Assignment5_AbhishekPandit

Abhishek Pandit

26 November 2019

R Markdown

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.

```
library(tm)
```

```
## Warning: package 'tm' was built under R version 3.5.3
```

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

```
## Warning: package 'NLP' was built under R version 3.5.2
```

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

```
## Warning: package 'wordcloud' was built under R version 3.5.3
```

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

```
## Warning: package 'RColorBrewer' was built under R version 3.5.2
```

```
library(wordcloud2)
```

```
## Warning: package 'wordcloud2' was built under R version 3.5.3
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.3
```

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

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
## Warning: package 'tibble' was built under R version 3.5.3
```

```
## Warning: package 'tidyr' was built under R version 3.5.2
```

```
## Warning: package 'readr' was built under R version 3.5.2
```

```
## Warning: package 'purrr' was built under R version 3.5.2
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
## Warning: package 'stringr' was built under R version 3.5.2
```

```
## Warning: package 'forcats' was built under R version 3.5.2
```

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

```
library(topicmodels)
```

```
## Warning: package 'topicmodels' was built under R version 3.5.3
```

```
library(textdata)
```

```
## Warning: package 'textdata' was built under R version 3.5.3
```

```
library(tidytext)
```

```
## Warning: package 'tidytext' was built under R version 3.5.3
```

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.
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

First, we load the CSV file.

```
platforms <- read.csv("D:/Dropbox/Chicago/Courses/UML/Problem-Set-5/Party_ Platforms_Data/platforms.csv
#Fortunately, the tm library contains most of the tools needed for Q2
docs <- VCorpus(DirSource("D:/Dropbox/Chicago/Courses/UML/Problem-Set-5/Party_ Platforms_Data")) %>%
tm_map(tolower) %>%
tm_map(removeWords, stopwords('english')) %>%
tm_map(removeNumbers) %>%
tm_map(removePunctuation) %>%
tm_map(stripWhitespace) %>%
tm_map(removeWords, c("will", "and", "the"))%>%
tm_map(PlainTextDocument)
#Now we create document-term matrices
democ_dtm <- DocumentTermMatrix(docs[1]) #This is the d16.txt file
repub_dtm <- DocumentTermMatrix(docs[3]) #This is the r16.txt file
```

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?

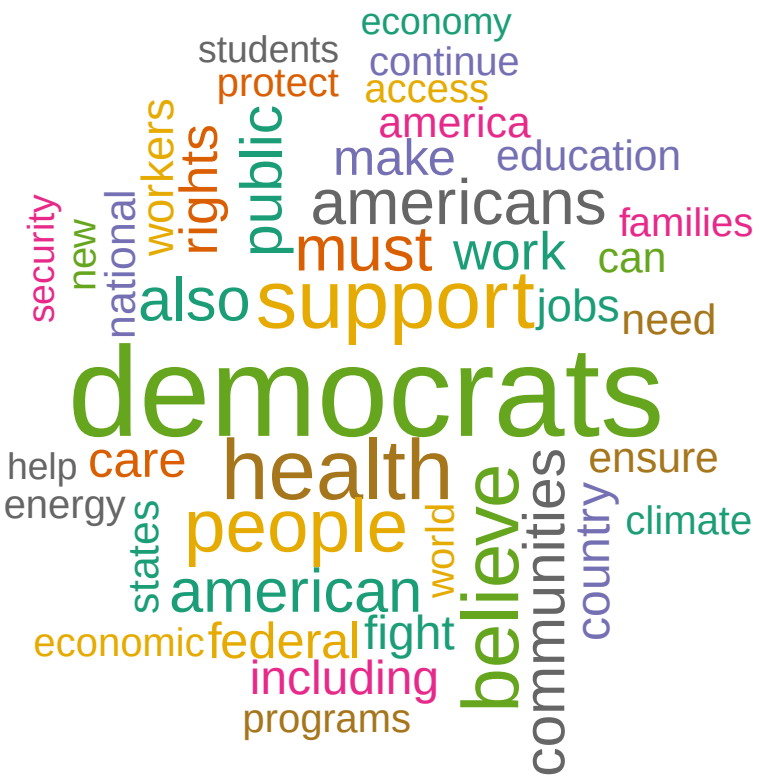
First, we reshape the Document-Term Matrices to make this possible.

```
democ_frequency <- sort(colSums(as.matrix(democ_dtm)),
                   decreasing=TRUE)
repub_frequency <- sort(colSums(as.matrix(repub_dtm)),
                   decreasing=TRUE)
```

Now the actual word clouds

```
set.seed(2019)
wordcloud(names(democ_frequency),
          democ_frequency,
          min.freq = 40,
          max.words = 200,
```

```
          colors = brewer.pal(8, "Dark2"),
          random.order = FALSE,
          rot.per = 0.30,
          random.color = TRUE)
```



Then for the Republicans

```
wordcloud(names(repub_frequency), repub_frequency, min.freq = 40,
          max.words = 250,
          colors = brewer.pal(8, "Dark2"),
          random.order = FALSE,
          rot.per = 0.30,
          random.color = TRUE)
```

```
## Warning in wordcloud(names(repub_frequency), repub_frequency, min.freq =
## 40, : president could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(repub_frequency), repub_frequency, min.freq =
## 40, : americans could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(repub_frequency), repub_frequency, min.freq =
## 40, : economic could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(repub_frequency), repub_frequency, min.freq =
## 40, : security could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(repub_frequency), repub_frequency, min.freq =
## 40, : economy could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(repub_frequency), repub_frequency, min.freq =
```

```
## 40, : military could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(repub_frequency), repub_frequency, min.freq =
## 40, : country could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(repub_frequency), repub_frequency, min.freq =
## 40, : religious could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(repub_frequency), repub_frequency, min.freq =
## 40, : americas could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(repub_frequency), repub_frequency, min.freq =
## 40, : energy could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(repub_frequency), repub_frequency, min.freq =
## 40, : oppose could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(repub_frequency), repub_frequency, min.freq =
## 40, : amendment could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(repub_frequency), repub_frequency, min.freq =
## 40, : america could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(repub_frequency), repub_frequency, min.freq =
## 40, : every could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(repub_frequency), repub_frequency, min.freq =
## 40, : freedom could not be fit on page. It will not be plotted.

## Warning in wordcloud(names(repub_frequency), repub_frequency, min.freq =
## 40, : families could not be fit on page. It will not be plotted.
```

Many of the most frequent words could not be plotted on this graph for the republicans. The frequent words are frequent in very high numbers, and therefore can't be fit in the cloud as per their inferred importance

SUMMARY: Based on the word clouds alone, it seems like the Democrats emphasize their party affiliation more conspicuously than for the Republicans. For the latter, the top priorities appear to be around 'government'. 'states' and 'federal', with the prominence of terms like "President" that surprisingly do not even appear in the list for Democrats. They therefore seem to focus more on the political system itself, as well as the people involved in the system. Conversely, the democrats speak of a number of issues that could be addressed via the system- such as health, energy, and communities. The Republicans also repeated their frequent words far more frequently. This lends an air of cogency and focus to their aspirations. Conversely, the Democrats seem somewhat more dispersed in their interests for good governance

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).*

For this section, we will leverage Tibbles to chain some of the required methods on both are datasets.

```
democ_tbl <- tibble(txt=as.character(docs[1]))
repub_tbl <- tibble(txt=as.character(docs[2]))
```

Now for both dictionaries on the Democrats data, followed by the same on Republicans The Afinn dictionary doesn't provide automatic methods to tabulate frequencies. But since positive and negatve scores lie on either side of 0, it can be tabulated into a matrix.

```
democ_bing <- democ_tbl %>%
unnest_tokens(word, txt) %>%
inner_join(get_sentiments("bing")) %>%
count(sentiment, sort = TRUE)
```

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

```
#Afinn
democ_afinn <- democ_tbl %>%
unnest_tokens(word, txt) %>%
inner_join(get_sentiments("afinn"))
```

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

```
matx_democ_afinn <- matrix(c(sum(democ_afinn$value >0),
                             sum(democ_afinn$value <0)),
                           ncol=2,
                           byrow=TRUE)
```

The same for Republicans

```
#Republicans
repub_bing <- repub_tbl %>%
unnest_tokens(word, txt) %>%
inner_join(get_sentiments("bing")) %>%
count(sentiment, sort = TRUE)
```

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

```
#Afinn
repub_affin <- repub_tbl %>%
unnest_tokens(word, txt) %>%
inner_join(get_sentiments("afinn"))
```

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

```
matx_repub_afinn <- matrix(c(sum(repub_affin$value >0),
                             sum(repub_affin$value <0)),
                           ncol=2,
                           byrow=FALSE)
matx_democ_afinn <- matrix(c(sum(democ_afinn$value >0),
                             sum(democ_afinn$value <0)),
                           ncol=2,
                           byrow=FALSE)
```

Now we inspect the matrices.

First, using the Bing dictionary for Democrats

```
democ_bing
```

```
## # A tibble: 2 x 2
##   sentiment     n
##   <chr>     <int>
## 1 positive   1367
## 2 negative    804
```

Now for Republicans

```
repub_bing
```

```
## # A tibble: 2 x 2
##   sentiment     n
##   <chr>     <int>
## 1 positive    520
## 2 negative    266
```

Now we check for and compare the ratio of positive to negative sentiments for the Bing Scores

```
democ_ratio_bing = democ_bing$n[1]/democ_bing$n[2]
repub_ratio_bing = repub_bing$n[1]/repub_bing$n[2]
print(democ_ratio_bing)
```

```
## [1] 1.700249
```

```
print(repub_ratio_bing)
```

```
## [1] 1.954887
```

Similarly, for the Affin dictionary

```
democ_ratio_afinn = matx_democ_afinn[1]/matx_democ_afinn[2]
repub_ratio_afinn = matx_repub_afinn[1]/matx_repub_afinn[2]
print(democ_ratio_afinn)
```

```
## [1] 2.145975
```

```
print(repub_ratio_afinn)
```

```
## [1] 2.685714
```

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?

Based on these results, it seems that the ratio of positive to negative sentiments is higher (under both the Afinn and Bing dictionaries) for the Democratic Party. This fits with what I have observed in media and

research about conservative preference for Republicans (trying to protect old values). This seems to stand in contrast to the seeming openness to policy innovations from the Democrats. For example, they ushered in firsts in the form of an African American as well as female candidate to run for the post of President.

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).*

From the requirements of this section, it's clear we'll be repeating the same topic modeling exercise. To save us the effort, we write a function specifically for this purpose.

```
topic_model_plot <-function(num_topics,data,lda_seed){
  lda_model<-LDA(data,
                 k=num_topics,
                 method = 'Gibbs',
                 control = list(seed=lda_seed))
  topics_k<-tidy(lda_model,
                  matrix = "beta")
  top_terms <- topics_k%>%
    group_by(topic) %>%
    top_n(10, beta) %>%
    ungroup() %>%
    arrange(topic, -beta)%>%
    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()
  return(top_terms)
}

get_perplex<-function(num_topics,data,lda_seed){
  lda_model<-LDA(data,
                 k=num_topics,
                 control = list(seed=lda_seed))
  return(perplexity(lda_model))
}
```

Looping for the Models first

```
required_k <-c(5,10,25)
for(i in required_k) {
 nam <- paste("democ_dtm", i, sep = "")
 assign(nam, topic_model_plot(i, democ_dtm, 2019))
}

for(i in required_k) {
 nam <- paste("repub_dtm", i, sep = "")
```

```
 assign(nam, topic_model_plot(i, repub_dtm, 2019))
}
```

Then looping for perplexities

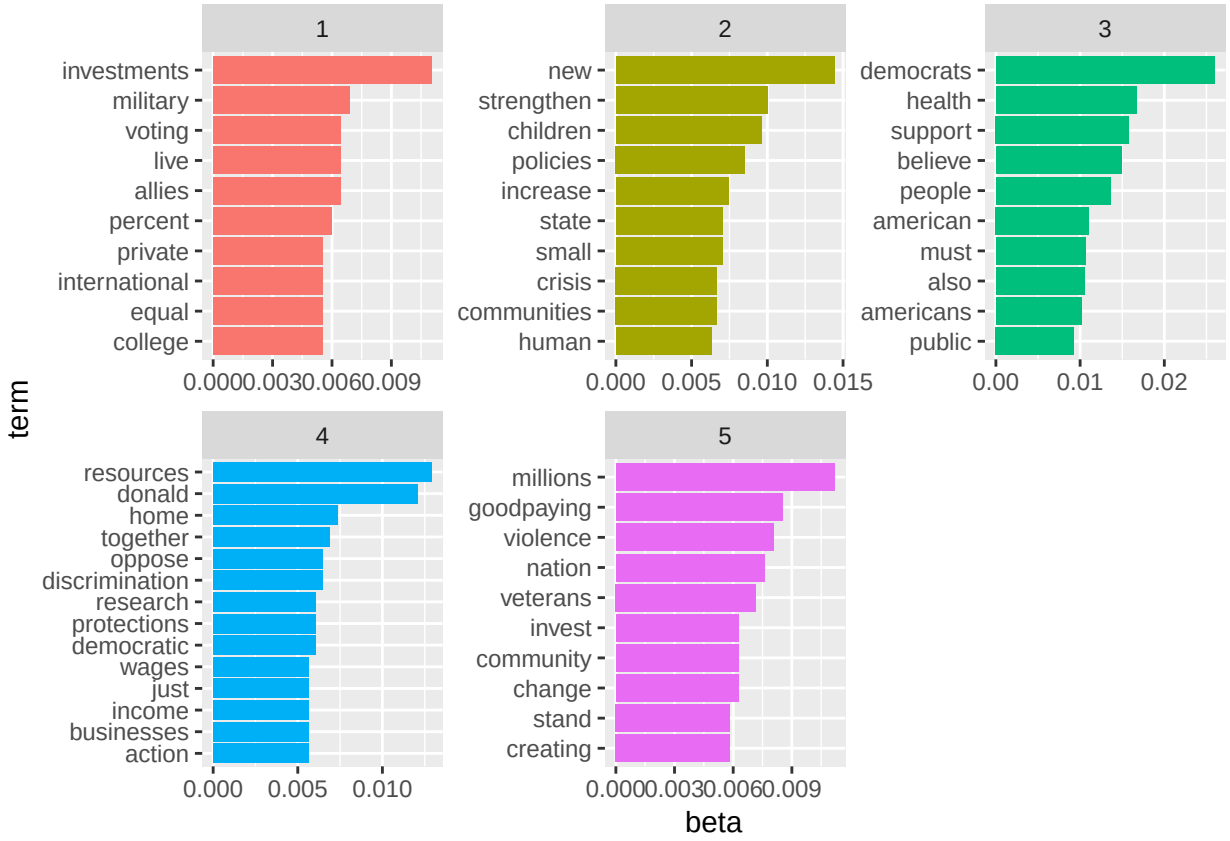
```
for(i in required_k){
  nam <- paste("perplex_democ", i, sep = "")
  assign(nam, get_perplex(i, democ_dtm, 2019))

}

for(i in required_k){
  nam <- paste("perplex_repub", i, sep = "")
  assign(nam, get_perplex(i, repub_dtm, 2019))
}
```

First for the Democrats:

```
democ_dtm5
```



Next, for the Republicans:

```
repub_dtm5
```

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

Of course, interpretation of topics almost by definition proves highly subjective. Nonetheless, some trends emerge.

Overall, we see that they seem to speak about different topics, with about one topic referring to the same content.

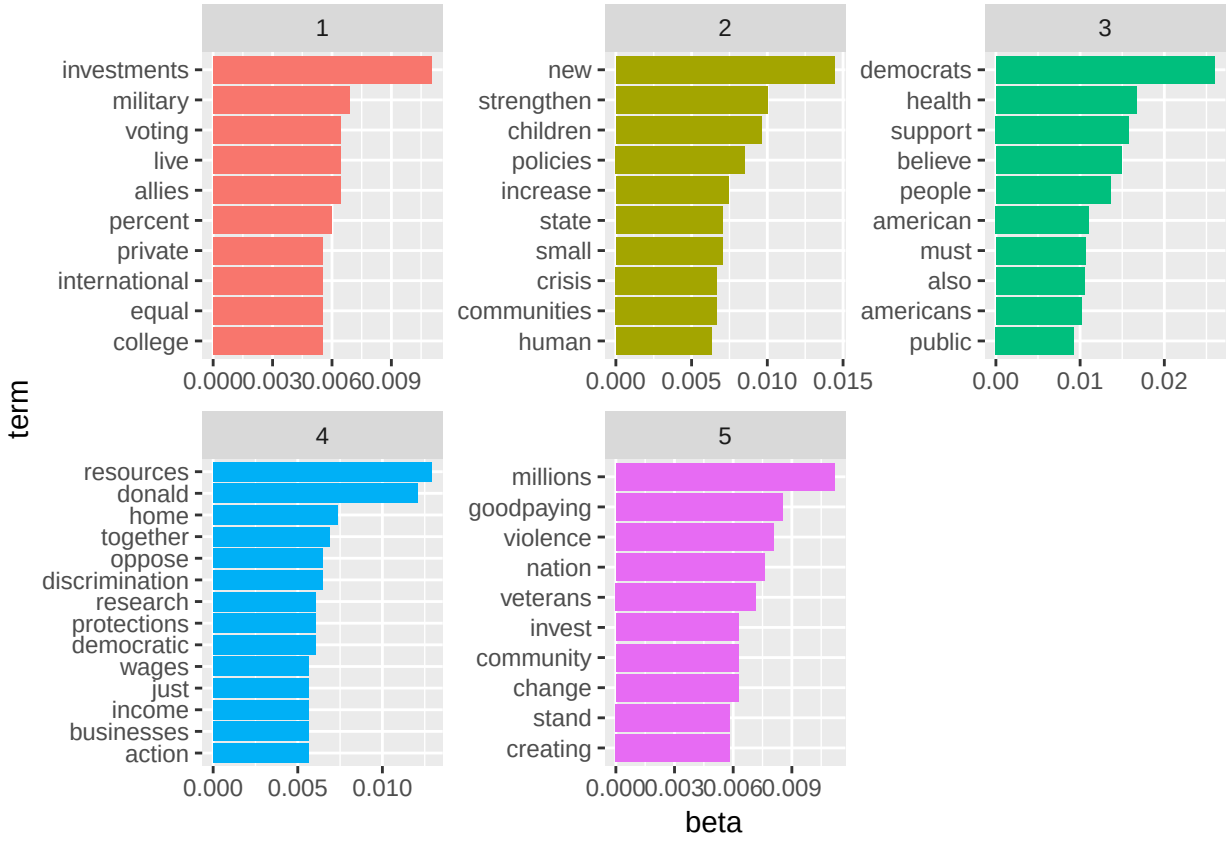
As we saw with the word frequencies and word clouds, most topics here (and especially Topics 1 & 3) reflect how the Republicans speak more about systems, while the Democrats refer more to the issues to be addressed through them.

Both groups speak about rights in differnt ways. Each also seems to favour one subject area not addressed by the other party. For example, Democrats appear to have a special focus on minorities and equality (as evident from Topic 5). Republicans seem to speak about regulation in a globalized world (Topic 4)

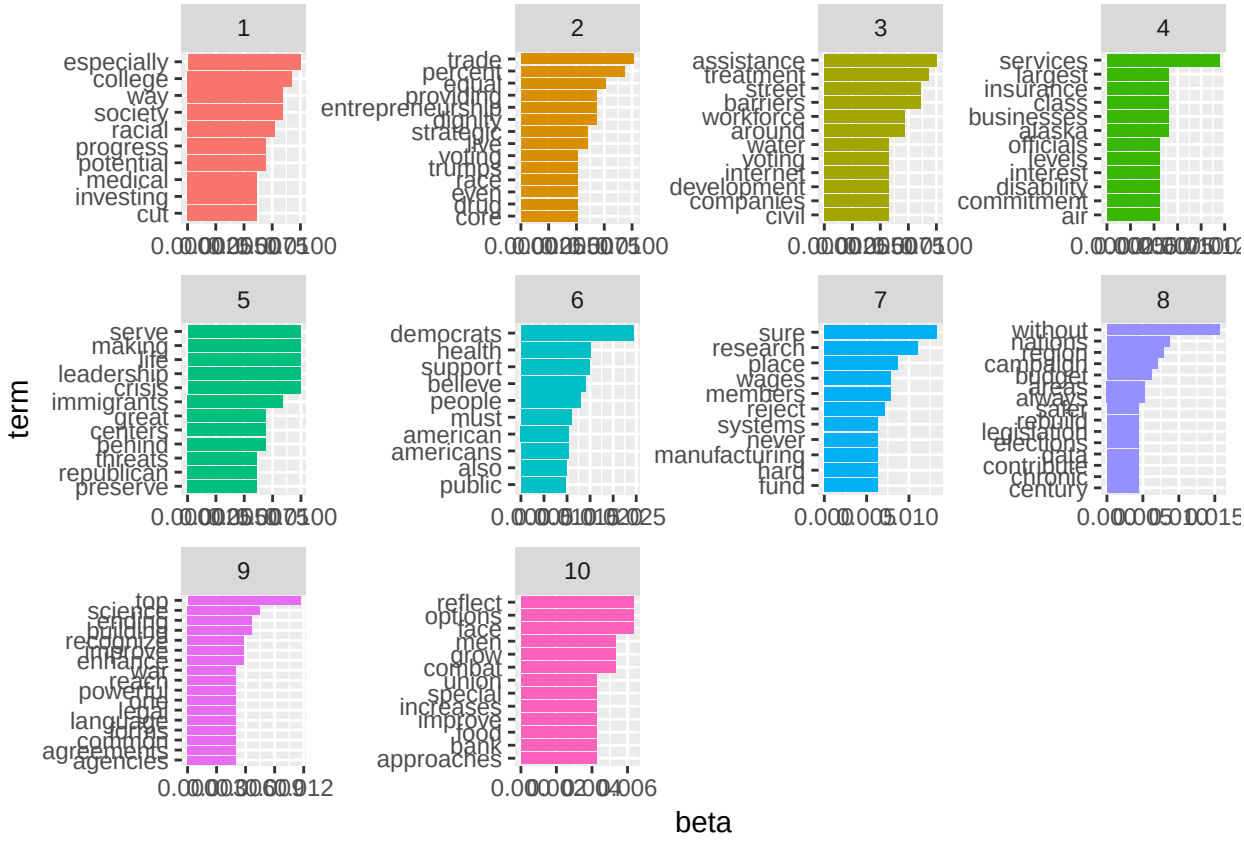
8. Fit 6 more topic models at the follow levels of `k` *for each party:* `5, 10, 25`*. Present the results however you'd like (e.g., visually and/or numerically).*

Since we have already generated all the plots on loop, the results will be presented visually.

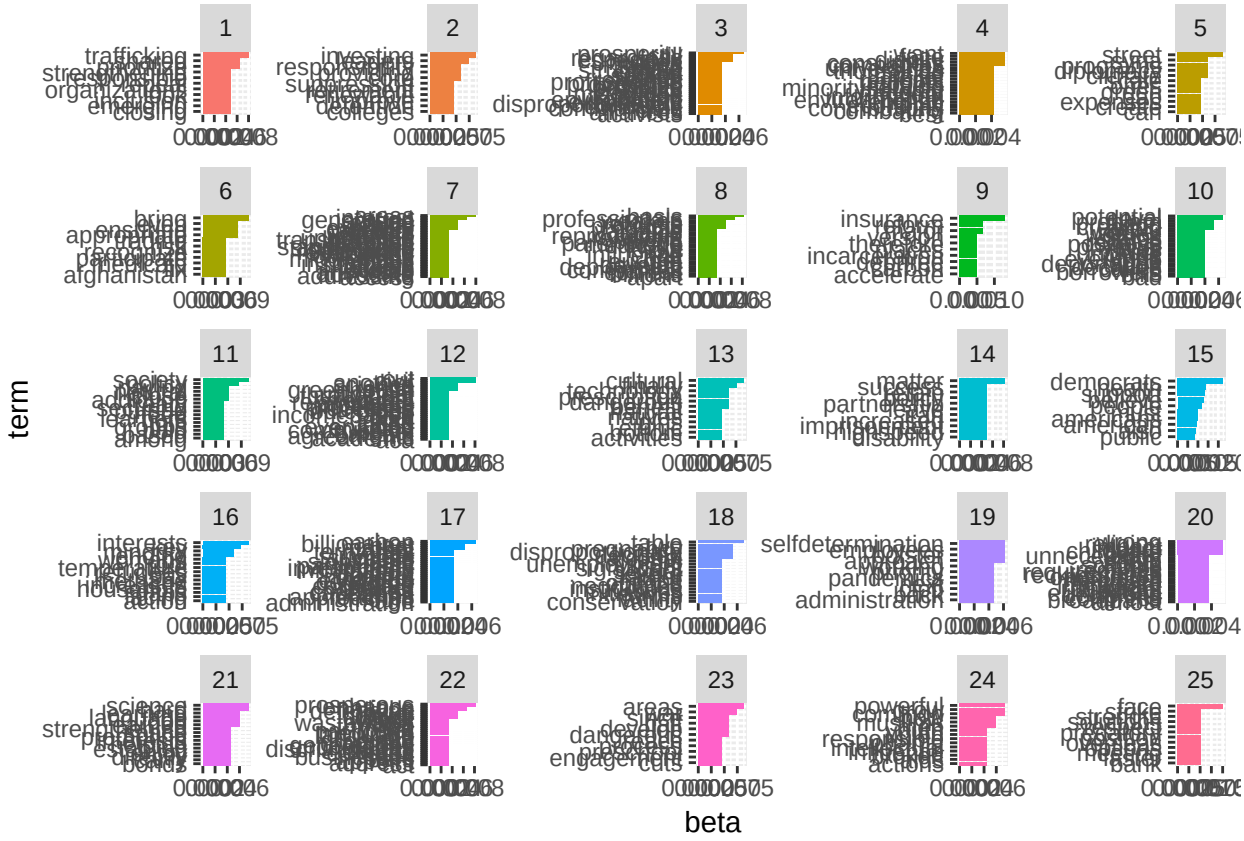
```
democ_dtm5
```


1
investments
military
voting
live
allies
percent
private
international
equal
college
0.000
0.003
0.006
0.009
2
new
strengthen
children
policies
increase
state
small
crisis
communities
human
0.000
0.005
0.010
0.015
3
democrats
health
support
believe
people
american
must
also
americans
public
0.00
0.01
0.02
4
resources
donald
home
together
oppose
discrimination
research
protections
democratic
wages
just
income
businesses
action
0.000
0.005
0.010
5
millions
goodpaying
violence
nation
veterans
invest
community
change
stand
creating
0.000
0.003
0.006
0.009
term
beta

```
democ_dtm10
```


1
especially
college
way
society
racial
progress
potential
medical
investing
cut
2
trade
percent
equal
providing
entrepreneurship
dignity
strategic
live
voting
trumps
face
even
drug
core
3
assistance
treatment
street
barriers
workforce
around
water
voting
internet
development
companies
civil
4
services
largest
insurance
class
businesses
alaska
officials
levels
interest
disability
commitment
air
5
serve
making
life
leadership
crisis
immigrants
great
centers
behind
threats
republican
preserve
6
democrats
health
support
believe
people
must
american
americans
also
public
7
sure
research
place
wages
members
reject
systems
never
manufacturing
hard
fund
8
without
nations
region
campaign
budget
green
always
safer
rebuild
legislation
elections
data
contribute
chronic
century
9
top
science
ending
building
recognize
improve
enhance
war
reach
powerful
home
legal
language
future
common
agreements
agencies
10
reflect
options
face
men
grow
combat
union
special
increases
improve
food
bank
approaches
term
beta

```
democ_dtm25
```

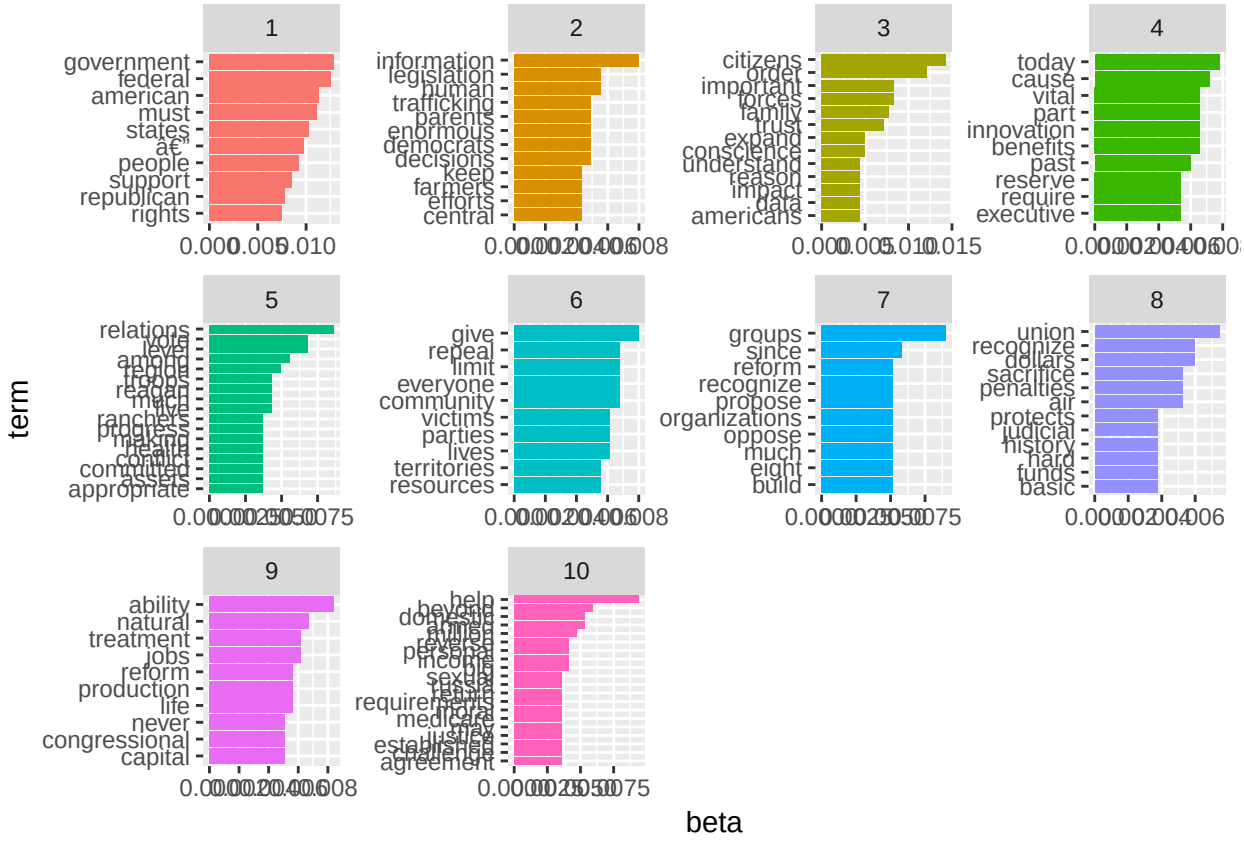


Now for Republicans: First, with 5 topics

```
repub_dtm5
```

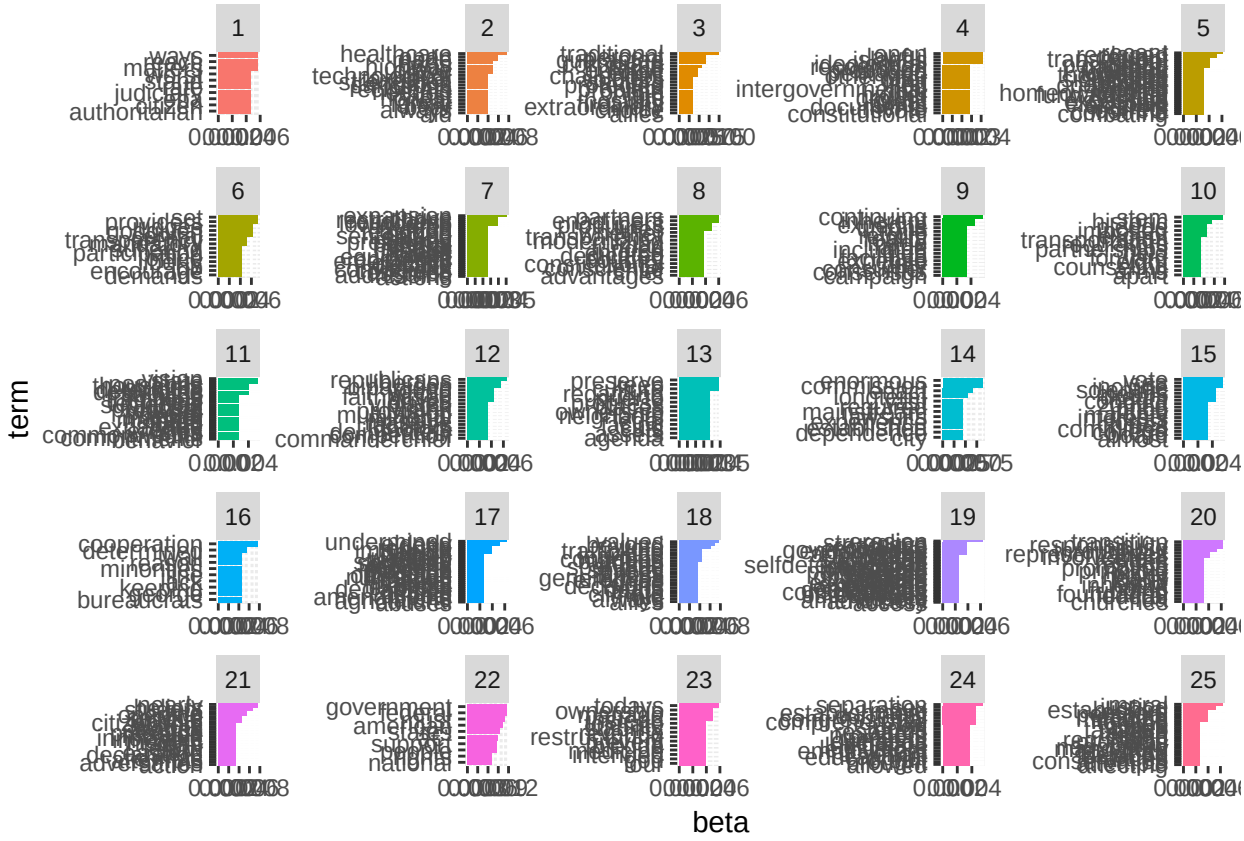
Next, with 10 topics

```
repub_dtm10
```



For 25 topics

```
repub_dtm25
```



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

A low perplexity is associated with a better fit for the Topic Model. So we now check for the lowest such score on all the models calculated above.

```
print(c("Perplexities from Democrat Topic Models at k=5,10 and 25:", perplex_democ5, perplex_democ10, p
```

```
## [1] "Perplexities from Democrat Topic Models at k=5,10 and 25:"
## [2] "1634.83951522735"
## [3] "1637.02329452451"
## [4] "1641.06488266786"
```

```
print(c("Perplexities from Republican Models at k=5,10 and 25:", perplex_repub5, perplex_repub10, perpl
```

```
## [1] "Perplexities from Republican Models at k=5,10 and 25:"
## [2] "2304.03319024035"
## [3] "2305.77754357422"
## [4] "2310.46828489502"
```

For both Democrats and Republicans, the model wih 25 topics has the lowest perplexity. If this alone became the grounds for selecting the number of topics, then we would choose 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?

Looking back at k=10, the topics are still not entirely clear from the top terms. However, the overall theme persists. The Democrats speak about matters around education, community, and health. For example, Topic 7 seems to refer to Minority groups (race, religion, etc). However, in other areas, such as economic and financial matters in Topic 4, we now have distinct topics emerging. The Democrats again speak about

government, federal systems and institutions. However, we now see a large number of verbs emerging in several topics (such as Topic 7 and 10), which seem to suggest an orientation towards action rather than discussion first.

Strangely, at k= 5, these same themes do not seem to emerge with a higher degree of clarity. Although perplexity is lower, the subjective interpretability seems to rise for 10 topics.

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)?

At a purely hypothetical level, the Republican party seems to be less concerned with its own name, as it is with government, rights and the federal system. They alsp have a higher proportion of action-oriented language, and belief in Presidents. If I were to use only this text to help me choose a decisive leader, I would pick the Republican party. Fortunately, this is all hypothetical.