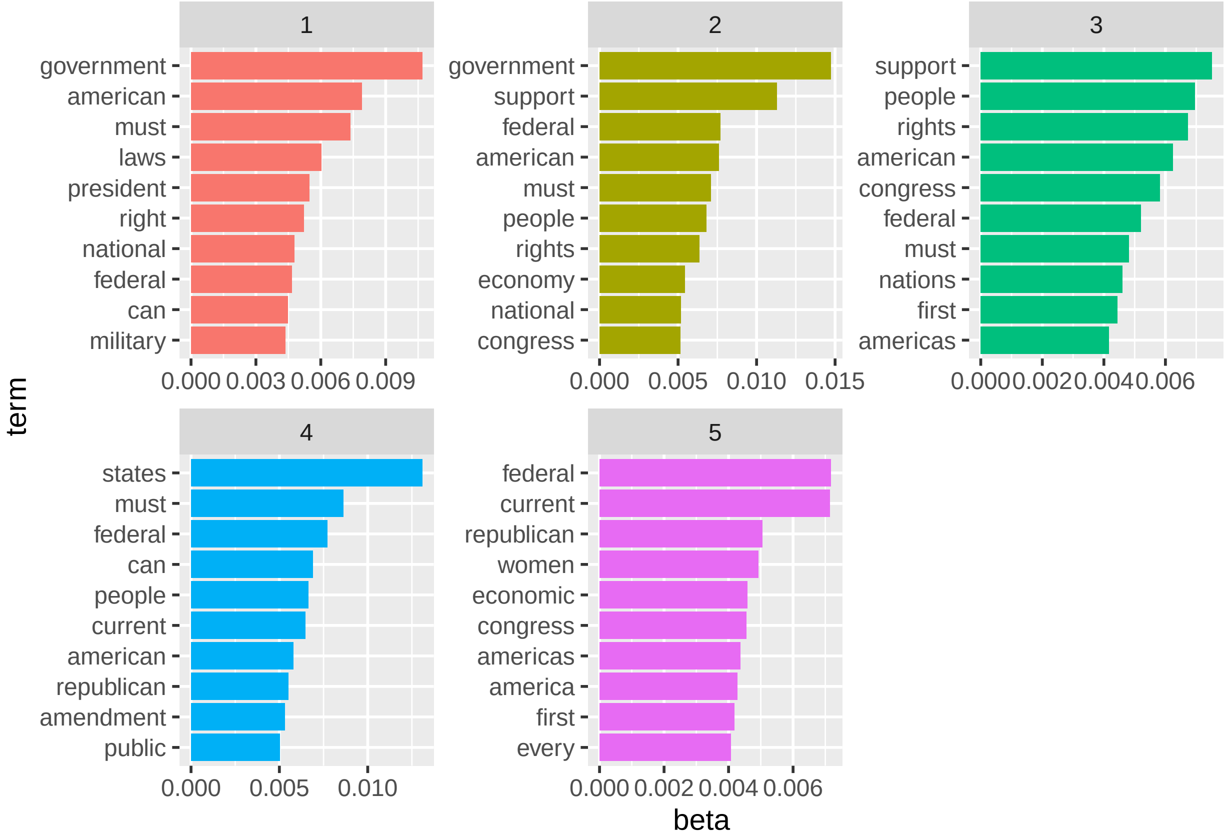
Problem 6:

```
#fitting a topic model to each of the parties
#democratic party
library(topicmodels)
library(tm)
# Preprocessing leaves behind a lot of white space, or extra spaces between words or lines
plat <- tm_map(plat, stripWhitespace)
plat <- tm_map(plat, PlainTextDocument) # final redefine for retaining the lataest preprocessing steps

dtm <- DocumentTermMatrix(plat[1])


dem_lda = LDA(dtm, k=5, control=list(seed=1234))

library(tidytext)
```

```
dem_tidy = tidy(dem_lda, matrix="beta")


dtm1 <- DocumentTermMatrix(plat[2])

rep_lda = LDA(dtm1, k=5, control=list(seed=1234))
rep_tidy = tidy(rep_lda, matrix="beta")


#presenting the results of the topic models

dem_terms <- dem_tidy %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

dem_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") +
  coord_flip() +
  scale_x_reordered()
```



```
rep_terms <- rep_tidy %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

rep_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") +
  coord_flip() +
  scale_x_reordered()
```



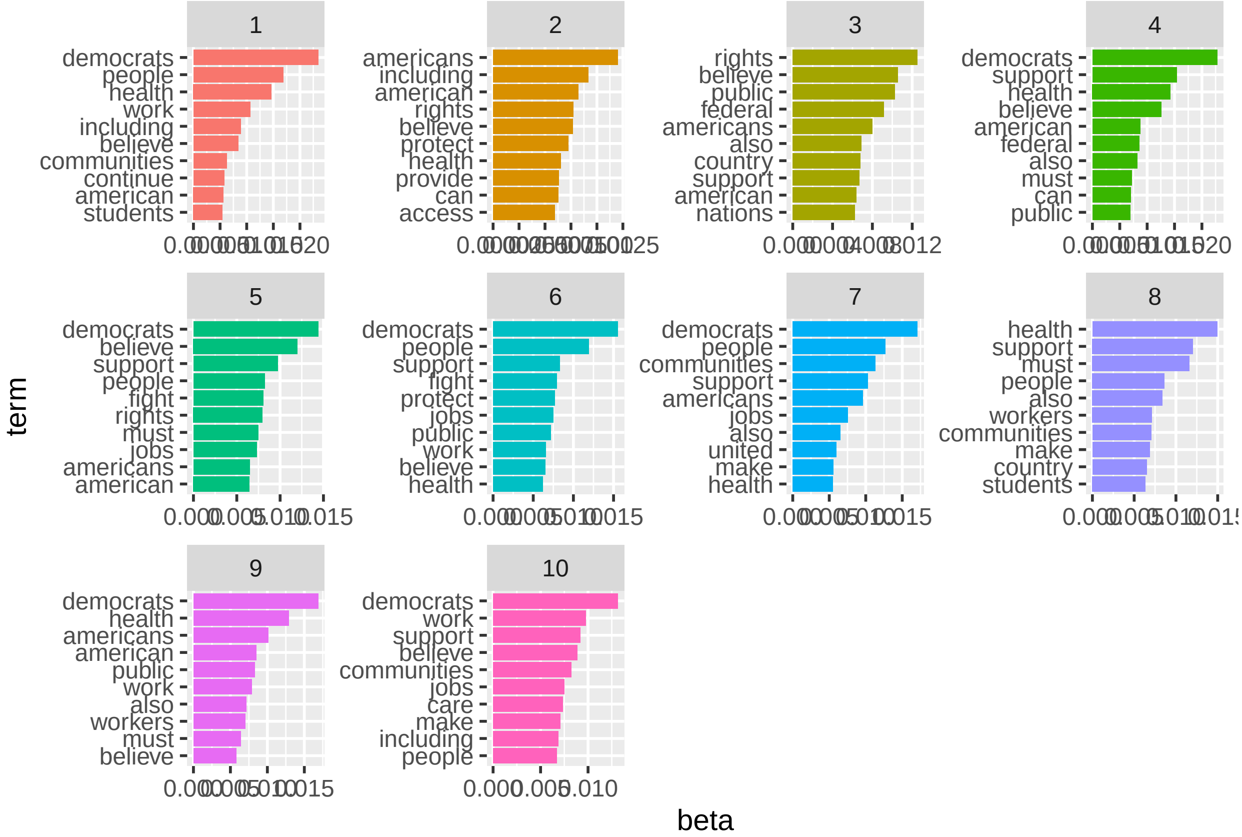
Problem 7:

The output of the LDA topic models for each of the parties reveal some key differences between the two parties' platforms that were not as noticeable in the word clouds earlier. First, in the 5 topics for democrats it seems that Democrats focus more on policy issues like "health", "jobs", "education", while in the Republican manifestos it seems they focus more on national rhetoric and government jargon using terms related to our democracy like "federal", "congress", "american", and "rights". This might be reflective on the more traditional values of the Republican party. Some similarities between the parties still exist though. Both parties' have topics that include frequent uses of "american" and "support".

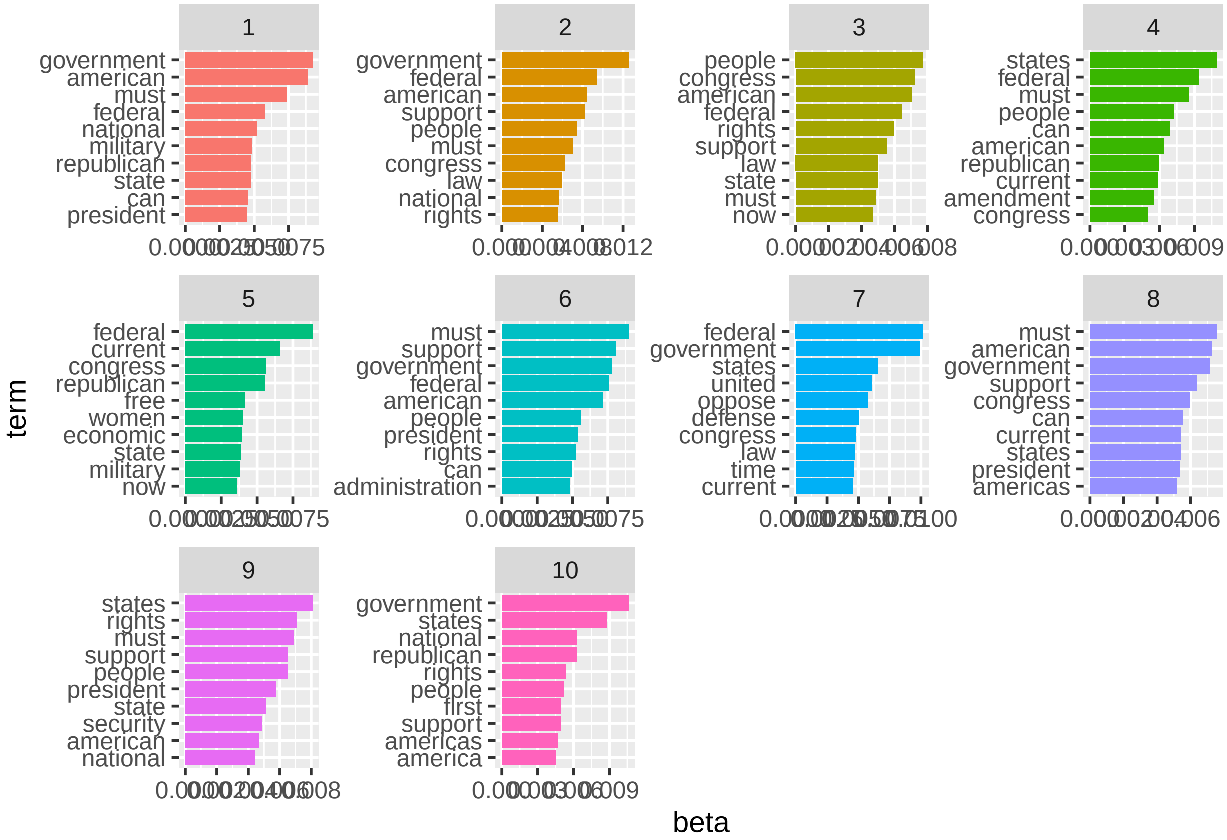
Problem 8:

```
dtm <- DocumentTermMatrix(plat[1])

dem_lda10 = LDA(dtm, k=10, control=list(seed=1234))
```

```
library(tidytext)
dem_tidy10 = tidy(dem_lda10, matrix="beta")


dtm1 <- DocumentTermMatrix(plat[2])

rep_lda10 = LDA(dtm1, k=10, control=list(seed=1234))
rep_tidy10 = tidy(rep_lda10, matrix="beta")


#presenting the results of the topic models

dem_terms10 <- dem_tidy10 %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

dem_terms10 %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered()
```

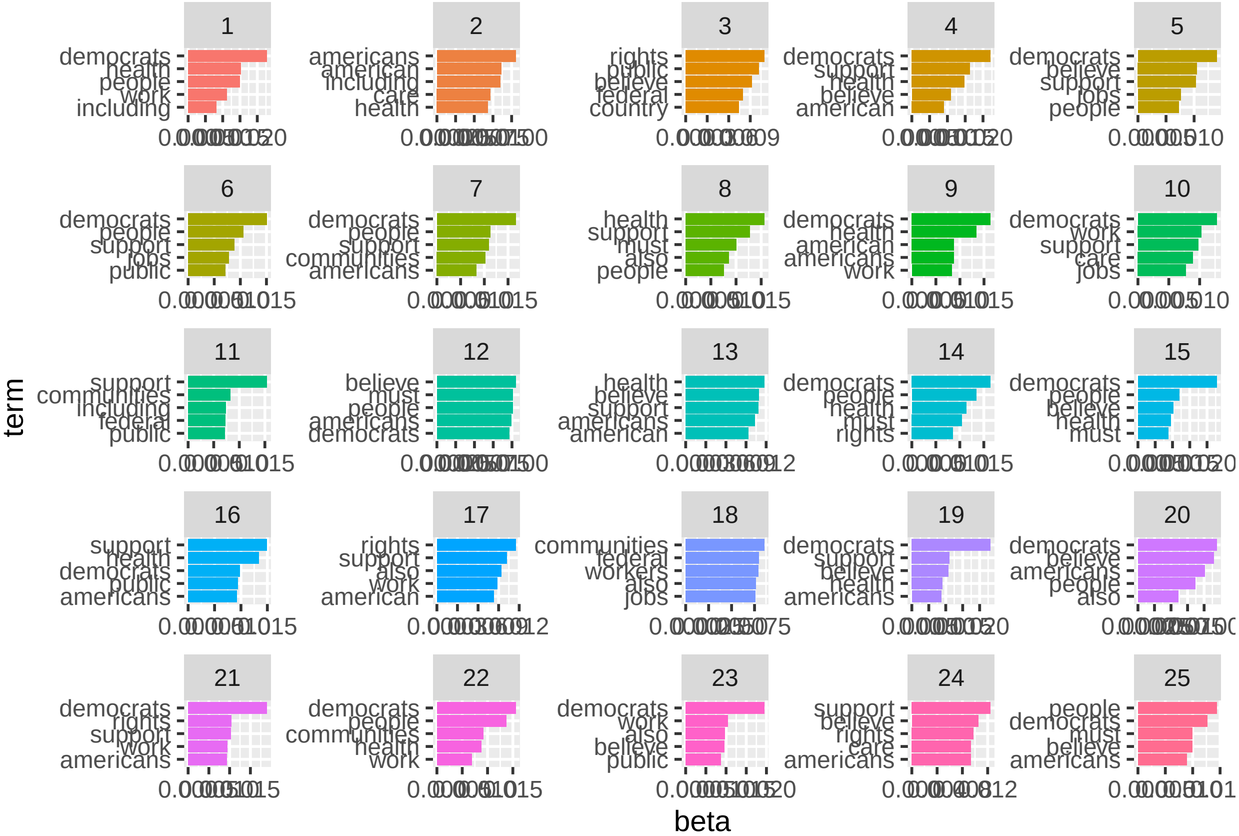


```r
rep_terms10 <- rep_tidy10 %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

rep_terms10 %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered()
```

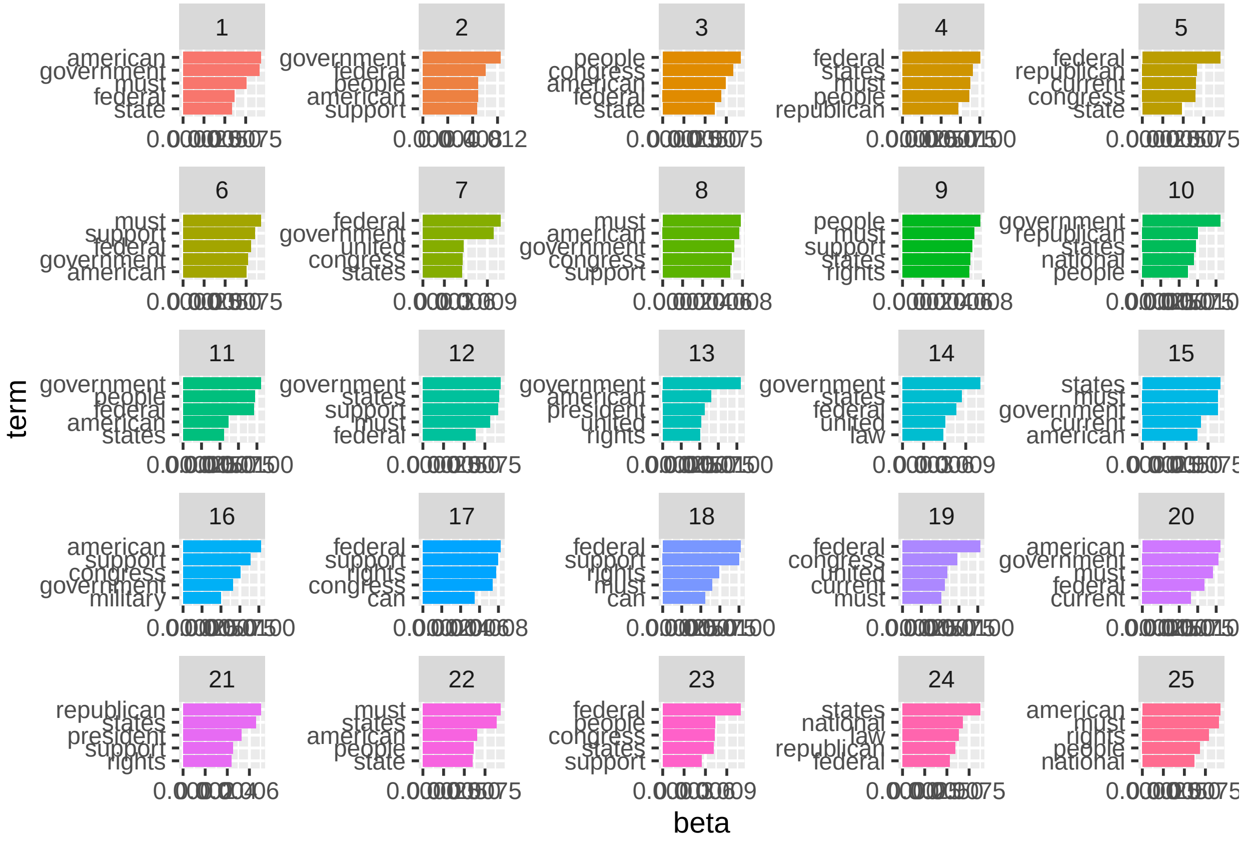


```r
dtm <- DocumentTermMatrix(plat[1])


dem_lda25 = LDA(dtm, k=25, control=list(seed=1234))

library(tidytext)
dem_tidy25 = tidy(dem_lda25, matrix="beta")


dtm1 <- DocumentTermMatrix(plat[2])

rep_lda25 = LDA(dtm1, k=25, control=list(seed=1234))
rep_tidy25 = tidy(rep_lda25, matrix="beta")
```

```r
#presenting the results of the topic models

dem_terms25 <- dem_tidy25 %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

dem_terms25 %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered()
```



```r
rep_terms25 <- rep_tidy25 %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

rep_terms25 %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered()
```



Problem 9:

```
#calculate perplexity of each model iteration
#for Dems
perplexity(dem_lda)
```

```
## [1] 1605.026
```

```
perplexity(dem_lda10)
```

```
## [1] 1605.656
```

```
perplexity(dem_lda25)
```

```
## [1] 1609.679
```

```
#for Reps
perplexity(rep_lda)
```

```
## [1] 2321.121
```

```
perplexity(rep_lda10)
```

```
## [1] 2322.482
```

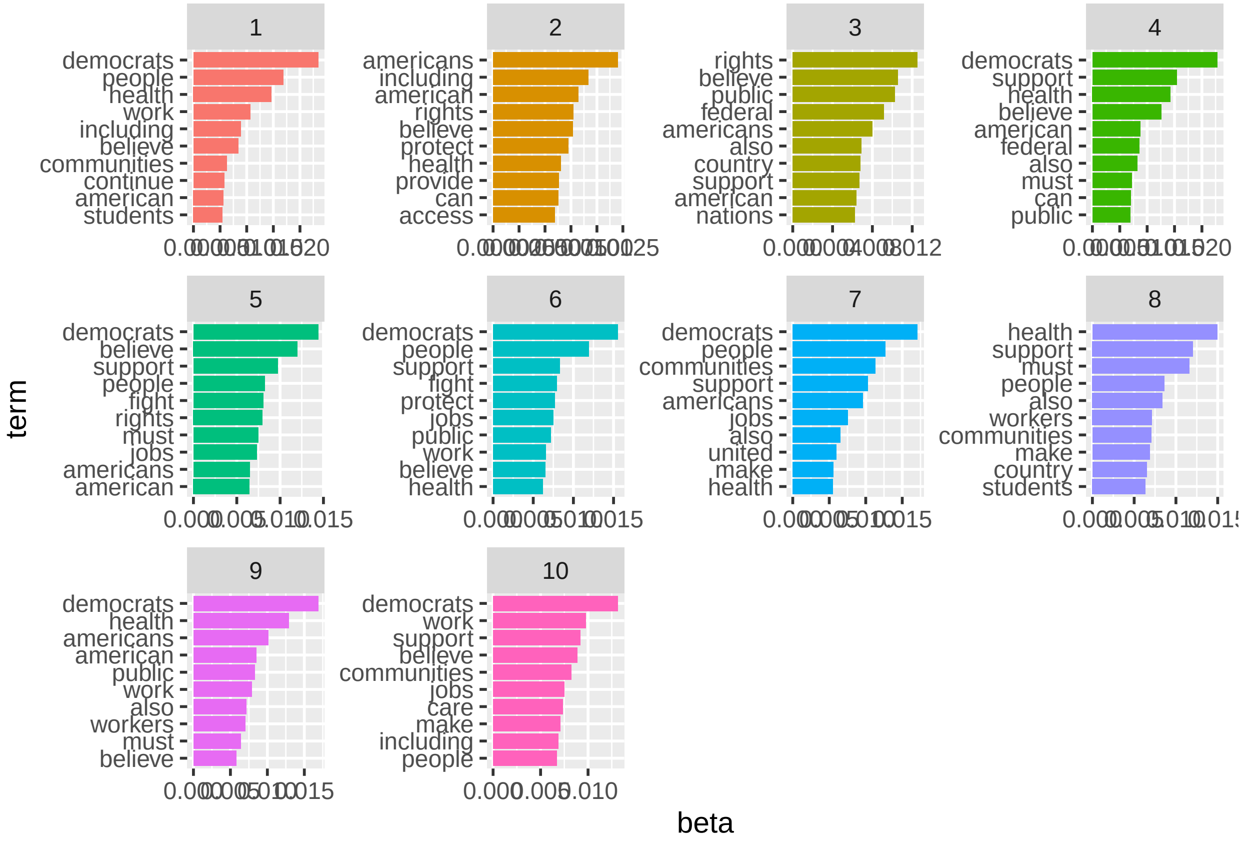
```r
perplexity(rep_lda25)
```

```
## [1] 2327.817
```

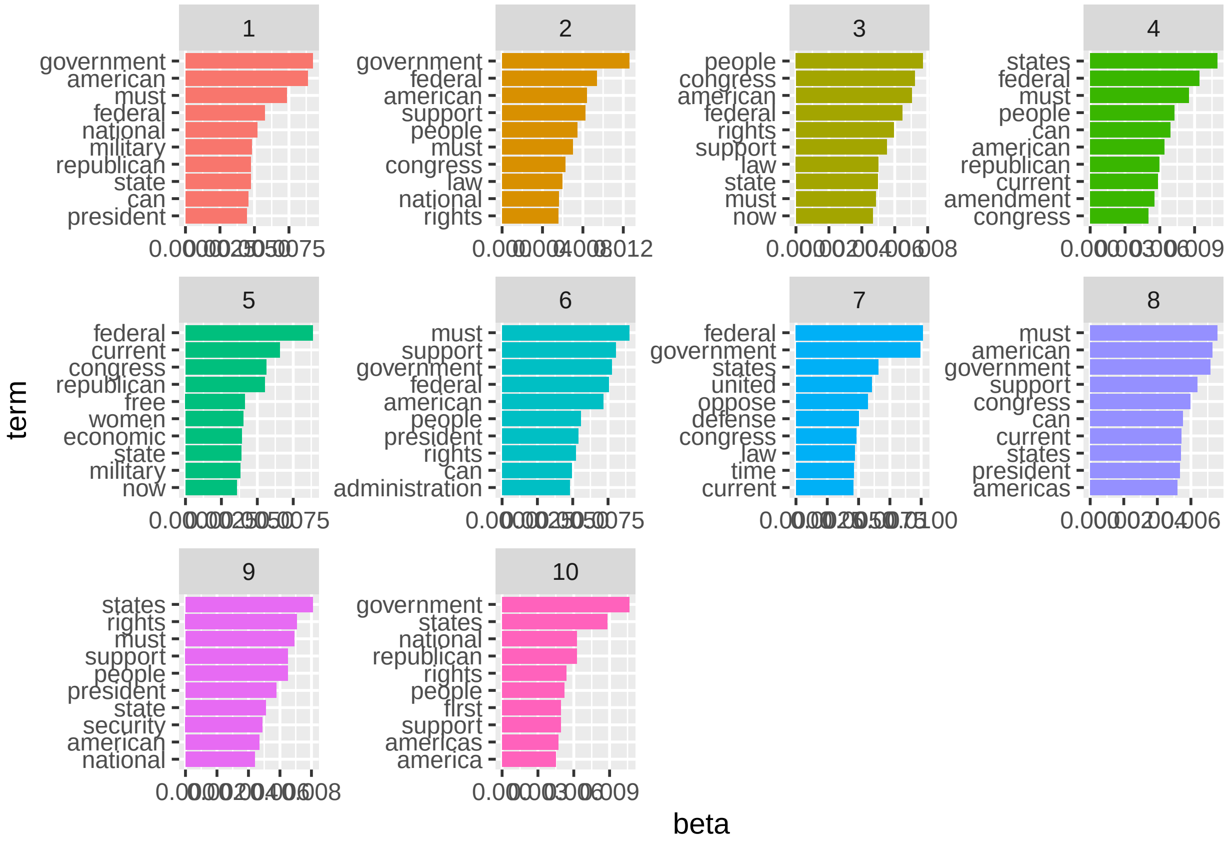
Since the model with the lowest perplexity score is generally considered to be the best model, I find in this analysis that the models with k-5 topics characterizing its tokens is the best fit for both the Republican and Democratic platforms.

Problem 10:

```r
dtm <- DocumentTermMatrix(plat[1])


dem_lda10 = LDA(dtm, k=10, control=list(seed=1234))

library(tidytext)
dem_tidy10 = tidy(dem_lda10, matrix="beta")


dtm1 <- DocumentTermMatrix(plat[2])

rep_lda10 = LDA(dtm1, k=10, control=list(seed=1234))
rep_tidy10 = tidy(rep_lda10, matrix="beta")


#presenting the results of the topic models

dem_terms10 <- dem_tidy10 %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

dem_terms10 %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered()
```



```
rep_terms10 <- rep_tidy10 %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

rep_terms10 %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered()
```



The k=10 topic model does not really contribute much more information about the differences between the parties than the simpler k=5 topic model that we favored in the previous questions. Alot of the topics seem to contain similar words for the Republican party especiallly like "federal", "government", and "states" and each additional topic is not telling us much more about the party. For the Democratic model, the k=10 topic model do give us more information on policy spheres that Democrats may be interested than the k=5 topic model. For instance, "students" and "jobs" and "workers" show up more in the additional topics in the k=10 model than the k=5. Overall, I do not think this model picks up differences more efficienctly since for the most part the new topics just contain new combinations of most of the words in the k=5 topics.

Problem 11:

Something that I value in selecting a party to affiliate myself with would be the extent to which the party has substantial policy goals and a positive outlook towards the future. Over the course of this analysis, I have noticed that in the topics and frequencies of the words in the Democratic manifestos, it seems that Democrats place more emphasis on policy related to jobs, health, students, women, and communities which are all things that I value. I also found over this analysis that the Democratic party was more optimistic and positive and focused more on what they supported than what they opposed. The Republican party on the other hand had more mentions of things they were against rather than things they were in support of.