PS3 - MACS40500

Alec MacMillen

11/25/2019

GENERAL NLP/PREPROCESSING

Question 1

```
# Load CSV of party platform data
platforms <- read_csv("Party Platforms Data/platforms.csv")
```

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

```
# Create
dem <- VCorpus(VectorSource(platforms$platform[1]))
rep <- VCorpus(VectorSource(platforms$platform[2]))

# Get summary of VCorpus objects
summary(dem)
```

```
##   Length Class             Mode
## 1 2      PlainTextDocument list
```

```
summary(rep)
```

```
##   Length Class             Mode
## 1 2      PlainTextDocument list
```

```
# Visually inspect the first few lines
writeLines(substr(as.character(dem[[1]]), 1, 300))
```

```
## In 2016, Democrats meet in Philadelphia with the same basic belief that animated the Continental Cong
##
## Under President Obama's leadership, and thanks to the hard work and determination of the American pe
```

Question 2

Create a document-term matrix and preprocess the platforms. Convert to lowercase, remove stopwords, remove numbers, and remove punctuation.

```r
# Remove punctuation
dem <- tm_map(dem, removePunctuation)
rep <- tm_map(rep, removePunctuation)
writeLines(substr(as.character(dem[[1]]), 1, 300))
```

```
## In 2016 Democrats meet in Philadelphia with the same basic belief that animated the Continental Cong
##
## Under President Obamas leadership and thanks to the hard work and determination of the American peopl
```

```r
# Remove more unique characters as necessary
for (j in seq(dem)) {
  dem[[j]] <- gsub("/", " ", dem[[j]])
  dem[[j]] <- gsub("â ", " ", dem[[j]])
  dem[[j]] <- gsub("@", " ", dem[[j]])
  dem[[j]] <- gsub("/u2028", " ", dem[[j]])
  dem[[j]] <- gsub("Ã¡", "a", dem[[j]])
  dem[[j]] <- gsub("â€", " ", dem[[j]])
}

for (j in seq(rep)) {
  rep[[j]] <- gsub("/", " ", rep[[j]])
  rep[[j]] <- gsub("â ", " ", rep[[j]])
  rep[[j]] <- gsub("@", " ", rep[[j]])
  rep[[j]] <- gsub("/u2028", " ", rep[[j]])
  rep[[j]] <- gsub("Ã¡", "a", rep[[j]])
  rep[[j]] <- gsub("â€", " ", rep[[j]])
}

writeLines(substr(as.character(dem[[1]]), 1, 300))
```

```
## In 2016 Democrats meet in Philadelphia with the same basic belief that animated the Continental Cong
##
## Under President Obamas leadership and thanks to the hard work and determination of the American peopl
```

```r
# Remove numbers, change to lowercase, and convert to plain text
dem <- tm_map(dem, removeNumbers)
dem <- tm_map(dem, tolower)

rep <- tm_map(rep, removeNumbers)
rep <- tm_map(rep, tolower)

writeLines(substr(as.character(dem[[1]]), 1, 300))
```

```
## in  democrats meet in philadelphia with the same basic belief that animated the continental congress
##
## under president obamas leadership and thanks to the hard work and determination of the american peopl
```

```r
# Remove stopwords
dem <- tm_map(dem, removeWords, stopwords("english"))
rep <- tm_map(rep, removeWords, stopwords("english"))

dem <- tm_map(dem, removeWords, c("will"))
```

```r
rep <- tm_map(rep, removeWords, c("will"))

writeLines(substr(as.character(dem[[1]]), 1, 300))
```

```
##   democrats meet  philadelphia    basic belief  animated  continental congress   gathered   years ag
##
##  president obamas leadership  thanks   hard work  determination   american people   come  long way
```

```r
# Stem the documents, strip whitespace, and convert to plaintext
dem_st <- tm_map(dem, stemDocument)
rep_st <- tm_map(rep, stemDocument)

# Strip whitespace
dem <- tm_map(dem, stripWhitespace)
dem_st <- tm_map(dem_st, stripWhitespace)
rep <- tm_map(rep, stripWhitespace)
rep_st <- tm_map(rep_st, stripWhitespace)

# Convert to plaintext
dem_st <- tm_map(dem_st, PlainTextDocument)
dem <- tm_map(dem, PlainTextDocument)
rep_st <- tm_map(rep_st, PlainTextDocument)
rep <- tm_map(rep, PlainTextDocument)

writeLines(substr(as.character(dem[[1]]), 1, 300))
```

```
##  democrats meet philadelphia basic belief animated continental congress gathered years ago many one
```

```r
# Convert to document-term matrix
# Documents are rows; terms are columns
dem_dtm <- DocumentTermMatrix(dem)
dem_dtm_st <- DocumentTermMatrix(dem_st)
rep_dtm <- DocumentTermMatrix(rep)
rep_dtm_st <- DocumentTermMatrix(rep_st)

# Find frequencies
freq_dem <- sort(colSums(as.matrix(dem_dtm)), decreasing=TRUE)
freq_dem_st <- sort(colSums(as.matrix(dem_dtm_st)), decreasing=TRUE)
freq_rep <- sort(colSums(as.matrix(rep_dtm)), decreasing=TRUE)
freq_rep_st <- sort(colSums(as.matrix(rep_dtm_st)), decreasing=TRUE)

head(freq_dem)
```

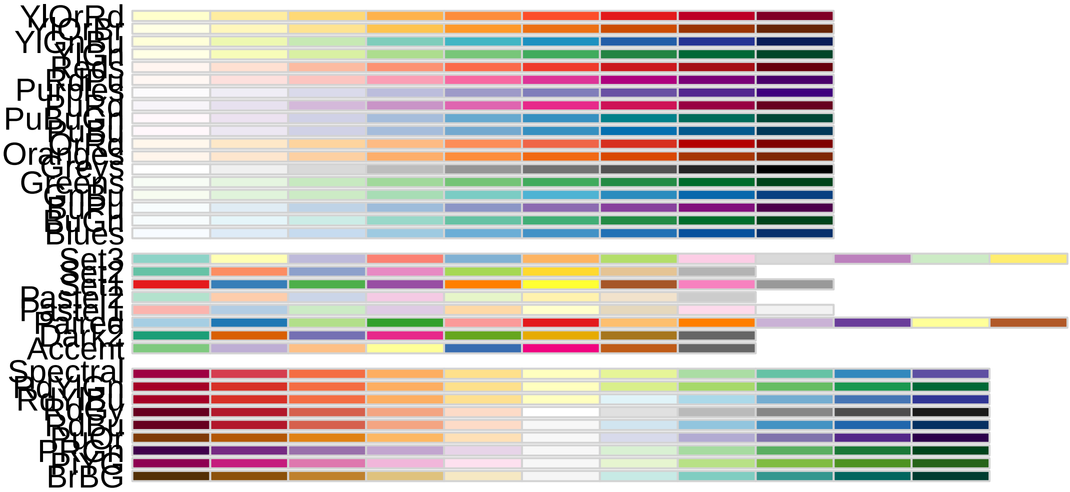
```
## democrats   workers   believe americans    people   support
##        46        36        29        26        24        24
```

```r
head(freq_rep)
```

```
##   american    federal government    economy     people        tax
##         28         26         25         19         19         18
```

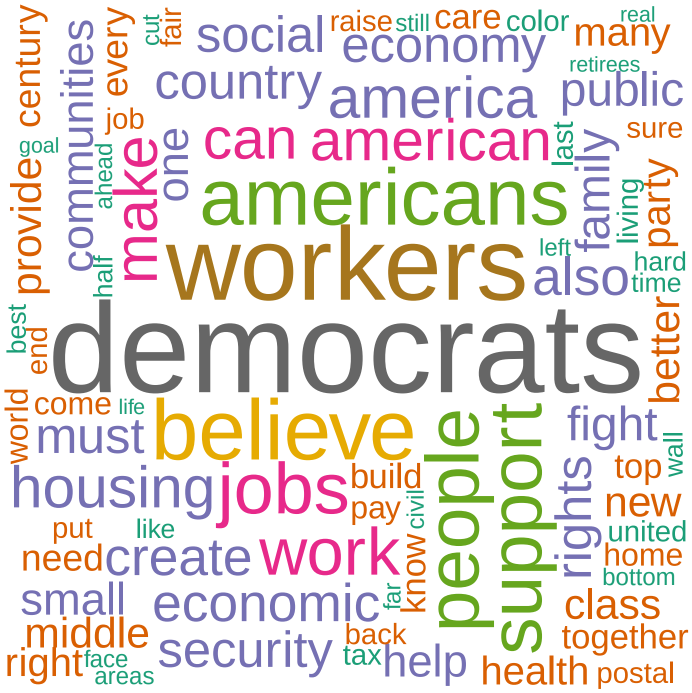
Even with a cursory look at the most commonly used words between the two parties, it is notable that Democrats reference "workers" quite frequently while Republicans mention "economy" and "tax".

```r
# Democratic party platform word cloud
set.seed(100)

display.brewer.all(n = NULL, type = "all", select = NULL,
                   exact.n = TRUE, colorblindFriendly = FALSE)
```

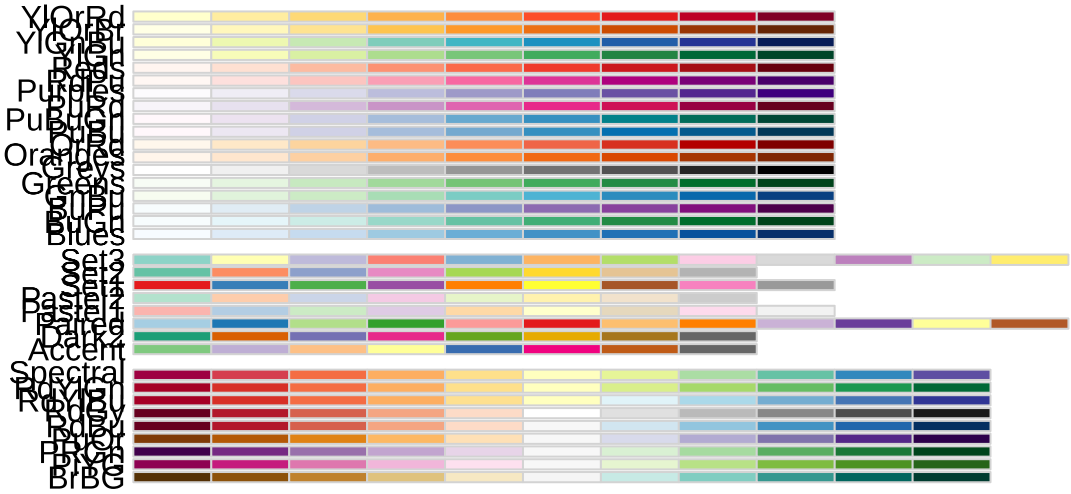


```r
layout(matrix(c(1, 2), nrow=2), heights=c(1, 6))
par(mar=rep(0, 4))
plot.new()
text(x=0.5, y=0.5, "Democratic Party Platform Term Frequency Wordcloud")
wordcloud(names(freq_dem), freq_dem,
          min.freq = 1,
          max.words = 300,
          random.order = FALSE,
          rot.per = 0.30,
          main = "Title",
          colors = brewer.pal(8, "Dark2"))
```

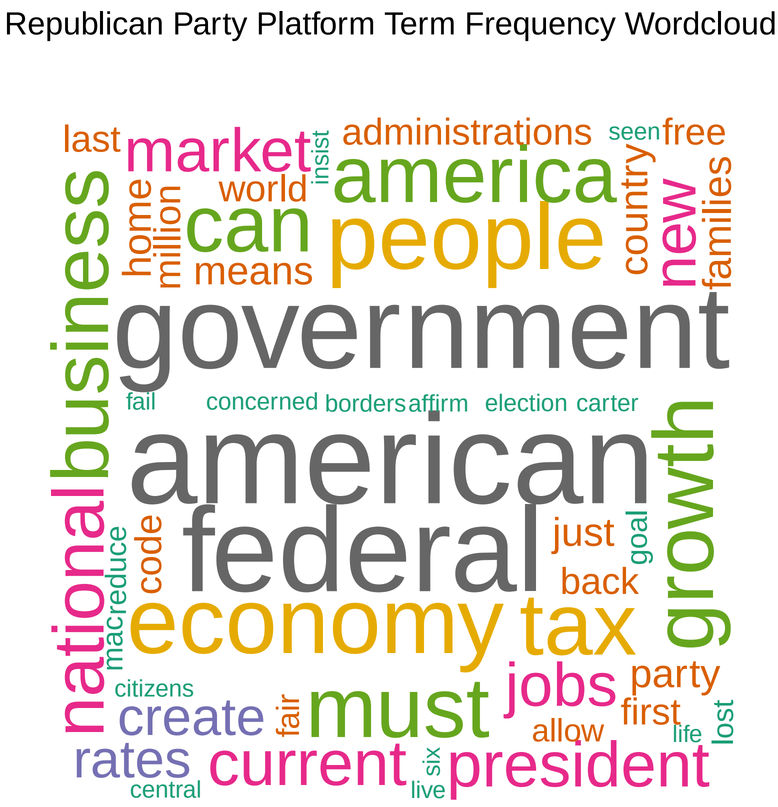
Democratic Party Platform Term Frequency Wordcloud



```
# Republican party word cloud
display.brewer.all(n = NULL, type = "all", select = NULL,
                   exact.n = TRUE, colorblindFriendly = FALSE)
```



```r
layout(matrix(c(1, 2), nrow=2), heights=c(1, 6))
par(mar=rep(0, 4))
plot.new()
text(x=0.5, y=0.5, "Republican Party Platform Term Frequency Wordcloud")
wordcloud(names(freq_rep), freq_rep,
          min.freq = 1,
          max.words = 300,
          random.order = FALSE,
          rot.per = 0.30,
          main = "Title",
          colors = brewer.pal(8, "Dark2"))
```


Republican Party Platform Term Frequency Wordcloud

The Democratic party platform word cloud appears to emphasize jobs and workers, with a focus on "family" and "communities". Housing and jobs are also frequently mentioned, suggesting that the Democratic party platform is focused on individual economic opportunities. On the other hand, the Republican party platform emphasizes words like "American", "tax", "trade", "business", and "national", indicating that Republicans are focused more on macro-scale economic policies and institutions.

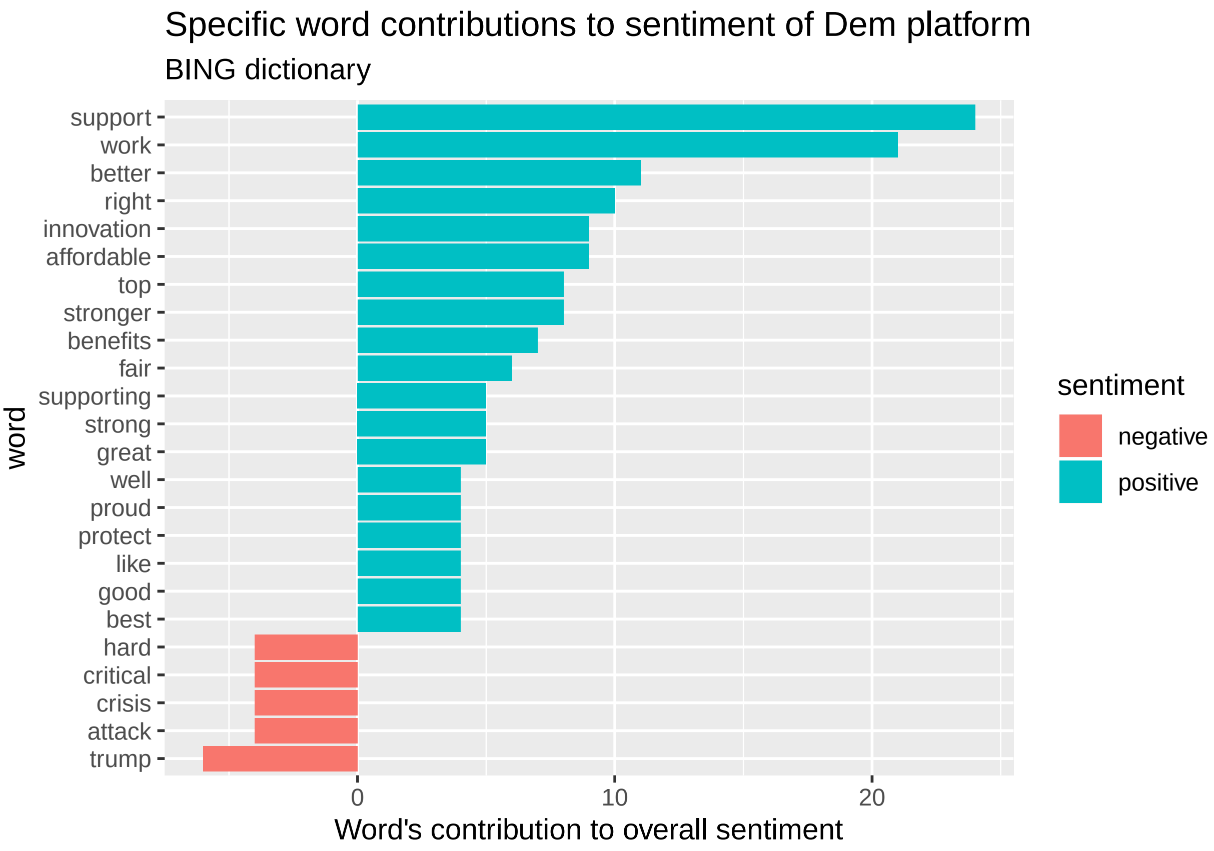
SENTIMENT ANALYSIS

Questions 4 and 5

```
freq_dem_t <- tibble("word" = names(freq_dem), "n" = freq_dem)
freq_rep_t <- tibble("word" = names(freq_rep), "n" = freq_rep)

bing <- get_sentiments("bing")
afinn <- get_sentiments("afinn")

dem_bing <- freq_dem_t %>%
  inner_join(bing, by = "word") %>%
  mutate(sentiment = ifelse(word == "trump", "negative", sentiment),
         # Recode Trump to negative sentiment in Dem party platform
         ntone = ifelse(sentiment == "positive", n, -n)) %>%
  summarize(total_tone = sum(ntone),
            total_words = sum(n))

dem_bing$total_tone / dem_bing$total_words
```

```
## [1] 0.3930348
```

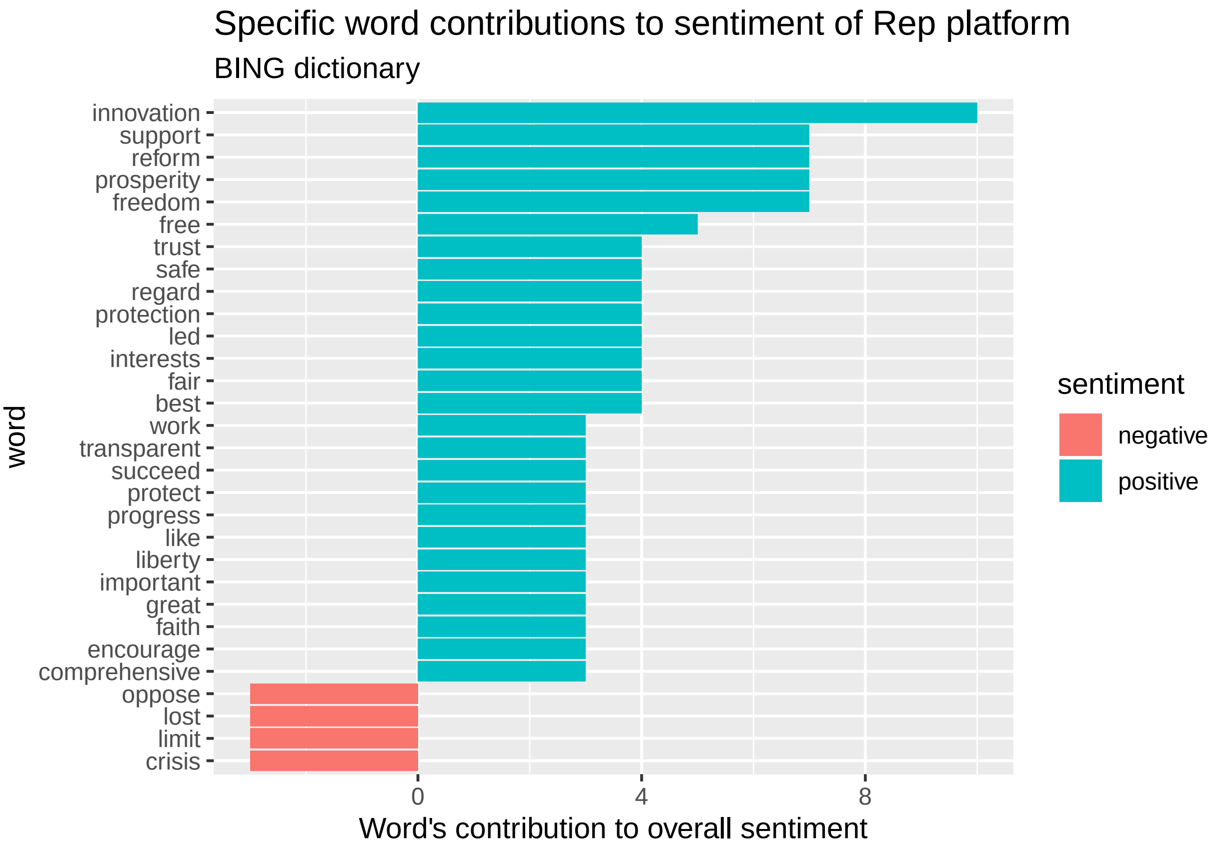
```
# Plot how specific words contribute to the bing sentiment of the platform
# Note that "Trump" here is counted as a positive word, when it probably
# shouldn't be.
freq_dem_t %>%
  inner_join(bing, by = "word") %>%
  filter(n > 3) %>%
  mutate(sentiment = ifelse(word == "trump", "negative", sentiment),
         # Recode Trump to negative in Dem party platform
         n = ifelse(sentiment == "positive", n, -n),
         word = reorder(word, n)) %>%
  ggplot(aes(word, n, fill = sentiment)) +
  geom_col() +
  coord_flip() +
  labs(y = "Word's contribution to overall sentiment",
       title = "Specific word contributions to sentiment of Dem platform",
       subtitle = "BING dictionary")
```



```
rep_bing <- freq_rep_t %>%
  inner_join(bing, by = "word") %>%
  mutate(ntone = ifelse(sentiment == "positive", n, -n)) %>%
  summarize(total_tone = sum(ntone),
            total_words = sum(n))

rep_bing$total_tone / rep_bing$total_words
```

```
## [1] 0.2189974
```

```
freq_rep_t %>%
  inner_join(bing, by = "word") %>%
  filter(n > 2) %>%
  mutate(n = ifelse(sentiment == "positive", n, -n),
         word = reorder(word, n)) %>%
  ggplot(aes(word, n, fill = sentiment)) +
  geom_col() +
  coord_flip() +
  labs(y = "Word's contribution to overall sentiment",
       title = "Specific word contributions to sentiment of Rep platform",
       subtitle = "BING dictionary")
```

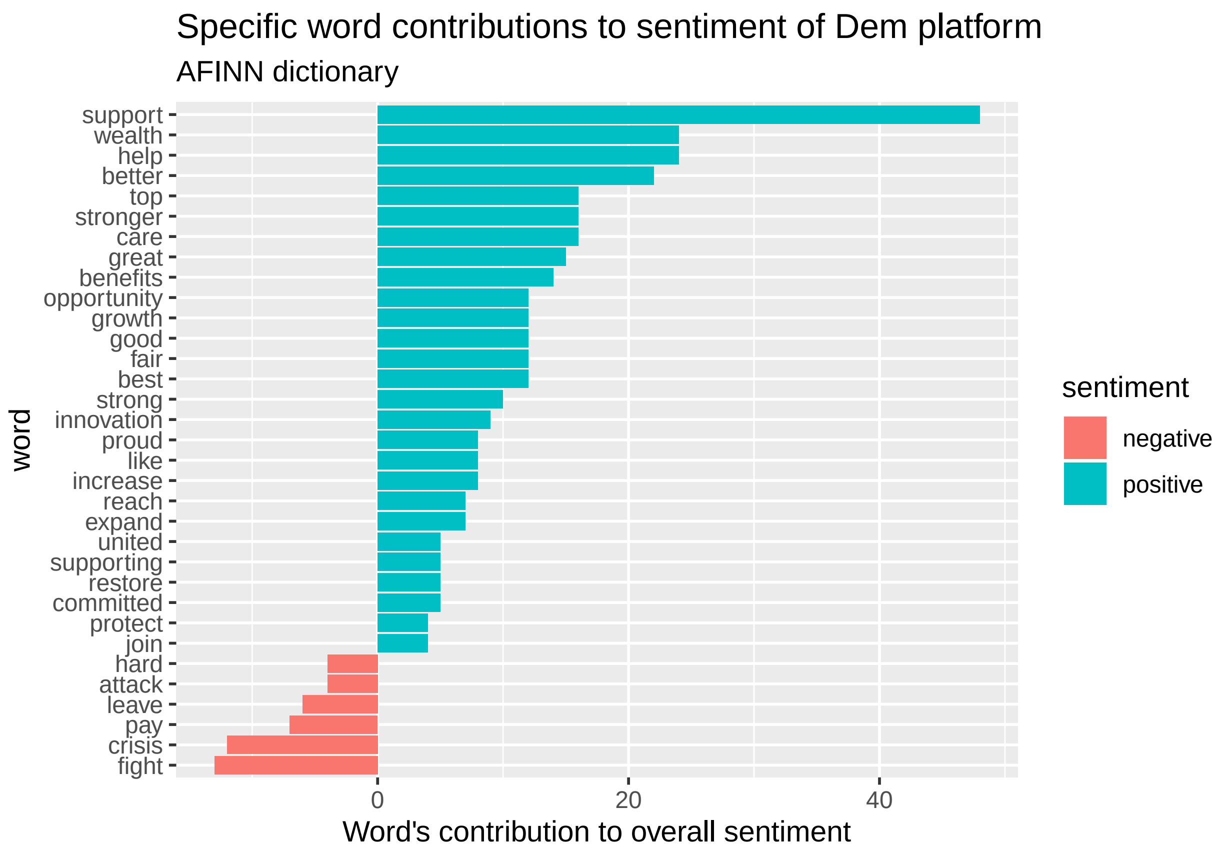


According to the positive/negative encoding done by the Bing dictionary, the Democratic party platform has a tone value that is positive and nearly twice the magnitude of the Republican party's platform. This difference holds true even after we reclassify "trump" as a negative word in the Democratic party platform, because we can reasonably assume Democrats mention Trump to attack him (while the opposite is true for Republicans, so we leave "trump" as a positive word in their party platform). This suggests that the Democrat platform uses more positive words and contains a generally more optimistic outlook about the state of the party, its goals, and the future. Interestingly, we see that "support" and "innovation" are two of the strongest positive words in both platforms, while “crisis” is among the strongest negative words in each. We can check this overall finding using the AFINN dictionary

```
dem_afinn <- freq_dem_t %>%
  inner_join(afinn, by = "word") %>%
  mutate(score = n*value) %>%
  summarize(totscore = sum(score),
            totwords = sum(n))

dem_afinn$totscore / dem_afinn$totwords
```

```
## [1] 0.9073634
```

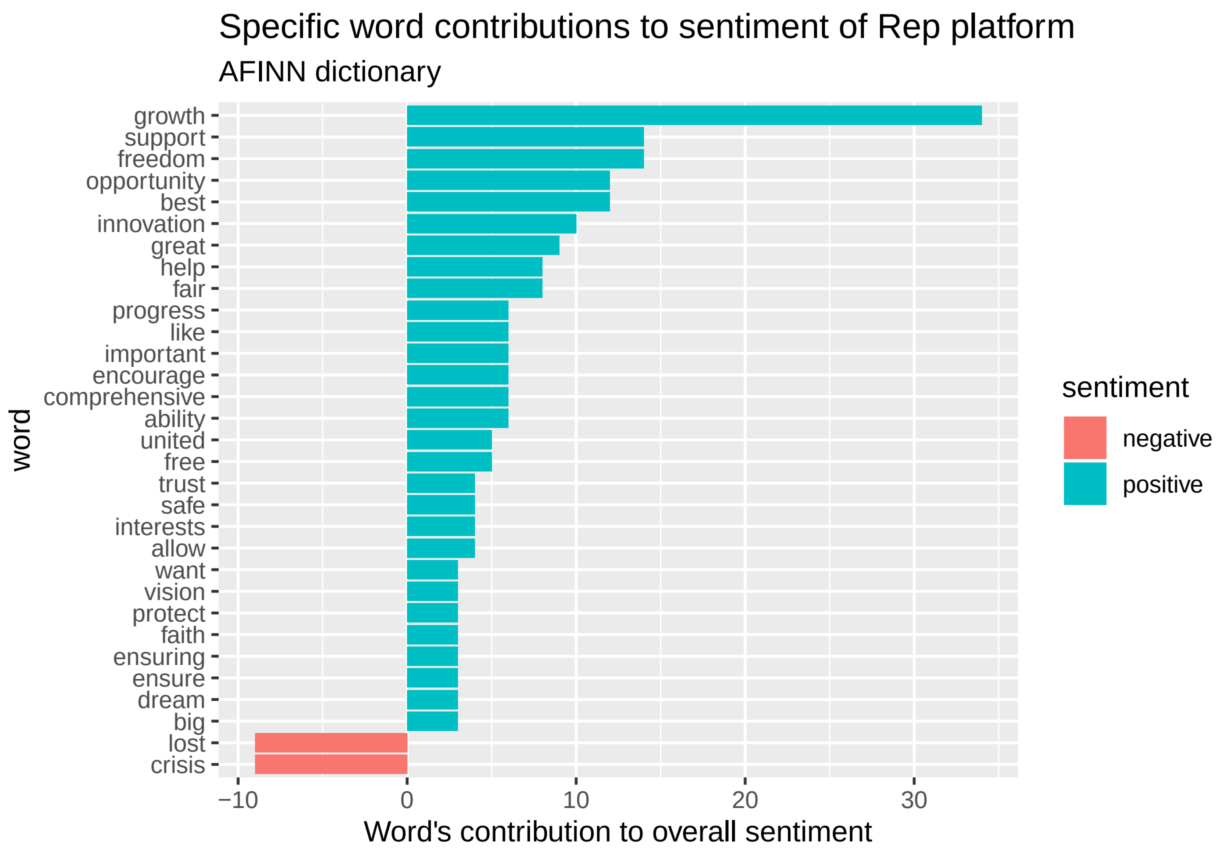
```
# Plot how specific words contribute to the bing sentiment of the platform
# Note that "Trump" here is counted as a positive word, when it probably
# shouldn't be.
freq_dem_t %>%
  inner_join(afinn, by = "word") %>%
  filter(n > 3) %>%
  mutate(value = ifelse(word == "trump", -value, value),
         # Recode Trump to negative in Dem party platform
         score = n*value,
         sentiment = ifelse(score >= 0, "positive", "negative"),
         word = reorder(word, score)) %>%
  ggplot(aes(word, score, fill = sentiment)) +
  geom_col() +
  coord_flip() +
  labs(y = "Word's contribution to overall sentiment",
       title = "Specific word contributions to sentiment of Dem platform",
       subtitle = "AFINN dictionary")
```



```
rep_afinn <- freq_rep_t %>%
  inner_join(afinn, by = "word") %>%
  mutate(score = n*value) %>%
  summarize(totscore = sum(score),
            totwords = sum(n))

rep_afinn$totscore / rep_afinn$totwords
```

```
## [1] 0.6051873
```

```
freq_rep_t %>%
  inner_join(afinn, by = "word") %>%
  filter(n > 2) %>%
  mutate(score = n*value,
         sentiment = ifelse(score >= 0, "positive", "negative"),
         word = reorder(word, score)) %>%
  ggplot(aes(word, score, fill = sentiment)) +
  geom_col() +
  coord_flip() +
  labs(y = "Word's contribution to overall sentiment",
       title = "Specific word contributions to sentiment of Rep platform",
       subtitle = "AFINN dictionary")
```



Using the AFINN dictionary, we again observe that the Democratic party platform has a more positive overall sentiment. This comports with my understanding of the two parties and their positions. Democrats tend to favor progressive policies that actively seek to change the future in positive ways, while Republicans support conservative policies that may play on pessimism or fear about the future to promote adherence to the status quo or even a return to policies from previous eras.

TOPIC MODELS

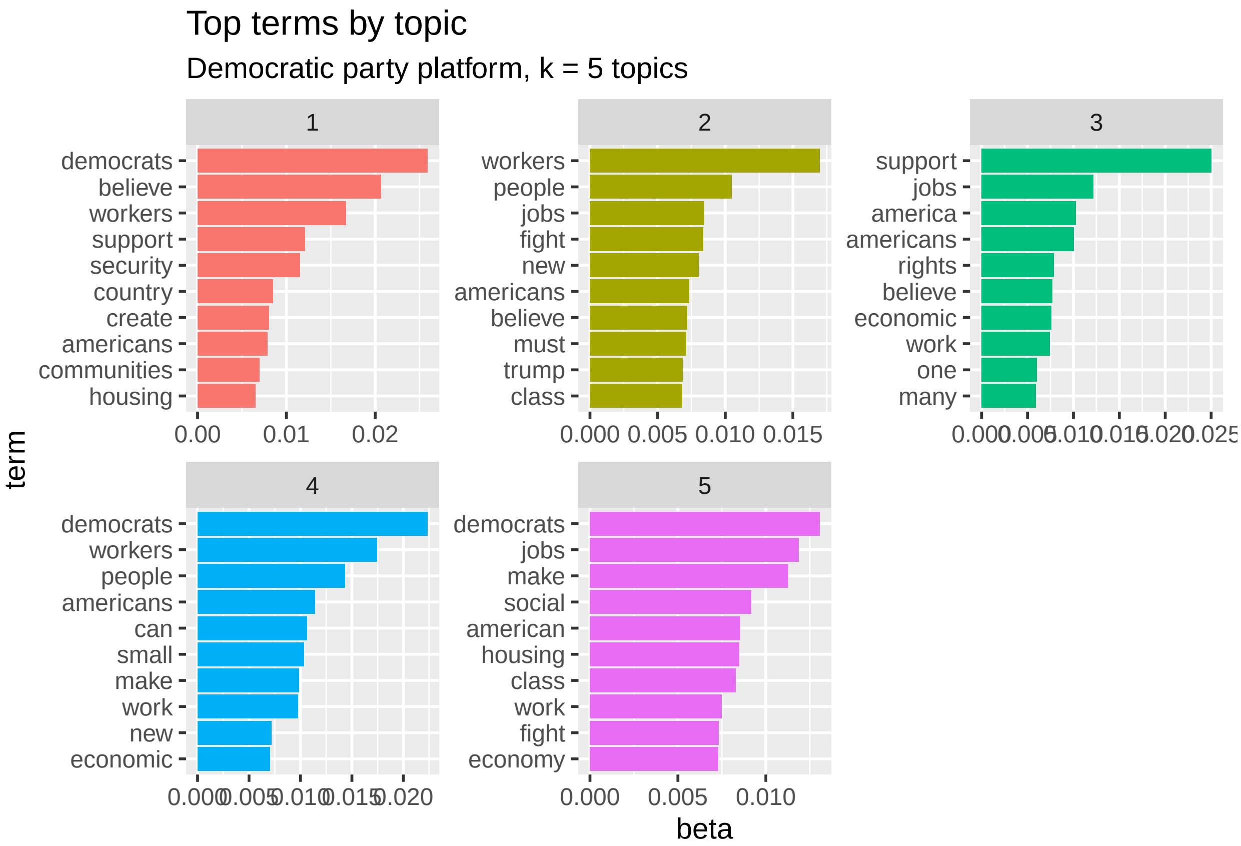
Question 6

Start by fitting topic models using the latent Dirichlet allocation algorithm with k = 5 for each party.

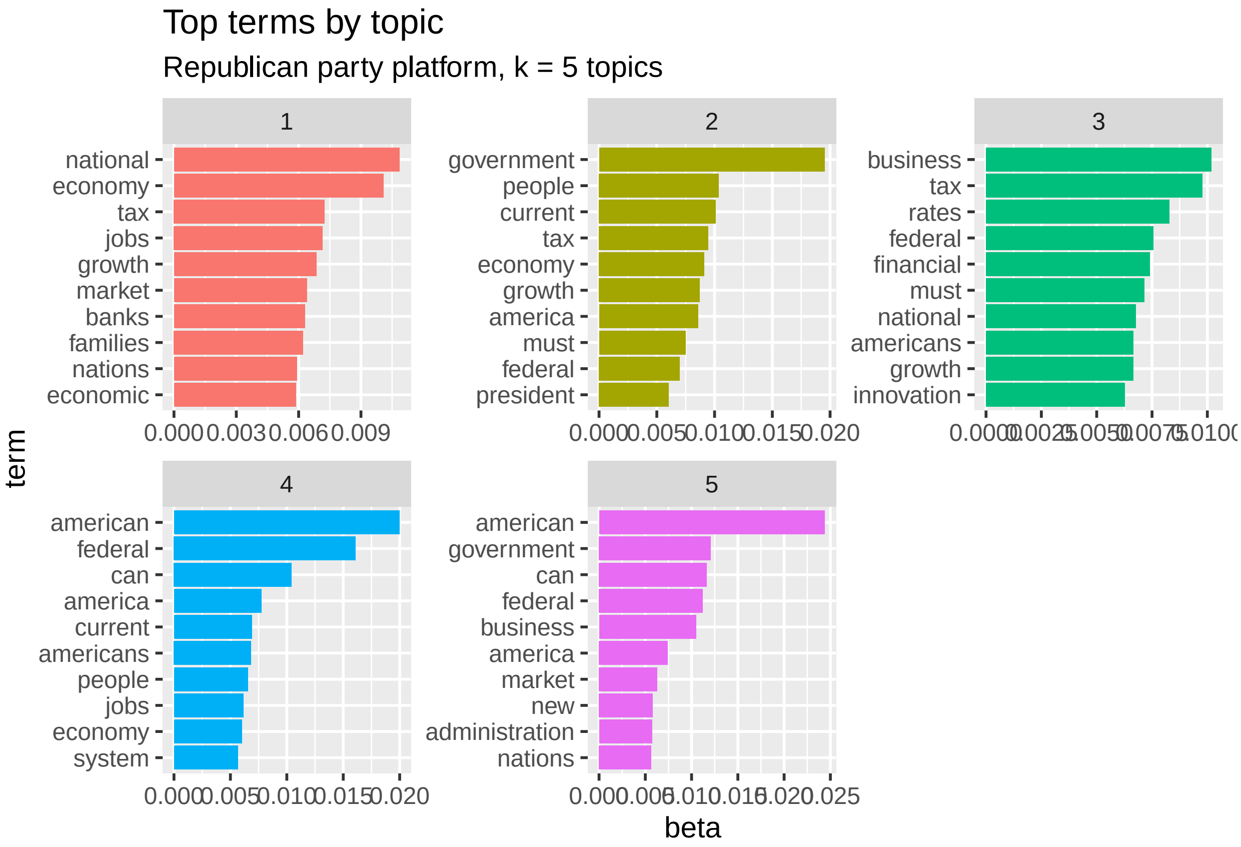
```
dem_t5 <- topicmodels::LDA(dem_dtm, k = 5, control = list(seed = 101))
rep_t5 <- topicmodels::LDA(rep_dtm, k = 5, control = list(seed = 101))

dem_topics <- tidy(dem_t5, matrix = "beta")
rep_topics <- tidy(rep_t5, matrix = "beta")
```

```
# Code adapted from source at: https://www.tidytextmining.com/topicmodeling.html
dem_top_terms <- dem_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, desc(beta))
```

```
dem_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") +
  coord_flip() +
  scale_x_reordered() +
  labs(title = "Top terms by topic",
       subtitle = "Democratic party platform, k = 5 topics")
```



```
rep_top_terms <- rep_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, desc(beta))

rep_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") +
  coord_flip() +
  scale_x_reordered() +
  labs(title = "Top terms by topic",
       subtitle = "Republican party platform, k = 5 topics")
```



Question 7

The 5 topics produced from the Democratic platform have a lot of overlap: some form of “work” or “workers” appears in all 5, and “jobs” and “economy” are very common across the 5 topics as well. As observed before, the Democratic party platform appears to put emphasis on individual economic rights and opportunities, especially with references to “class”, “housing”, and “fight.” These topic divisions and the terms that define them appear to reinforce the image of the Democratic party as the party of working people.

On the other hand, the Republican platform gives rise to topics that focus on the macroeconomy and systemic/technocratic policy, with numerous references to “business”, “tax”, “national”, “federal”, “administration”, “growth”, and “market.” These topics suggest that, while the economy is the most important policy area for Republicans as well as Democrats, Republicans take a much more corporatist/institutional approach to economic policy, preferring to focus on big picture national policies.

Question 8

Start by fitting the 6 new additional topic models.

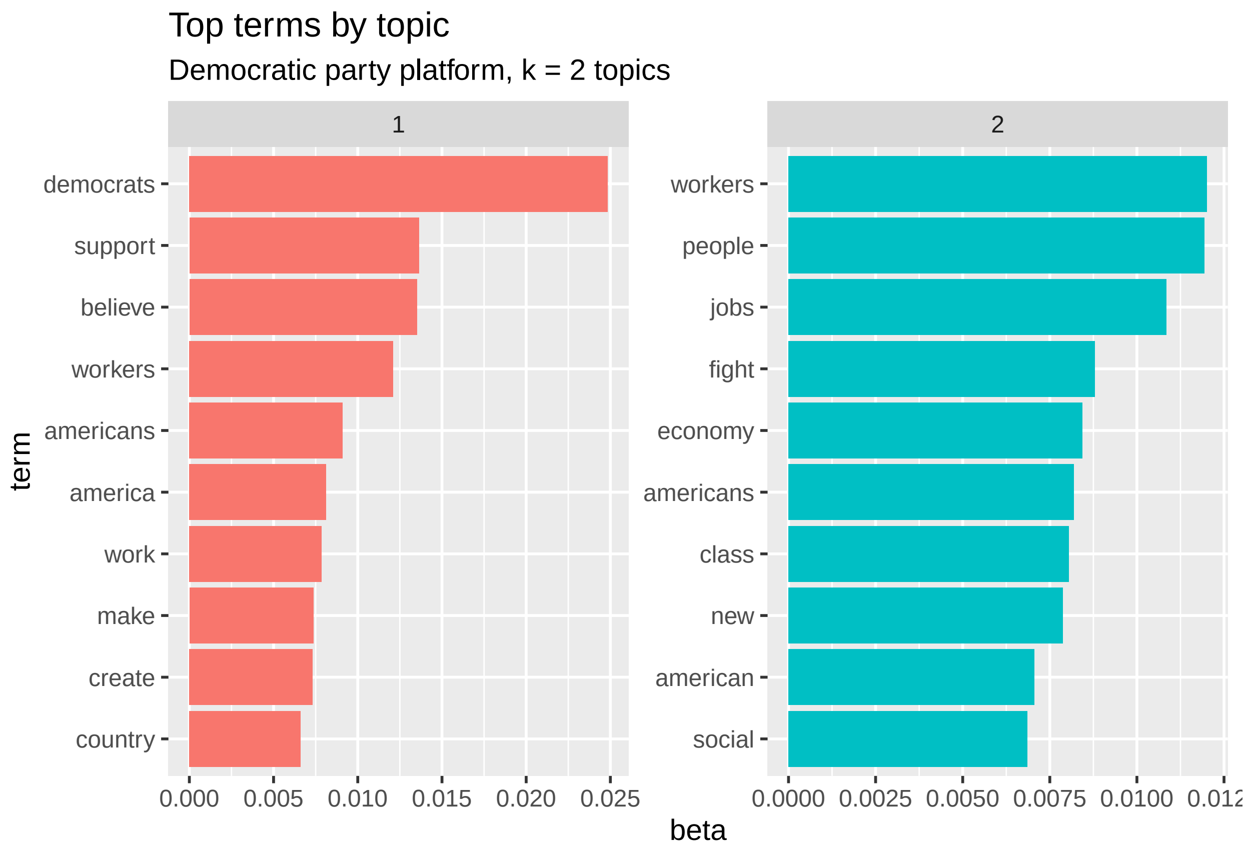
```
# I'm assuming that, since we already fit k = 5, the instruction to fit k = 5
# again was a typo. So I'll fit k = 2 in its place.
dem_t2 <- topicmodels::LDA(dem_dtm, k = 2, control = list(seed = 101))
rep_t2 <- topicmodels::LDA(rep_dtm, k = 2, control = list(seed = 101))

dem_t10 <- topicmodels::LDA(dem_dtm, k = 10, control = list(seed = 101))
rep_t10 <- topicmodels::LDA(rep_dtm, k = 10, control = list(seed = 101))
```

```
dem_t25 <- topicmodels::LDA(dem_dtm, k = 25, control = list(seed = 101))
rep_t25 <- topicmodels::LDA(rep_dtm, k = 25, control = list(seed = 101))
```

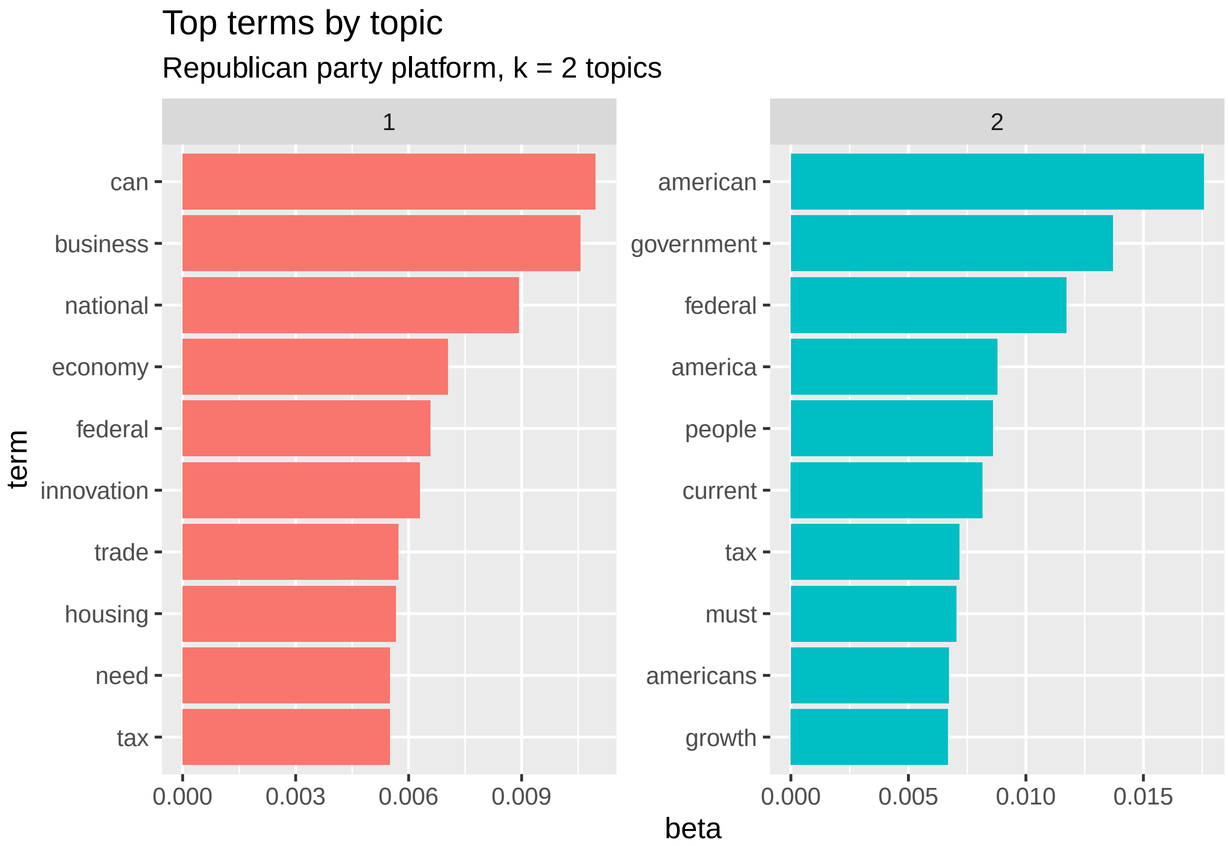
Create a function to make visual plotting of the topics more convenient.

```
topics_visual <- function(model, party, k) {
  topics <- tidy(model, matrix = "beta")
  top_terms <- topics %>%
    group_by(topic) %>%
    top_n(10, beta) %>%
    ungroup() %>%
    arrange(topic, desc(beta))

  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") +
    coord_flip() +
    scale_x_reordered() +
    labs(title = "Top terms by topic",
         subtitle = paste0(party, " party platform, k = ", k, " topics"))
}
```

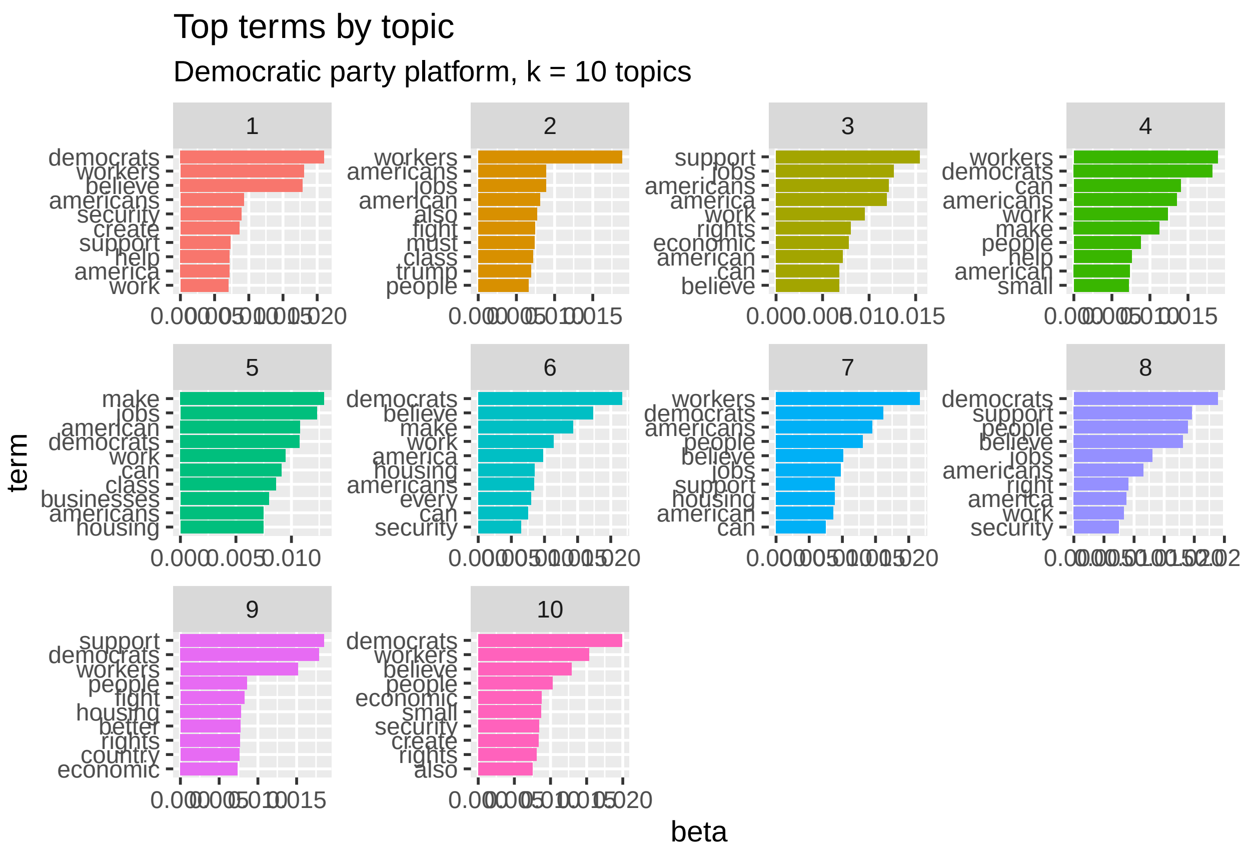
```
topics_visual(dem_t2, "Democratic", "2")
```



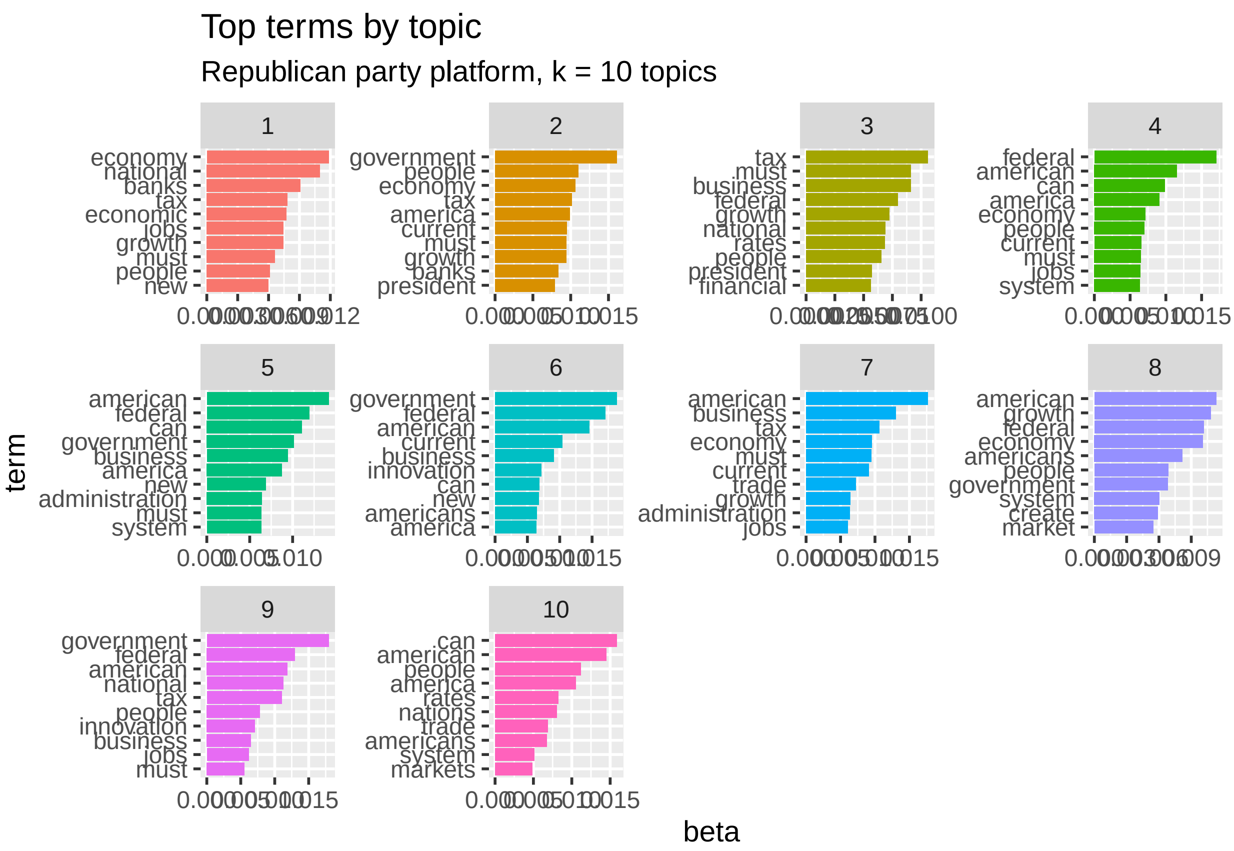
```r
topics_visual(rep_t2, "Republican", "2")
```



```r
topics_visual(dem_t10, "Democratic", "10")
```



```r
topics_visual(rep_t10, "Republican", "10")
```



The two-topic model from the Democratic platform reveals two well-bifurcated topics: on one hand, there is a more "aspirational" topic discussed in broad terms that frequently references America and its citizens, along with generally positive but less specific verbs like "make", "create", "support", and "believe." On the other hand, there is a much more specific topic that calls out economic policy buzzwords that define the Democrats' ideals, including "social", "class", "jobs", and "fight". The same is true on the Republican side - the first topic has much more specific terms that have to do with Republicans' views on the economy, while the second topic is much more anodyne and vague. The 10-topic models follow roughly the same path, although there is much duplication of terms across topics which makes distinguishing between topics a little more opaque.

Visual representation of 25 topics is difficult, so instead we'll find the top 3 terms from each of the 25 selected topics and then count the number of times (out of 25) a given term is in the top 3 terms for a given topic.

```
tidy(dem_t25, matrix = "beta") %>%
  group_by(topic) %>%
  top_n(3, beta) %>%
  group_by(term) %>%
  summarize(n = n()) %>%
  arrange(desc(n))
```

```
## # A tibble: 15 x 2
##    term           n
##    <chr>      <int>
##  1 democrats     16
##  2 workers       14
##  3 believe        9
##  4 americans     8
##  5 support       7
##  6 people        4
##  7 america       3
##  8 can           3
##  9 work          3
## 10 jobs          2
## 11 make          2
## 12 also          1
## 13 american      1
## 14 housing       1
## 15 must          1
```

```
tidy(rep_t25, matrix = "beta") %>%
  group_by(topic) %>%
  top_n(3, beta) %>%
  group_by(term) %>%
  summarize(n = n()) %>%
  arrange(desc(n))
```

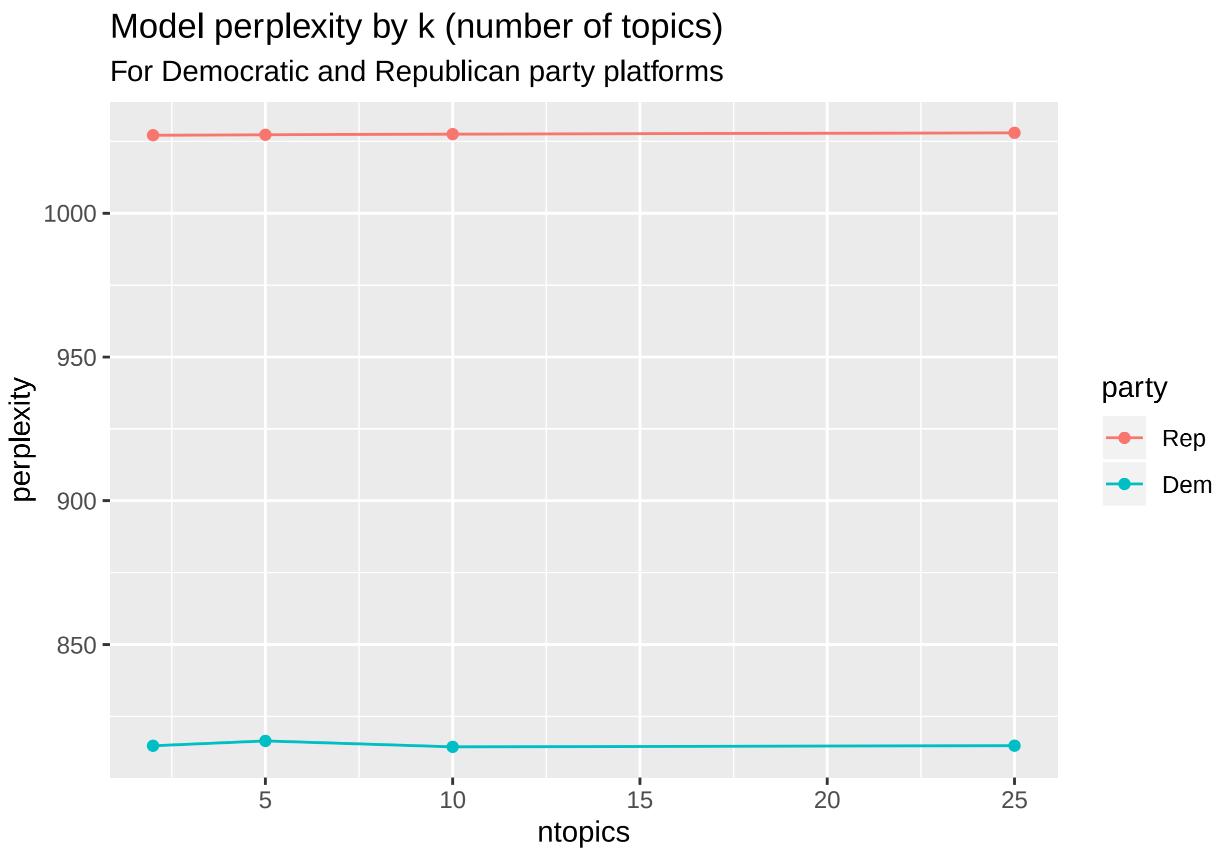
```
## # A tibble: 13 x 2
##    term           n
##    <chr>      <int>
##  1 federal       14
##  2 american      12
##  3 government    12
##  4 economy        6
##  5 growth         6
##  6 business       5
##  7 america        4
##  8 must           4
##  9 people         4
## 10 tax            3
## 11 can            2
## 12 national       2
## 13 innovation     1
```

The most interesting finding to emerge from the 25-topic models are the specific words that appear in the highest proportion of the topics. It is intriguing that Democrats' party platform explicitly includes the word "Democrats", while Republicans do not reference their own party name to nearly the same degree. Democrats reference "workers" far more than most other terms, and "believe" and "support" are also quite prominent, bolstering the inference that the Democrats' party platform was designed explicitly with working people in mind and contains a more aspirational tone. On the other hand, Republicans mention "federal", "american", and "government" most often, perhaps in service of making the case for smaller, less intrusive federal government.

Question 9

Now we'll calculate the perplexity of the models we've trained and use that measure to determine which model provides the best fit to the data.

```r
ntopics <- c(2, 2, 5, 5, 10, 10, 25, 25)
party <- rep(c("Dem", "Rep"), 4)
perplexities <- map(c(dem_t2, rep_t2, dem_t5, rep_t5, dem_t10, rep_t10, dem_t25, rep_t25), perplexity)

perplex_df <- tibble(ntopics = ntopics, party = party, perplexity = unlist(perplexities)) %>%
  mutate(ntopics = as.numeric(ntopics),
         party = factor(party, c("Rep", "Dem")))

ggplot(perplex_df, aes(ntopics, perplexity, color = party)) +
  geom_point() +
  geom_line() +
  labs(title = "Model perplexity by k (number of topics)",
       subtitle = "For Democratic and Republican party platforms")
```



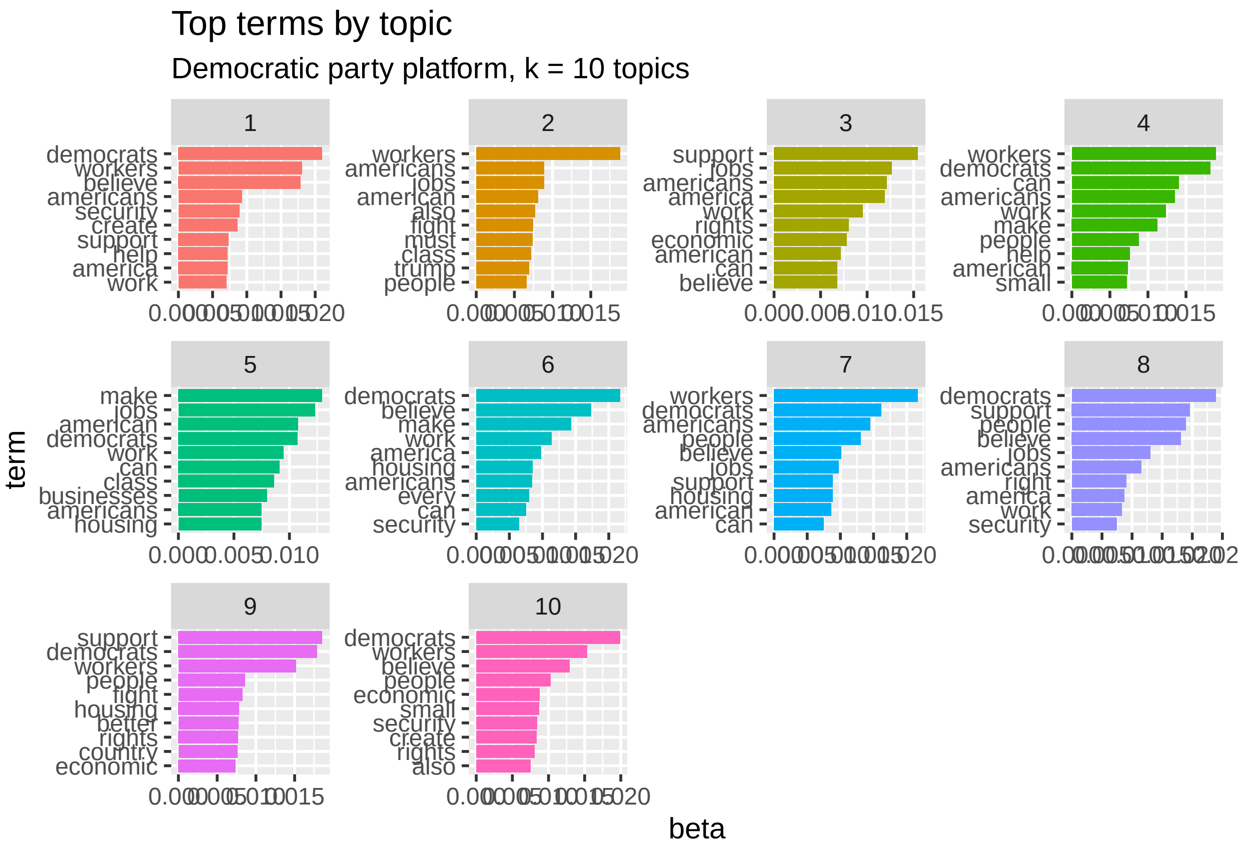
```r
perplex_df
```

```
## # A tibble: 8 x 3
##   ntopics party perplexity
##     <dbl> <fct>      <dbl>
## 1       2 Dem         815.
## 2       2 Rep        1027.
## 3       5 Dem         816.
## 4       5 Rep        1027.
## 5      10 Dem         814.
## 6      10 Rep        1028.
## 7       25 Dem          815.
## 8       25 Rep         1028.
```

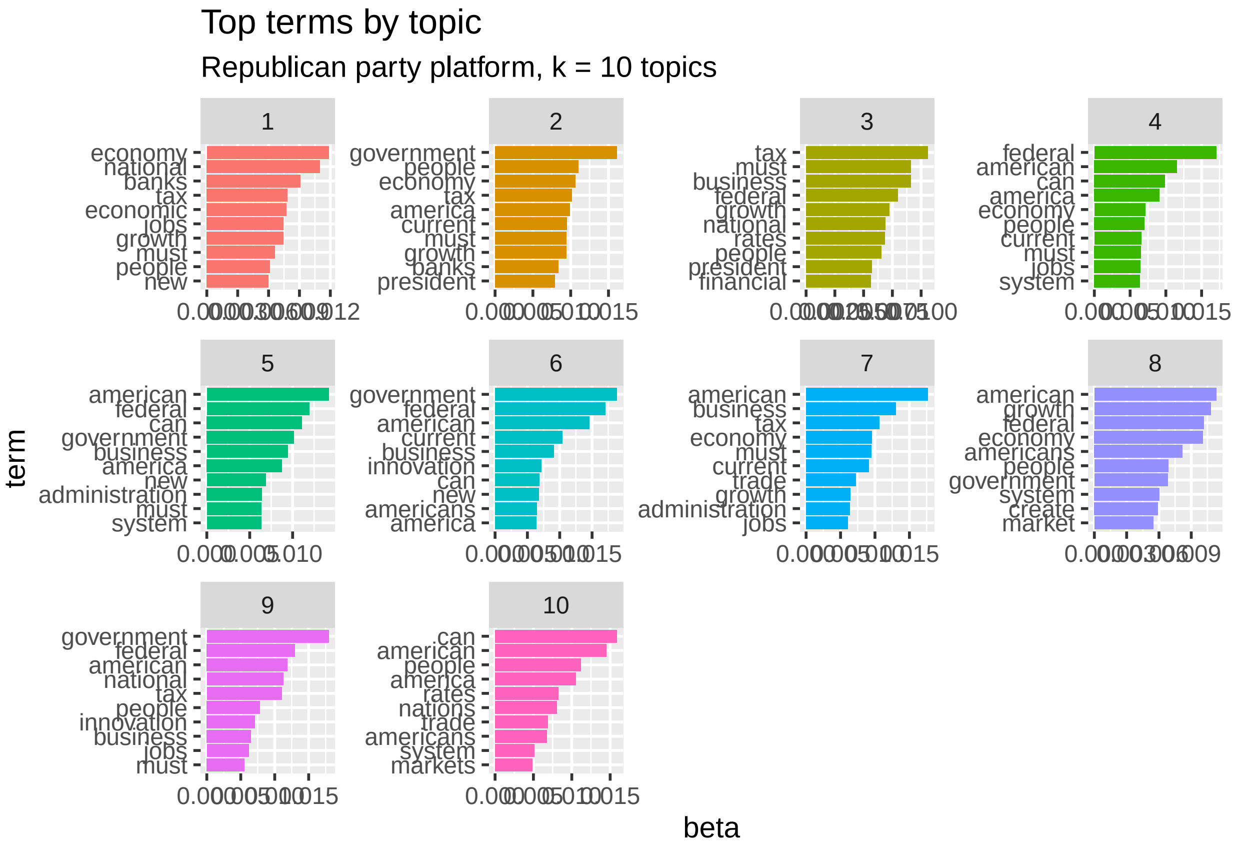
The perplexity scores are remarkably similar across all k-topic models within each party platform. Technically speaking, k = 10 is the best model fit for the Democratic platform, while k = 2 is the best model fit for the Republican platform, but these differences are extremely granular.

Question 10

```
topics_visual(dem_t10, "Democratic", "10")
```



```
topics_visual(rep_t10, "Republican", "10")
```



I don't necessarily believe that the k = 10 models pick up differences between the party platforms any more efficiently than the k = 2 or k = 5 models, because there is quite a bit of repetition across topics. There seems to be less diverse, "new" information because the same amount of information from the fewer-topic models is now spread out and in some cases duplicated in the k = 10 case.

CONCLUSION

Question 11

The Democrats' party platform appears to focus on individual economic opportunity, workers' rights, and explicit class differences. The Republicans' party platform seems to espouse a more institutional view of economic policy, focusing on tax, trade, and the government's role in the economy. My experience from family history and posterior beliefs from studying economics tell me that trickle-down economics don't provide the benefits their supporters claim they do. Income and wealth inequality are serious problems in America and the anecdotal and scholarly evidence I've consumed suggest that the social safety net, minimum wage, and reasonable government regulation (in jobs, housing, etc.) are necessary to provide a baseline level of stability and security for American families. Furthermore, I think progressive policies have the potential to create a brighter, more optimistic future, which is also reflected in the tone of the Democrats' party platform. The Democratic paradigm for the economy that has emerged from this textual analysis aligns more closely with my own, so I would support the Democrats in 2020.