PS 5 WFP

EDA

See pdf from python to see my answers to questions 1 & 2.

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
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------
## v ggplot2 3.2.1          v purrr   0.3.2
## v tibble  2.1.3          v dplyr   0.8.3
## v tidyr   0.8.99.9000    v stringr 1.4.0
## v readr   1.3.1          v forcats 0.4.0

## -- Conflicts ------------------------------------------------------------- ti
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(wordcloud)
```

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

```r
library(tidytext)

library(topicmodels)
library(tm)
```

```
## Loading required package: NLP

##
## Attaching package: 'NLP'

## The following object is masked from 'package:ggplot2':
##
##     annotate
```

```r
dem_plat <- read_csv("dem_doc_matrix.csv") %>%
  select(-X1)
```

```
## Warning: Missing column names filled in: 'X1' [1]

## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   word = col_character(),
##   count = col_double()
## )
```
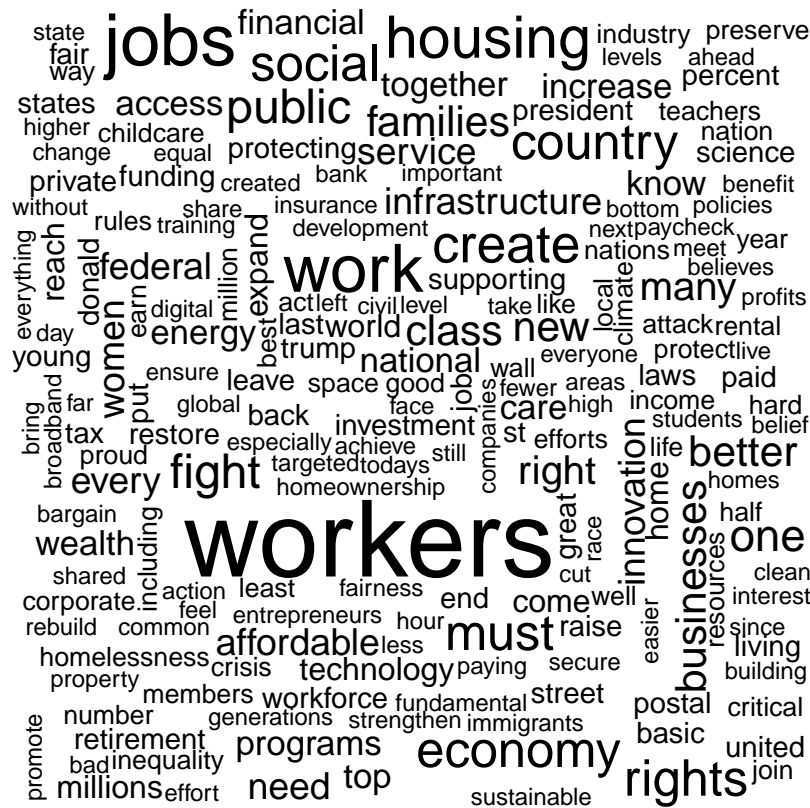
```
rep_plat <- read_csv("repub_doc_matrix.csv") %>%
  select(-X1)
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   word = col_character(),
##   count = col_double()
## )
```

Democratic platform wordcloud

```
wordcloud(dem_plat$word, dem_plat$count)
```


Republican party word cloud

```
wordcloud(rep_plat$word, rep_plat$count)
```

The democratic platform seems to emphasize the economy, people, support, and rights. The republican platform emphasizes laws, finanical institutions, growth and more administration/federal type terms.

SENTIMENT ANALYSIS

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).

Get the sentiment scores

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

affin <- get_sentiments("afinn")
```

Mean affin score by party

Democrats
```
mean(left_join(dem_plat, affin)$value, na.rm = TRUE)
```

```
## [1] 0.4971098
```
```
mean(left_join(rep_plat, affin)$value, na.rm = TRUE)
```

```
## [1] 0.270936
```

Bing Score distribution by party

Democratic

```
table(left_join(dem_plat, bing)$sentiment)
```

```
##
## negative positive
##       88      114
```

Republican

```
table(left_join(rep_plat, bing)$sentiment)
```

```
##
## negative positive
##      122      130
```

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?

The democratic party platform has a higher average AFFIN score and a higher proportion of bing "positive" terms than the republican party. This fits with my conception that the democratic party is more progressive (in the sense of change/future will be better than today) compared to the more conservative republican party (things were better historically, change/progressive ideas about the future are bad)

TOPIC MODELS

6. With a general sense of sentiments of the party platforms (i.e., the tones related to how parties talk about their roles in the political future), 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).

```
dem_tdm <- dem_plat %>%
  pivot_wider(names_from = "word", values_from = "count")


topic_and_display <- function(tdm, k){
  LDA <- LDA(tdm, k = k)

  LDA %>%
    tidy(matrix = "beta") %>%
    group_by(topic) %>%
    top_n(10, beta) %>%
    ungroup() %>%
    arrange(topic, -beta)
```

```
}

topic_and_plot <- function(tdm, k){
  LDA <- LDA(tdm, k = k)

  top_10 <- LDA %>%
    tidy(matrix = "beta") %>%
    group_by(topic) %>%
    top_n(10, beta) %>%
    ungroup() %>%
    arrange(topic, -beta)

  top_10 %>%
    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()
}


topic_and_plot(dem_tdm, k =5)
```
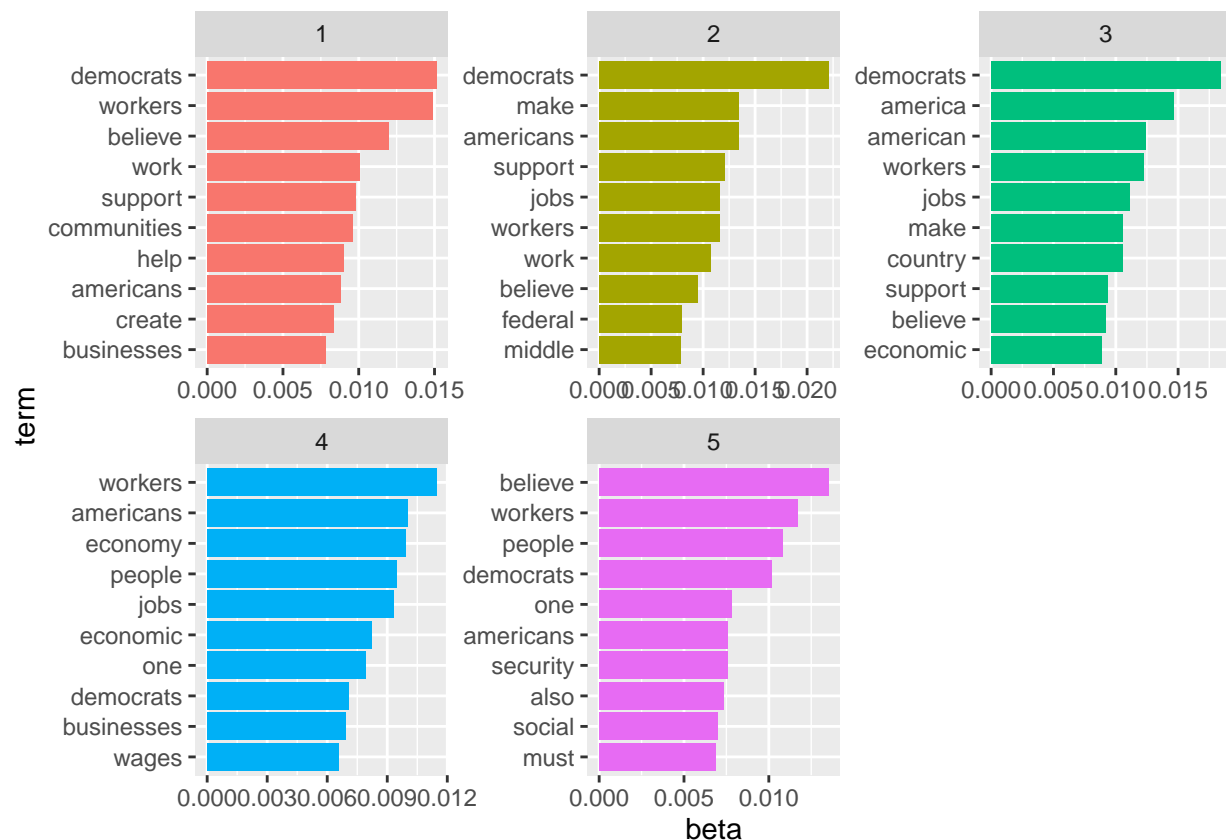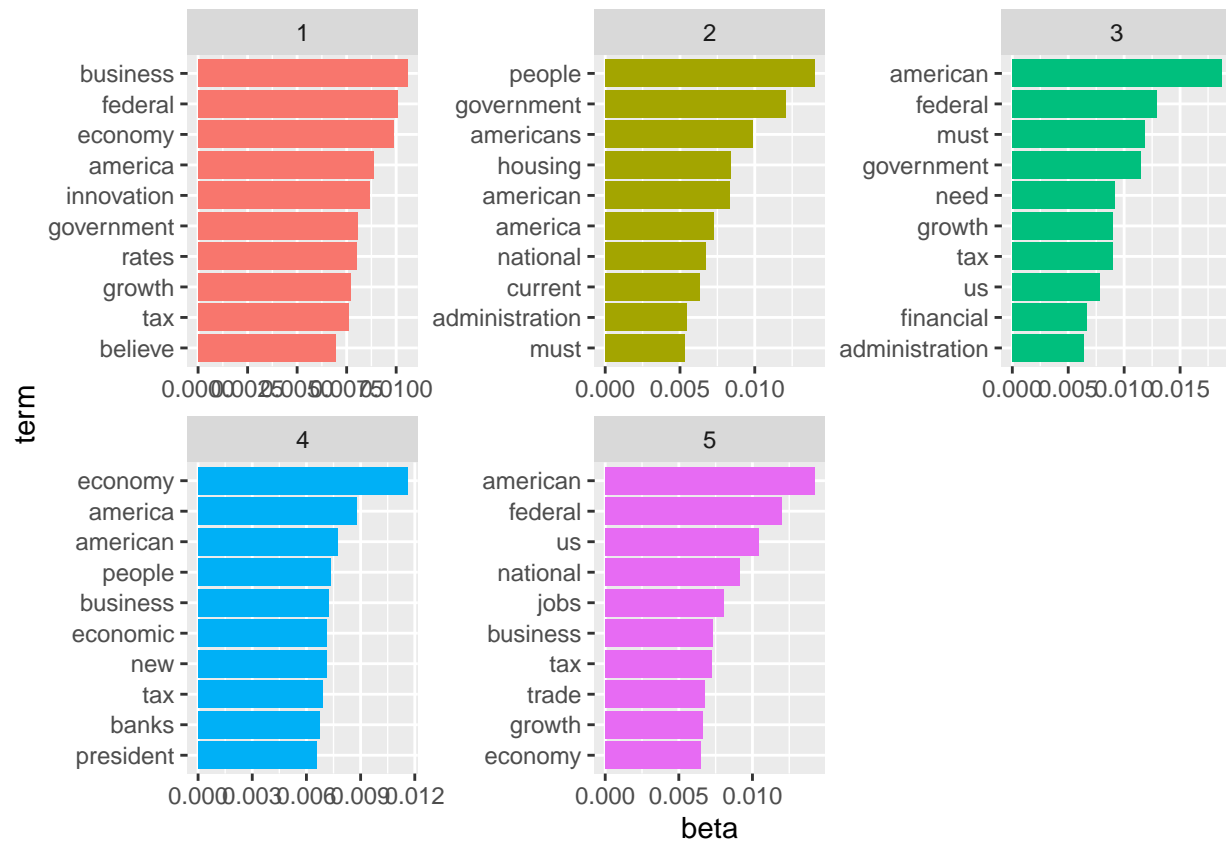


```
rep_tdm <- rep_plat %>%
  pivot_wider(names_from = "word", values_from = "count")
```
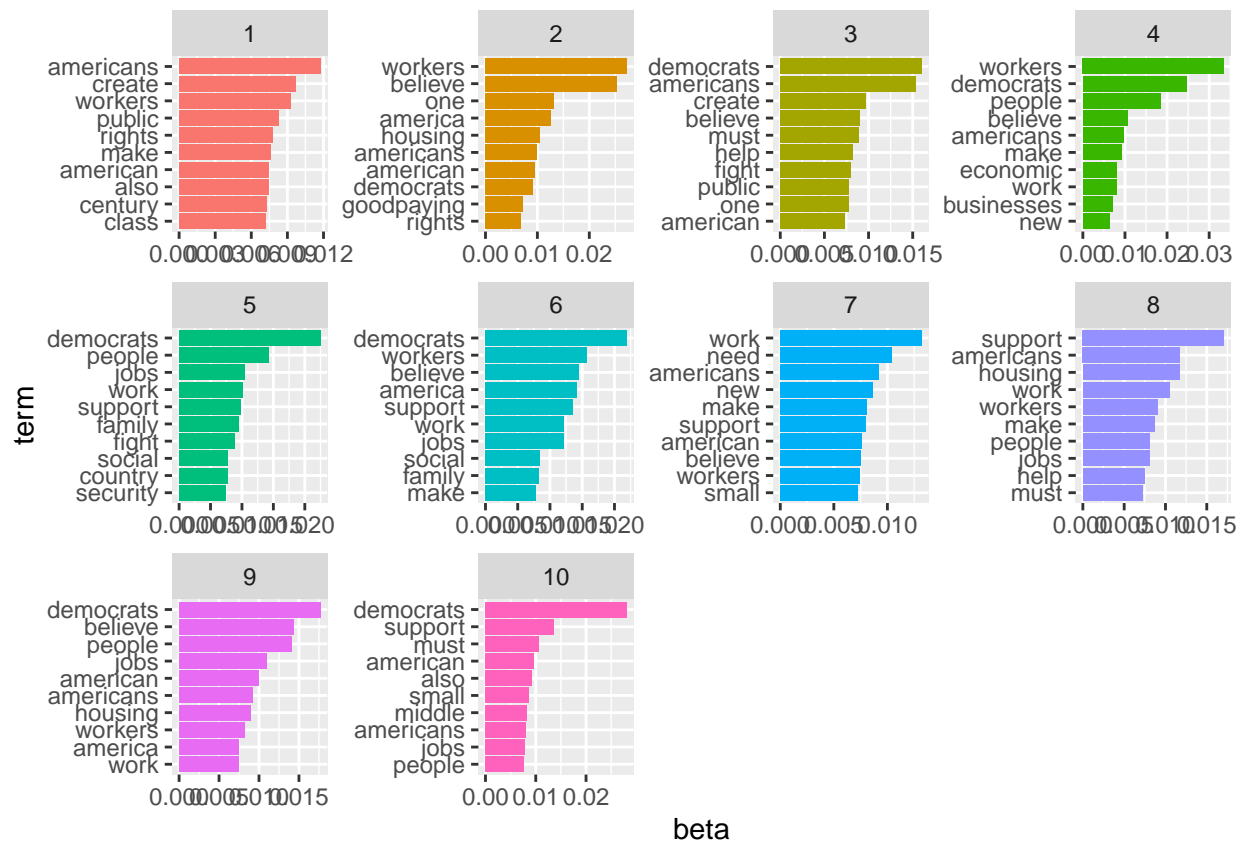
```
topic_and_plot(rep_tdm, k = 5)
```


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

For the democrats, the topics seem to me to be focused on creating jobs and supporting americans who work

For the republicans, the focus is much more on trade, businesses, banks/institutions, and the economy at large.

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).

```
topic_and_plot(dem_tdm, k =10)
```
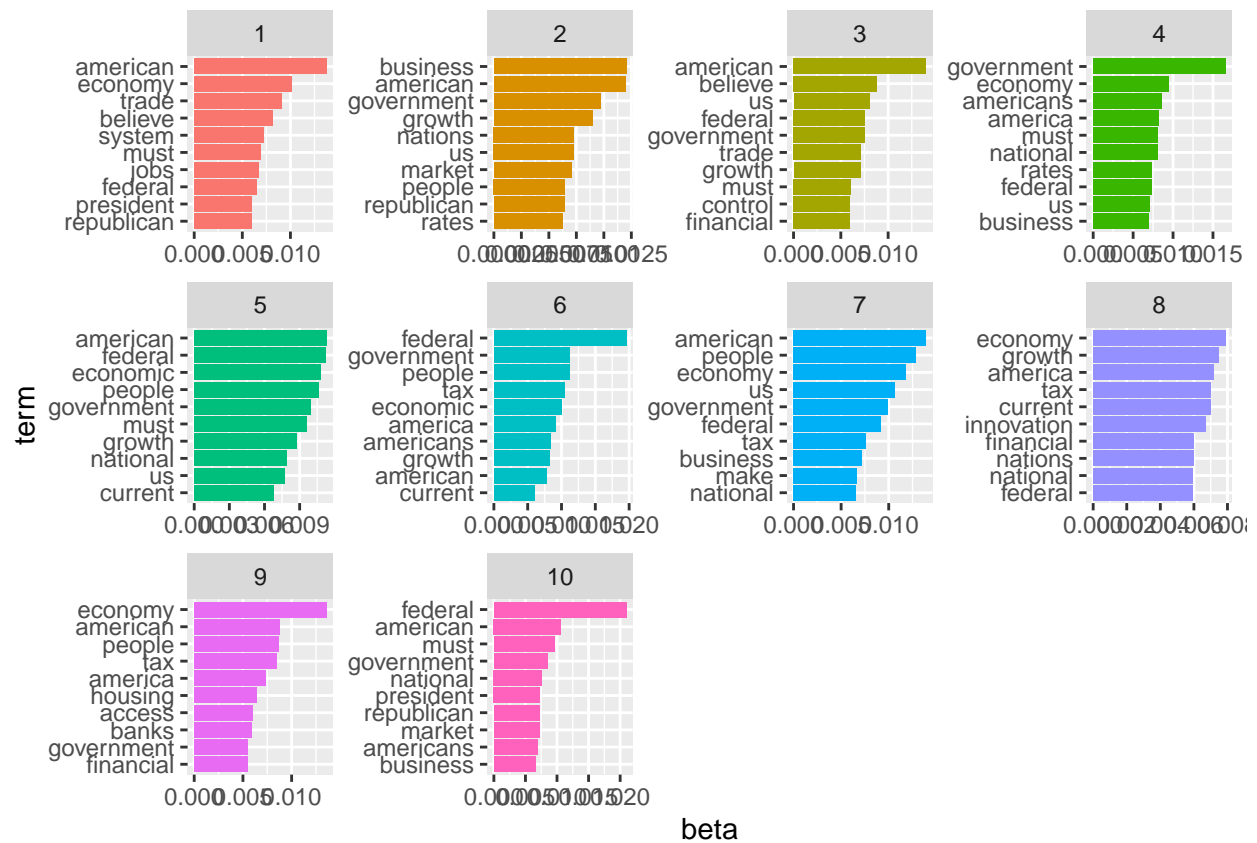
```
topic_and_plot(dem_tdm, k =25)
```

```
topic_and_plot(rep_tdm, k =10)
```

```
topic_and_plot(rep_tdm, k =25)
```

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

```
topic_perplex <- function(tdm, k){
  LDA <- LDA(tdm, k = k)
  perplexity(LDA)
}
```

```
for (k in c(5, 10, 25)) {
  print(topic_perplex(dem_tdm, k))
}
```

```
## [1] 808.6276
## [1] 808.7799
## [1] 809.1333
```

The democratic topic model is minimized at k = 5

```
for (k in c(5, 10, 25)) {
  print(topic_perplex(rep_tdm, k))
}
```

```
## [1] 1033.55
## [1] 1031.157
## [1] 1031.573
```

The repbulican topic model is minimized at k = 5 as well.

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?

```
topic_and_plot(dem_tdm, k =10)
```
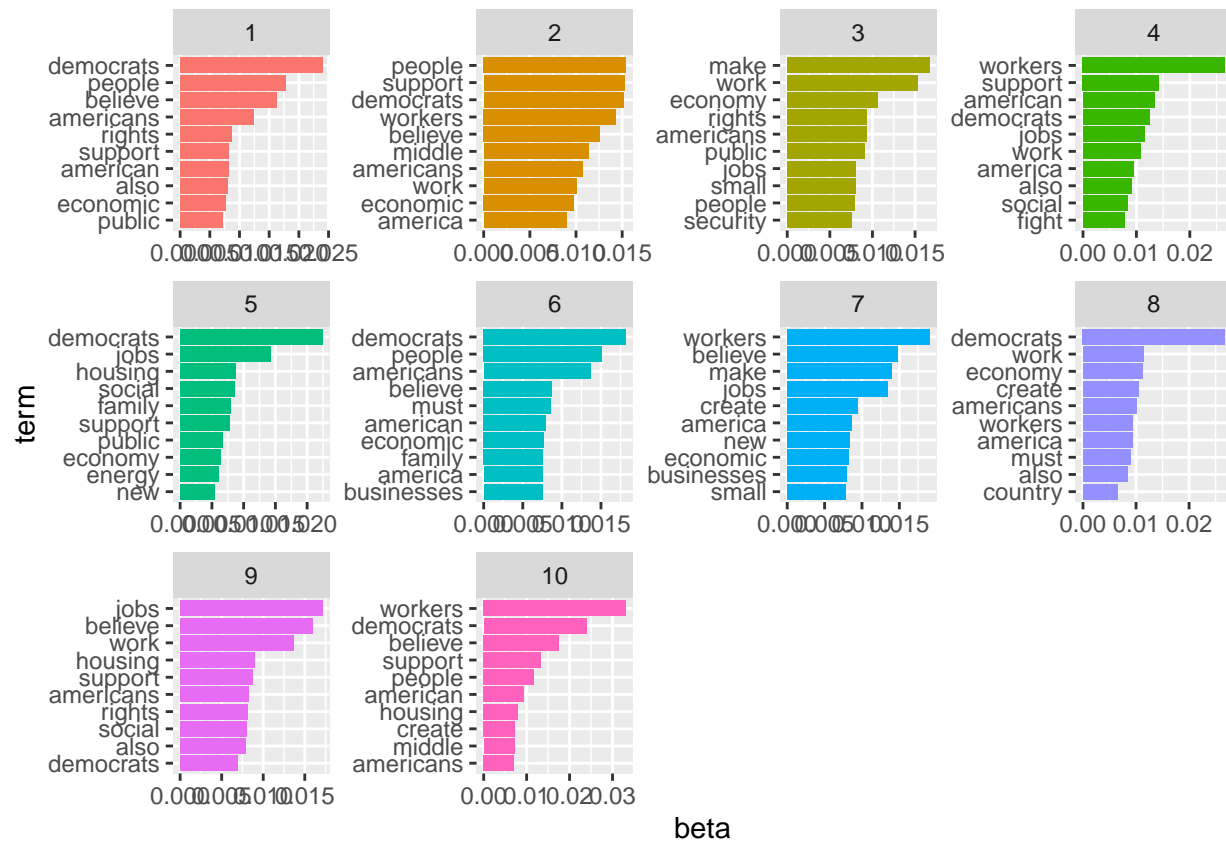


```
topic_and_plot(rep_tdm, k =10)
```
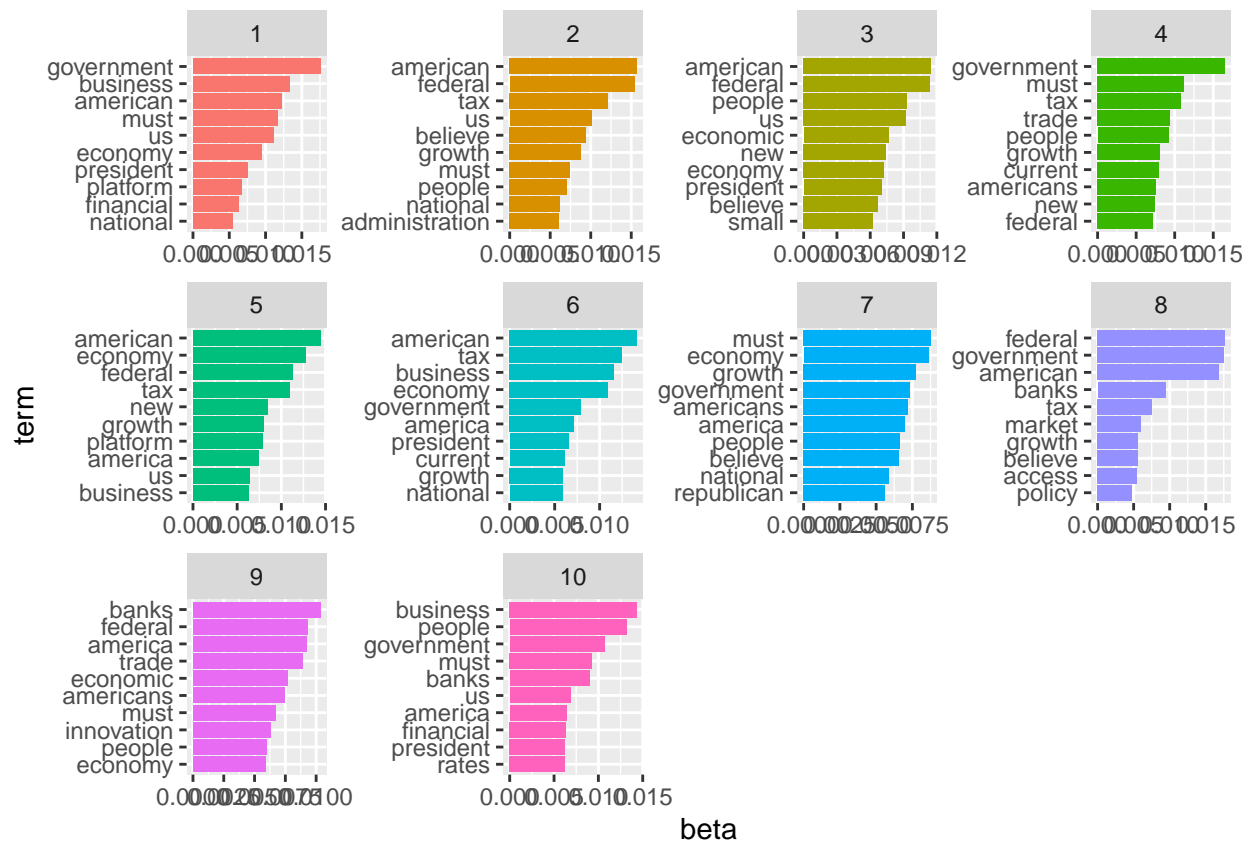
A major theme is that the democratic topics revolve around workers and wages (people level) and the republican topics are focused on higher levels (businesses, federal, taxes, the economy at large).

I think this major distinction was fairly apparent at k = 5

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)?

I guess I'm into the democrats because the analyses seem to have revealed they are more focused on helping people directly, rather than abstract concepts/institutions. However, that could just be my biases leading to subjective interpretation of the results.