HW5

Arun Pandian

11/25/2019

1.

Loading the data and displaying the first

```
suppressMessages((library(tm)))
```

```
## [1] "tm"        "NLP"       "stats"     "graphics"  "grDevices" "utils"
## [7] "datasets"  "methods"   "base"
```

```
suppressMessages(library(grid))
suppressMessages(library(wordcloud))
suppressMessages(library(wordcloud2))
suppressMessages(library(tidyverse))
texts <- file.path("~", "Desktop", "texts")
texts<-"/Users/arun/Dropbox/2019 Fall/MACS 40800/hw/hw5/Problem-Set-5-master/Party Platforms Data/texts
docs <- VCorpus(DirSource(texts))
summary(docs)
```

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

```
#split <- strsplit(as.character(docs[1]),split=" ")
#split
#writeLines(as.character(docs))
dtm <- DocumentTermMatrix(docs)
frequency <- sort(colSums(as.matrix(dtm)),
                  decreasing=TRUE) # add number of times each term is used, and sorting based on freque
as.data.frame(frequency[1:50]) # most frequently used words
```

```
##             frequency[1:50]
## the                    3425
## and                    2895
## that                    818
## for                     747
## our                     691
## will                    646
## their                   426
## with                    425
## are                     356
## have                    289
## not                     241
## from                    230
## all                     222
## must                    217
## democrats               215
## support                 215
## should                  207
## has                     206
## american                202
```

```
## federal                 194
## who                     184
## they                    175
## more                    164
## people                  159
## its                     158
## health                  150
## believe                 147
## those                   147
## government              139
## which                   137
## this                    136
## public                  135
## national                129
## other                   123
## against                 121
## rights                  121
## can                     117
## states                  111
## also                    110
## americans               106
## new                     102
## than                    101
## make                     98
## work                     98
## been                     97
## republican               97
## economic                 96
## united                   96
## including                91
## president                90
```

As we can see from the table, we need to remove stop words. Inspection of texts (not shown here for brevity), also reveals the need to remove not only stopwords, lowercase text, whitespace, etc. but also remove combinations of punctuation with words.

2.

Clean the corpus and create two separate document-term matrices. This is an iterative process where I cleaned, looked at the frequency tables and word cloud results and then went back to clean again. In addition to the remarks above about punctuation containing words, I decided to remove "will", "also" and "must" since they both common across both parties and seemed to add very little semantic value at the outset. I have also joined together words that seemed to occur frequently together, that are one semantic concept.

```
docs <- tm_map(docs, tolower)
docs <- tm_map(docs, removePunctuation)
docs <- tm_map(docs, removeNumbers)

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("\n\"\",", " ", docs[[j]])
```

```r
  docs[[j]] <- gsub("\"\",", " ", docs[[j]])

}
docs <- tm_map(docs,
               removeWords,
               stopwords("english"))


docs <- tm_map(docs, PlainTextDocument) # redefine



for (j in seq(docs)) {
  docs[[j]] <- gsub("health care", "health-care", docs[[j]])
  docs[[j]] <- gsub("donald trump", "donald-trump", docs[[j]])
  docs[[j]] <- gsub("united states", "united-states", docs[[j]])
}

docs <- tm_map(docs, removeWords, c("will", "also", "must"))

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


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

```
## <<DocumentTermMatrix (documents: 1, terms: 3849)>>
## Non-/sparse entries: 3849/0
## Sparsity           : 0%
## Maximal term length: 22
## Weighting          : term frequency (tf)
```

```r
#Inspecting most frequent words in the democratic party corpus
frequency <- sort(colSums(as.matrix(dtm)),
                  decreasing=TRUE)
as.data.frame(frequency[1:50])
```

```
##             frequency[1:50]
## democrats               207
## support                 123
## believe                 117
## people                  107
## americans                90
## health                   88
## american                 86
## communities              80
## public                   79
## rights                   71
## work                     71
## make                     66
## federal                  64
## country                  60
## fight                    58
## including                57
```

```
## jobs                          55
## workers                       54
## america                       51
## ensure                        50
## can                           49
## education                     48
## national                      47
## need                          47
## access                        46
## continue                      46
## programs                      46
## protect                       46
## economic                      45
## world                         45
## energy                        43
## climate                       42
## families                      42
## health-care                   42
## new                           42
## economy                       41
## help                          41
## security                      40
## students                      40
## government                    39
## provide                       39
## women                         39
## efforts                       37
## right                         37
## committed                     36
## every                         36
## global                        36
## schools                       36
## build                         35
## end                           34
```

```r
dtm2 <- DocumentTermMatrix(docs[2])

#Inspecting most frequent words in the republican party corpus
frequency2 <- sort(colSums(as.matrix(dtm2)),
                  decreasing=TRUE)
as.data.frame(frequency2[1:50])
```

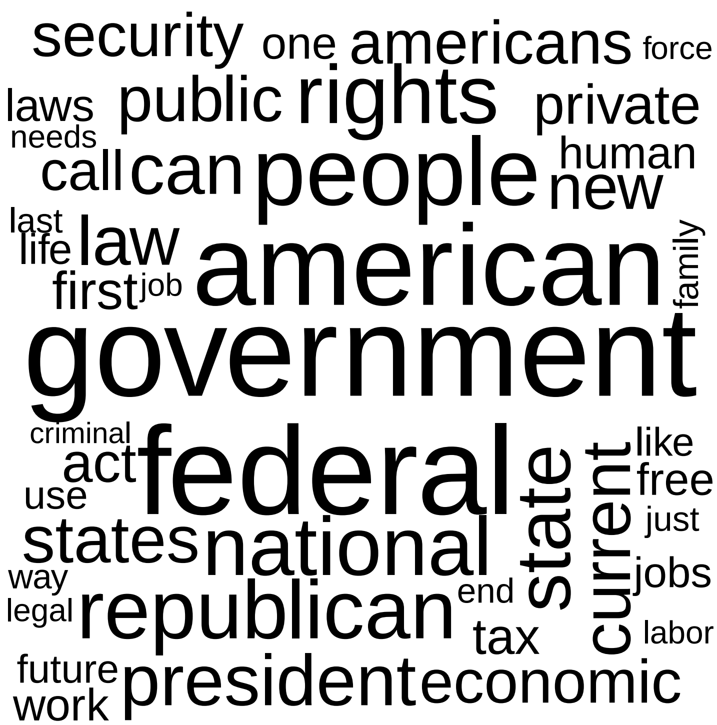
```
##                  frequency2[1:50]
## government                    137
## federal                       134
## american                      121
## support                       100
## people                         98
## national                       83
## republican                     83
## rights                         83
## congress                       81
## state                          74
## president                      70
## can                            69
```

```
## law                          66
## current                      65
## states                       63
## new                          60
## public                       60
## americans                    57
## economic                     56
## security                     56
## economy                      51
## military                     51
## administration               50
## private                      50
## act                          49
## education                    49
## world                        49
## call                         48
## country                      48
## religious                    48
## first                        47
## united-states                47
## right                        46
## americas                     45
## energy                       44
## nations                      44
## health                       43
## oppose                       43
## amendment                    42
## america                      42
## every                        42
## freedom                      42
## tax                          42
## families                     41
## urge                         40
## party                        38
## policies                     38
## protect                      38
## free                         37
## human                        37
```

3.

```r
wordcloud(names(frequency), frequency
          ,random.order = FALSE, max.words = 300)
```


```r
wordcloud(names(frequency2), frequency2,
          random.order = FALSE, max.words = 200)
```



These word clouds as well as the frequency tables printed above, suggest that the democrats seem to take on more agency in their data generation, being that 'democrats' is a word that occurs very frequently. Health and health related topics also take up a prominent space. Interestingly, workers are also mentioned while there is not prominent mention of religion. Both the words "communities" and "Indians" are mentioned, suggesting an preoccupation with minority groups, at least in rhetoric. It has to be noted, "students" are also mentioned here prominently.

In the case of republicans, there is no prominent mentions of workers, there is prominent mention of religion. Additionally, they seem to give themselves less agency, as their most frequent term is government and not "republicans". This may suggest that that either they are abstract in their data generation process or they are doing more polemical work than positive work. Interestingly they mention rights, to I checked what kind of rights they talk about(and if this is different from the democrats). I report those results below.

The wordcloud(s) suggest then a difference in framing and content between the two parties.

```
#health, federal, donald
suppressMessages(library(quanteda))
kwic(x = as.character(docs),
      pattern = "health",
     window = 2)

kwic(x = as.character(docs),
     pattern = "donald",
     window = 2)

kwic(x = as.character(docs),
     pattern = "states",
      window = 2)
#confirming what indians are being talked about
kwic(x = as.character(docs),
       pattern = "indian",
       window = 2)
#checkgin if democrats and republicans talk about rights differently
kwic(x = as.character(docs[1]),
     pattern = "rights",
     window = 2)
kwic(x = as.character(docs[2]),
     pattern = "rights",
     window = 2)

kwic(x = as.character(docs[1]),
     pattern = "climate",
     window = 2)
```

The analysis of certain words in context revealed(not shown here for brevity, but code is shown above) that, Donald trump should be one word. Additionally, united states should also be tokenized together. Interestingly "health" seems to occur together with many words and therefore, I have not made any effort to group it with another word. Additionally, there is a difference between the "rights" that democrats and republicans talk about. Climate seems to occur with change mainly, which is to be expected. The democrats talk about it more frequently than the republicans. Democrats speak of workers rights, civil rights, voting rights, reproductive rights, lgbt rights; the republicans use this word differently, they typically seem to use it in the context of constitutional, natural, inalienable, property, individual and first amendment rights.

4 and 5.

For sentiment analysis, I used tidy text and modified the code found at https://www.tidytextmining.com/sentiment.html.

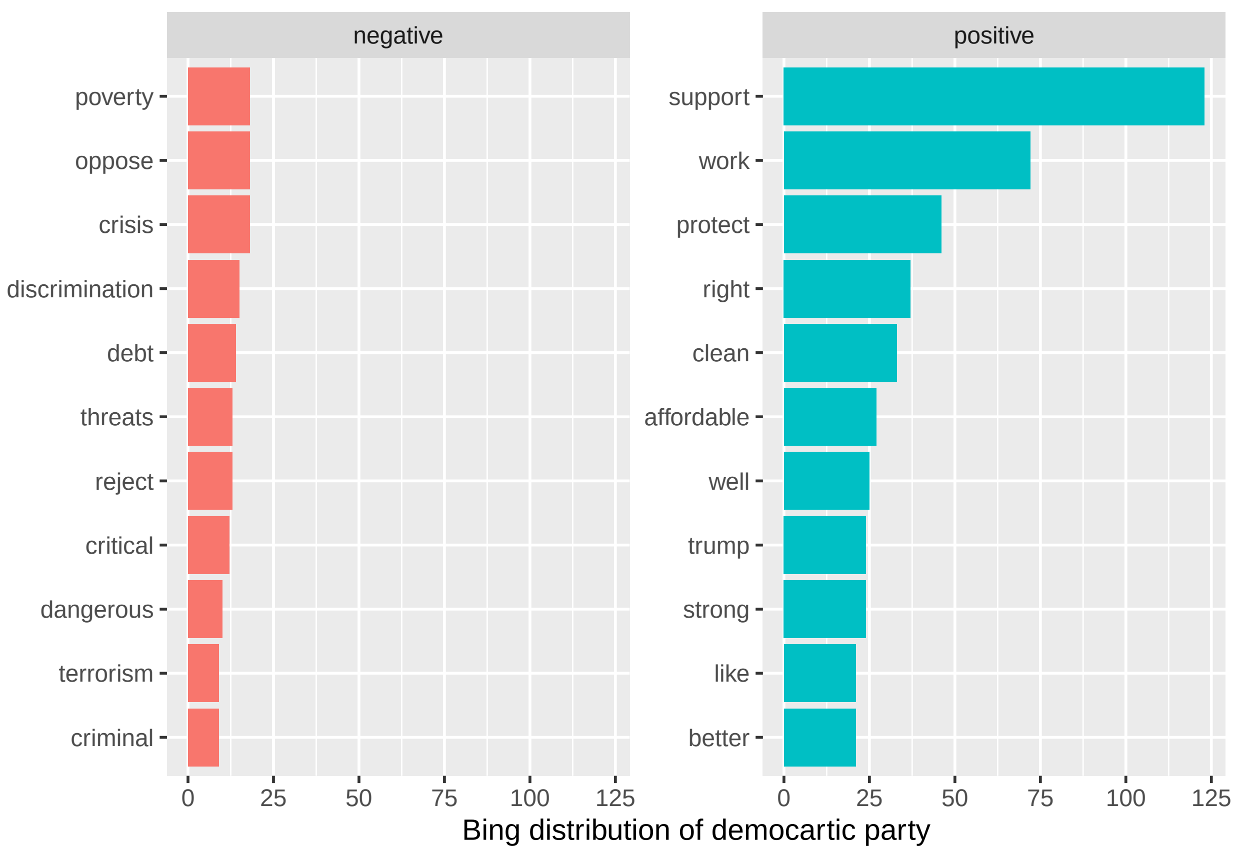
```
suppressMessages(library(tidytext))
suppressMessages(library(textdata))
suppressMessages(library(tidyr))
```

```
dem_corpus <- tidy(docs[1]) %>%unnest_tokens(word,text)
rep_corpus <- tidy(docs[2]) %>%unnest_tokens(word,text)


dem_bing_word_counts <- dem_corpus %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort = TRUE)
```

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

```
dem_bing_word_counts %>%
  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 = "Bing distribution of democartic party",
       x = NULL) +
  coord_flip()
```

```
## Selecting by n
```

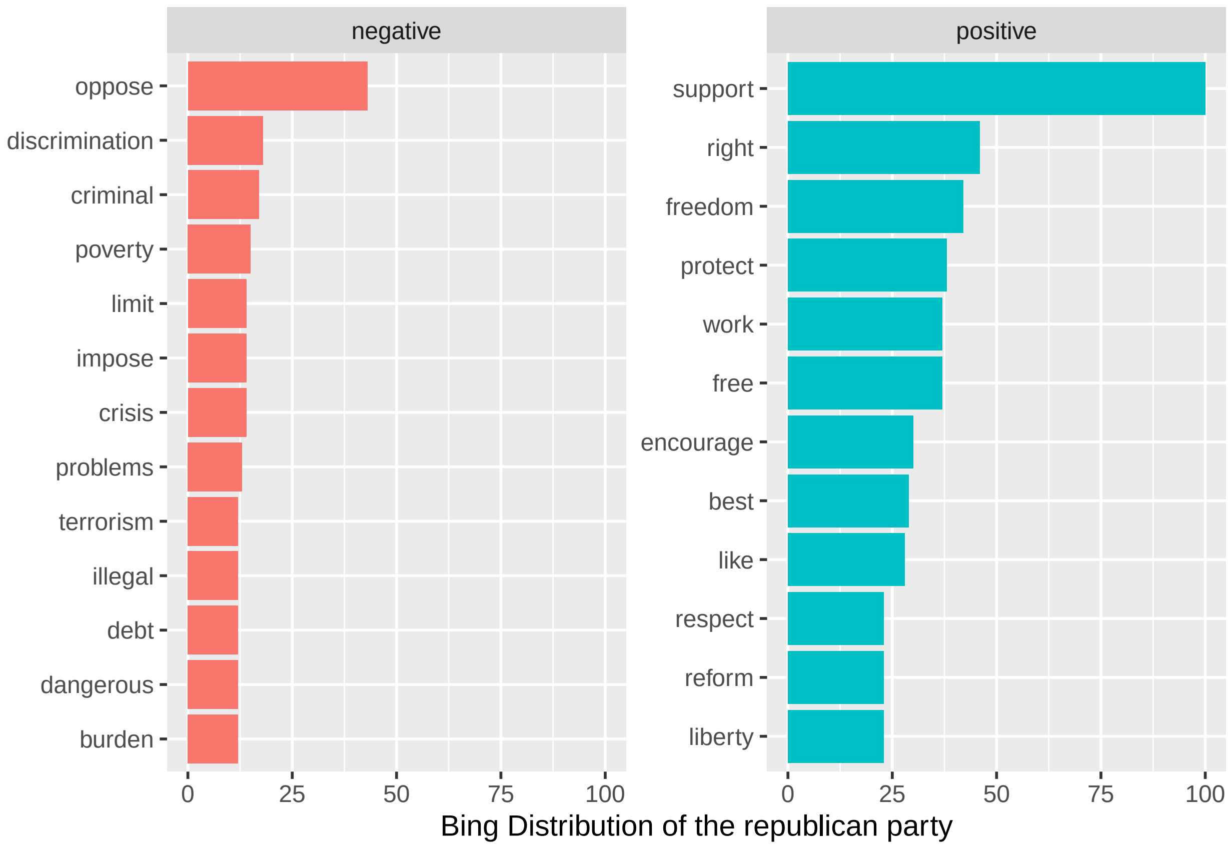



```
rep_bing_word_counts <- rep_corpus %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort = TRUE)
```

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

```
rep_bing_word_counts %>%
  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 = "Bing Distribution of the republican party",
       x = NULL) +
  coord_flip()
```

```
## Selecting by n
```



Just looking at the bing distributions, we can see that the democratic party uses negative words less frequently and positive words more frequently.

Let's now calculate some overall scores for both parties using Afinn:

```
dem_afinn_mean_score <-dem_corpus%>%
  inner_join(get_sentiments("afinn"))%>%summarise(sentiment=sum(value))/nrow(dem_corpus%>%inner_join(ge
```

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

```
#Democratic platform mean afinn score
dem_afinn_mean_score
```

```
##   sentiment
## 1  0.562851
```

```r
rep_afinn_mean_score <-rep_corpus%>%
  inner_join(get_sentiments("afinn"))%>%summarise(sentiment=sum(value))/nrow(rep_corpus%>%inner_join(ge
```

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

```r
#Republican platform mean afinn score
rep_afinn_mean_score
```

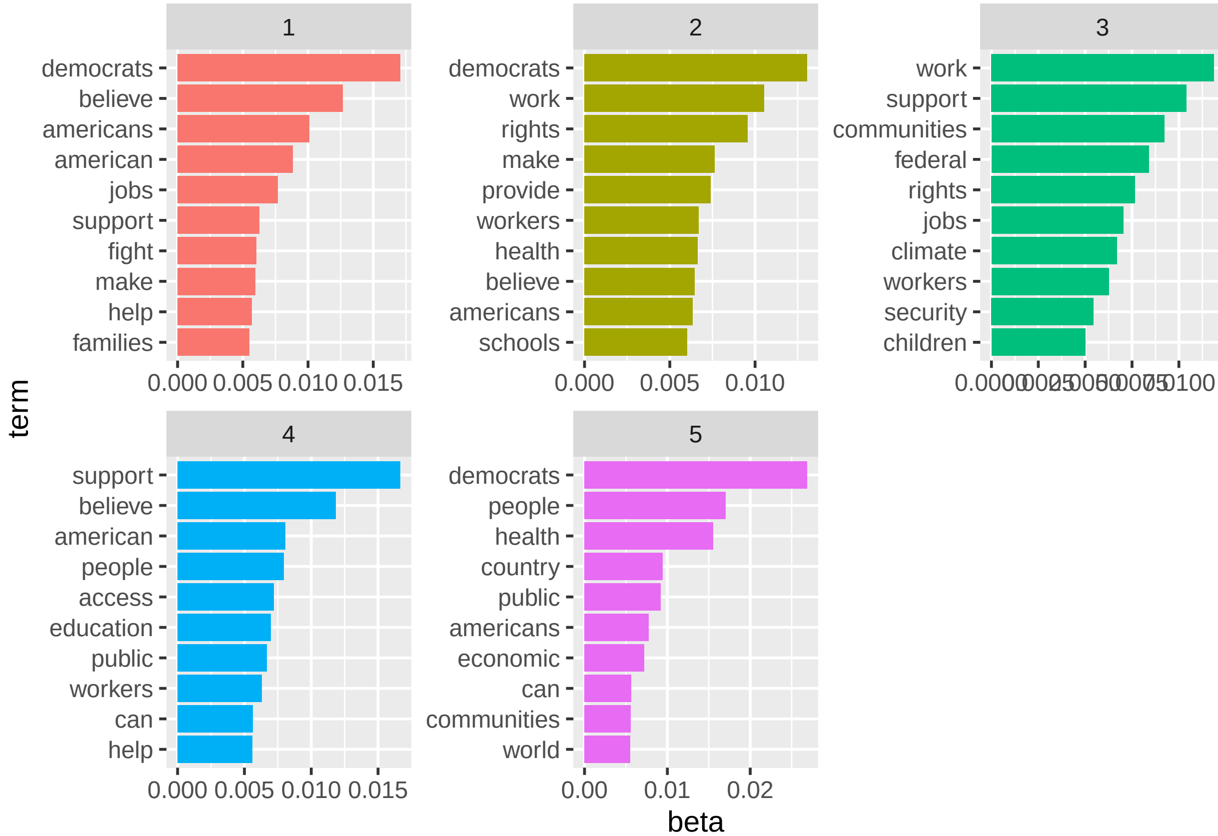
```
##   sentiment
## 1 0.3540724
```

Once again, the democrats seems to be more positive, when we calculate mean afinn scores. Therefore overall(using bing or afinn) the democrats are more positive in their platform.

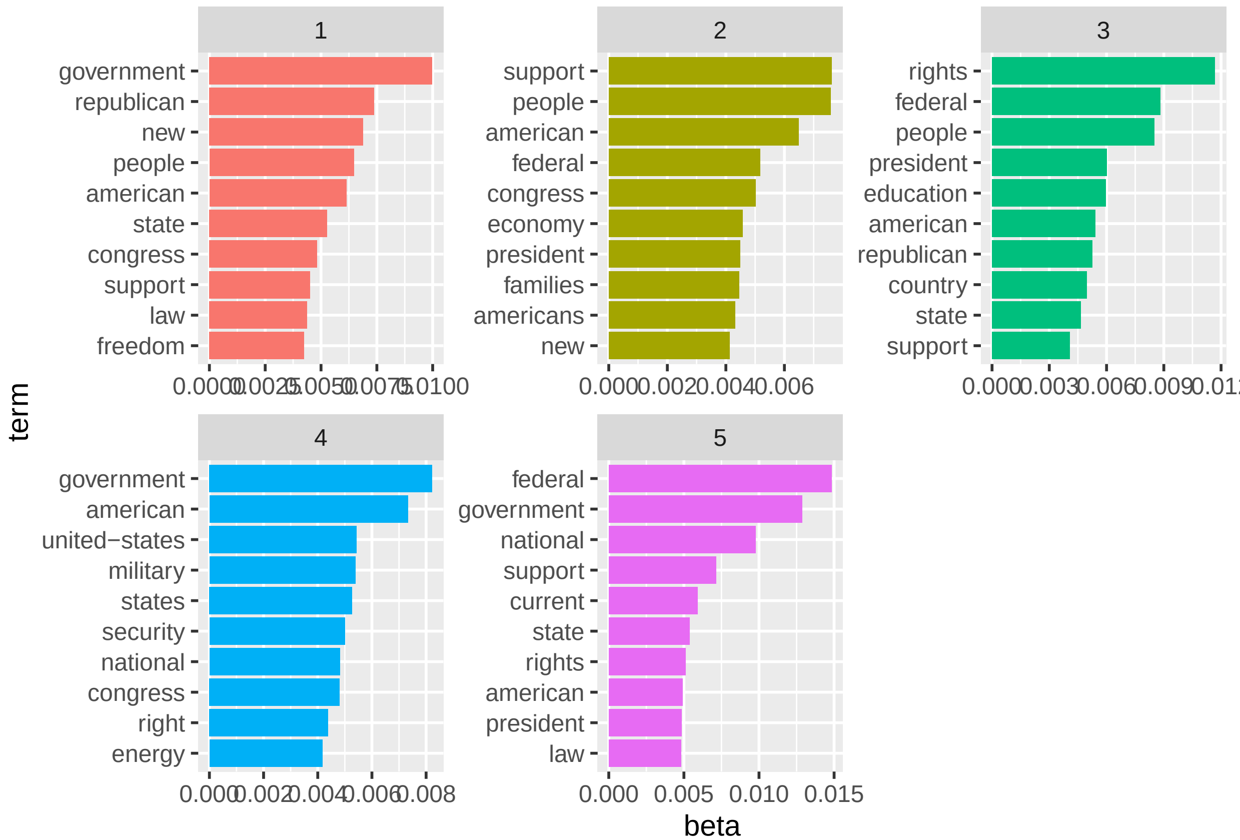
6.

For this problem, we are looking at documents as a bag of words, each of which can belong to multiple topics. For this purpose, I will use the tidy text package, and the documentation available online at (https://www.tidytextmining.com/topicmodeling.html). I will however not use the package for preprocessing as I have already done that above. First I will create the topic models and visualizations for democrats and then the same for republicans.

```r
library(topicmodels)

library(ggplot2)
library(dplyr)
plot_func <- function(dtm_x, k){

x_lda <- LDA(dtm_x, k=k, control= list(seed=5656))
x_topics <- tidy(x_lda, matrix = "beta")
#get top 10 terms in each topic and sort within topics
x_top_terms <- x_topics %>%
  group_by(topic)  %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
#plot
x_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()
}

plot_func(dtm,5)
```




```
plot_func(dtm2,5)
```



7.

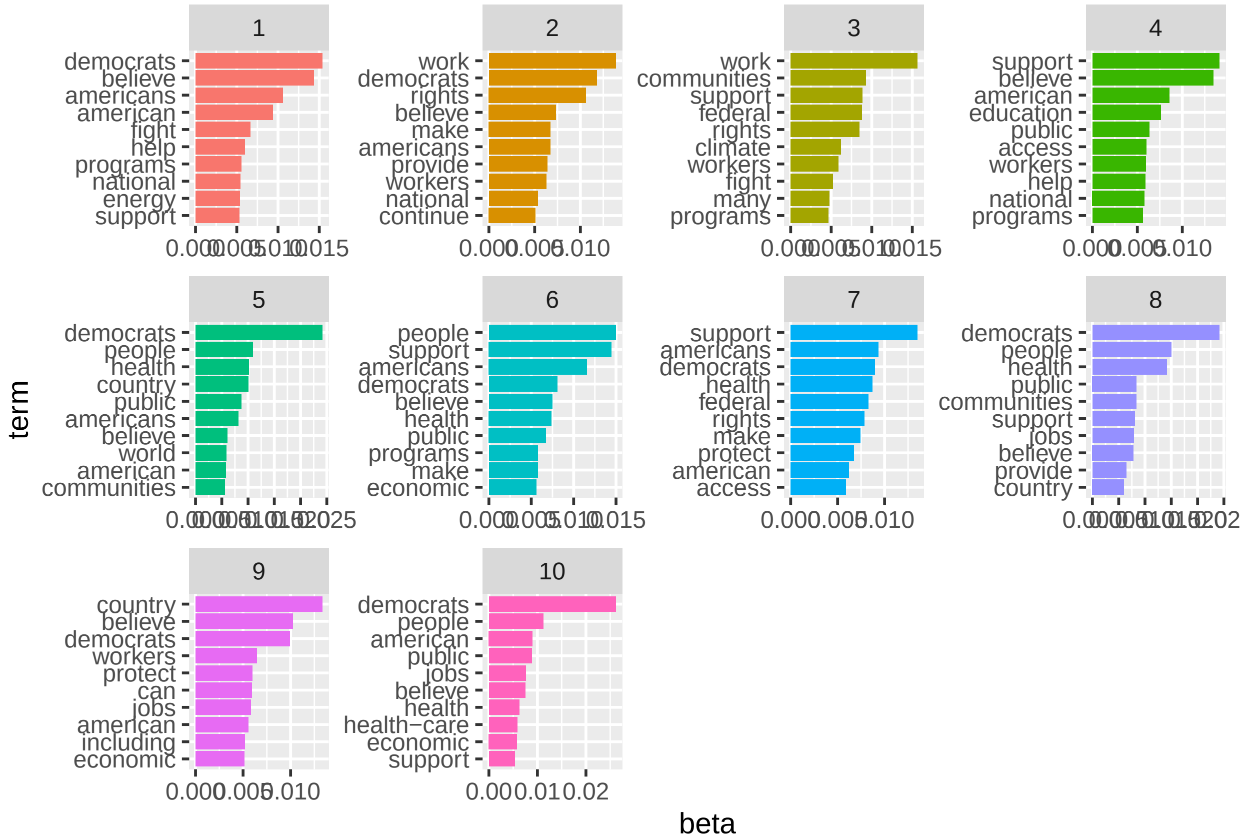
Topic modelling lends clarity to the differences I mentioned in section 3 between the parties. However k=5, might not be the ideal number of topics as it seems like multiple topics are grouped together. For instance, in topic three for the democrats, both climate change and workers are grouped together. In topic 4 again, workers are grouped together education(access?). In topic 5, health is grouped with world. It maybe that these topics are coherent, but at the surface level doesn't seem to be case(Unless one starts thinking about how these recurring concepts are integrated somehow into different issues).

For the republicans, there are recurring words across topics, but the integration of those words into different topics seems tenable. For instance, national (security?) and energy occur together, while national (support?) also co-occurs in the same topic as state (rights). This seems much more tenable than health, world and economy occurring in the same topic. I think topic 3 for republicans is actually very coherent, talking about education in the context of states rights. Topic 1 seems to be about congress passing laws for promulgate 'freedom'. Topic two seems to be about the economy and american families although it's not clear if it's positive or negative. Topic 4 is about national security and energy (clearly related concepts). Topic 5 seems to be about states rights in general.

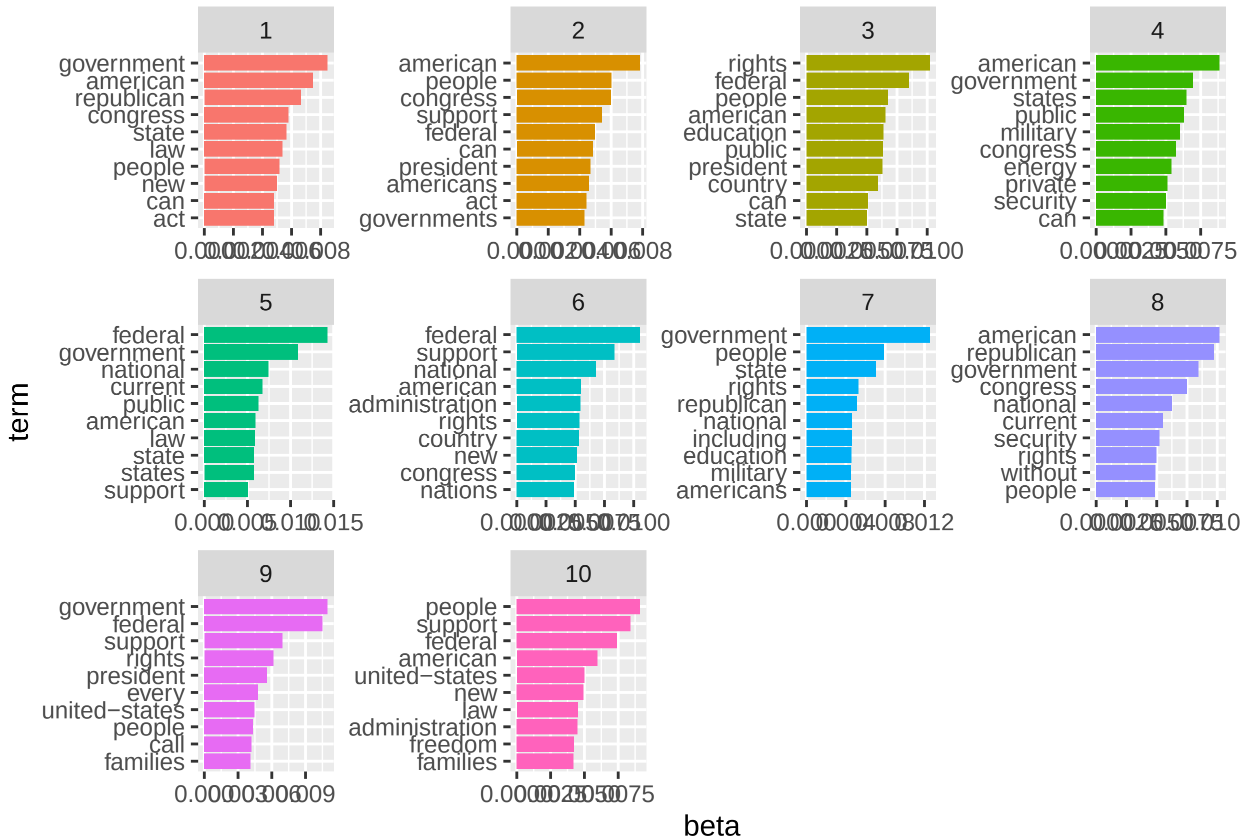
8.

Since topic models for k=5 are fitted, I will fit below the four models for k= 10,25.

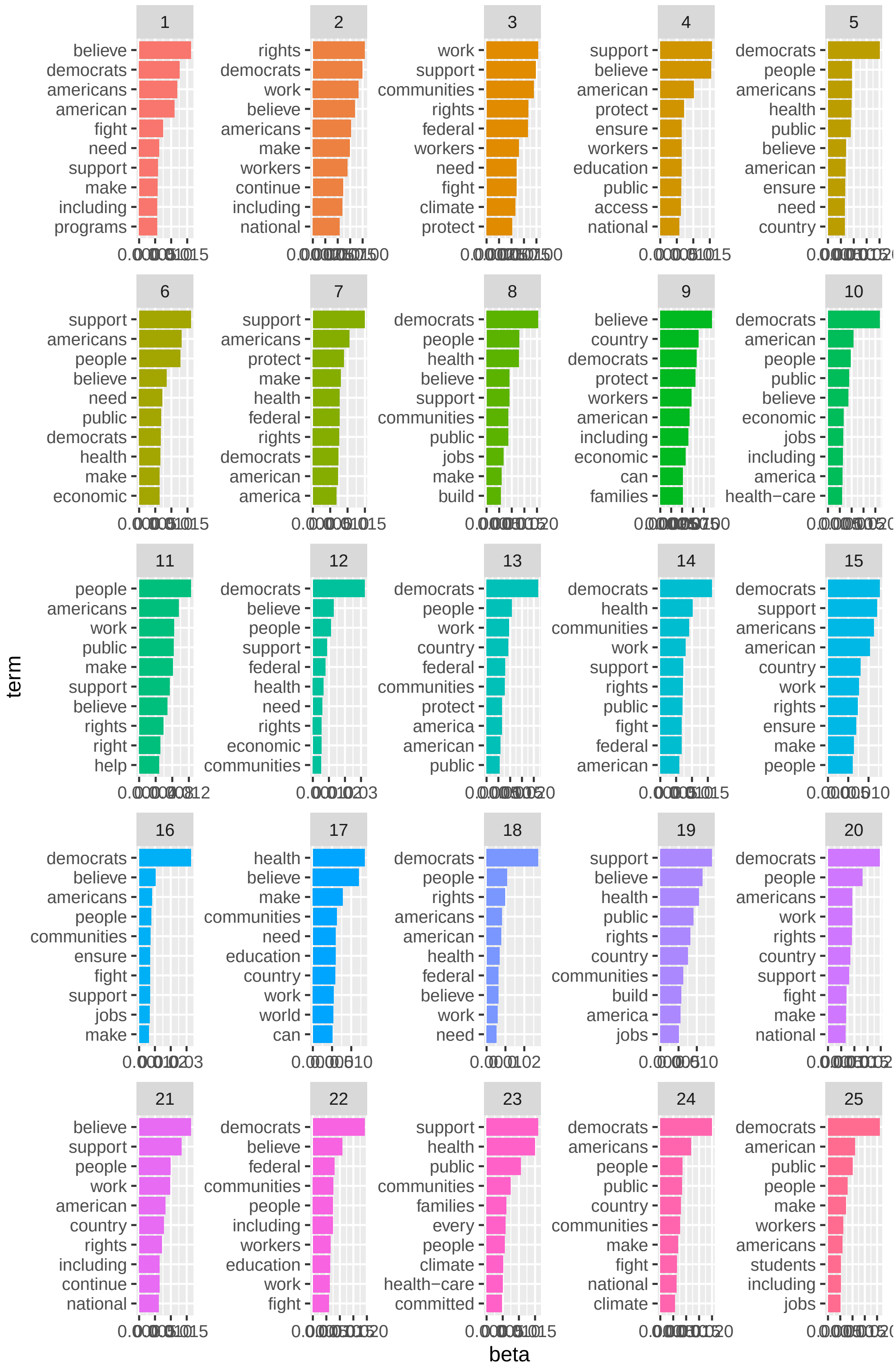
```
#democratic corpus at k=10
plot_func(dtm,10)
```




```
#republican corpus at k=10
plot_func(dtm2,10)
```

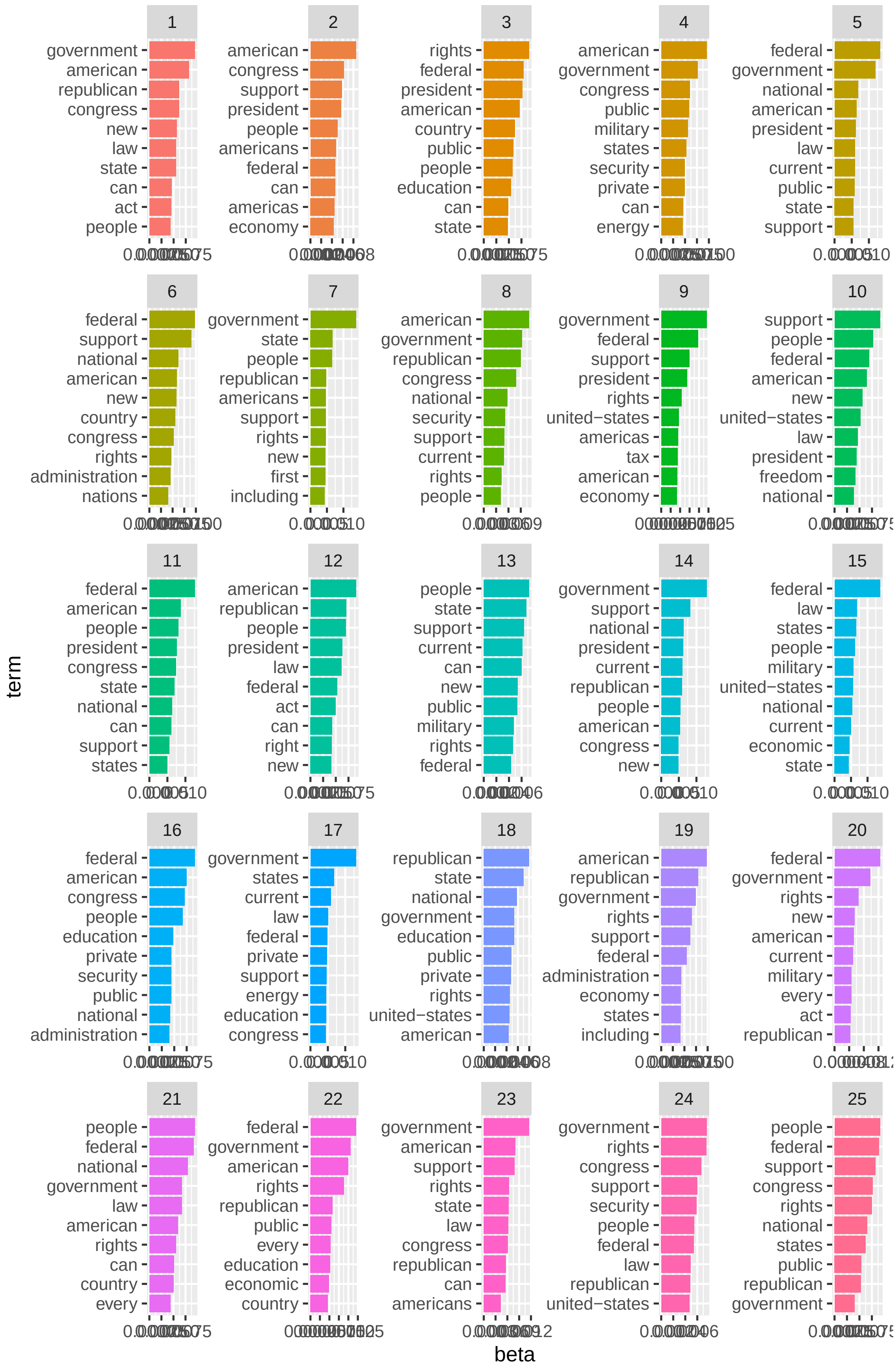



```r
#democratic corpus at k=25
plot_func(dtm,25)
```



1
believe
democrats
americans
american
fight
need
support
make
including
programs
2
rights
democrats
work
believe
americans
make
workers
continue
including
national
3
work
support
communities
rights
federal
workers
need
fight
climate
protect
4
support
believe
american
protect
ensure
workers
education
public
access
national
5
democrats
people
americans
health
public
believe
american
ensure
need
country
6
support
americans
people
believe
need
public
democrats
health
make
economic
7
support
americans
protect
make
health
federal
rights
democrats
american
america
8
democrats
people
health
believe
support
communities
public
jobs
make
build
9
believe
country
democrats
protect
workers
american
including
economic
can
families
10
democrats
american
people
public
believe
economic
jobs
including
america
health−care
11
people
americans
work
public
make
support
believe
rights
right
help
12
democrats
believe
people
support
federal
health
need
rights
economic
communities
13
democrats
people
work
country
federal
communities
protect
america
american
public
14
democrats
health
communities
work
support
rights
public
fight
federal
american
15
democrats
support
americans
american
country
work
rights
ensure
make
people
16
democrats
believe
americans
people
communities
ensure
fight
support
jobs
make
17
health
believe
make
communities
need
education
country
work
world
can
18
democrats
people
rights
americans
american
health
federal
believe
work
need
19
support
believe
health
public
rights
country
communities
build
america
jobs
20
democrats
people
americans
work
rights
country
support
fight
make
national
21
believe
support
people
work
american
country
rights
including
continue
national
22
democrats
believe
federal
communities
people
including
workers
education
work
fight
23
support
health
public
communities
families
every
people
climate
health−care
committed
24
democrats
americans
people
public
country
communities
make
fight
national
climate
25
democrats
american
public
people
make
workers
americans
students
including
jobs
term
beta

```r
#republican corpus at k=25
plot_func(dtm2,25)
```


1
government
american
republican
congress
new
law
state
can
act
people
2
american
congress
support
president
people
americans
federal
can
americas
economy
3
rights
federal
president
american
country
public
people
education
can
state
4
american
government
congress
public
military
states
security
private
can
energy
5
federal
government
national
american
president
law
current
public
state
support
6
federal
support
national
american
new
country
congress
rights
administration
nations
7
government
state
people
republican
americans
support
rights
new
first
including
8
american
government
republican
congress
national
security
support
current
rights
people
9
government
federal
support
president
rights
united−states
americas
tax
american
economy
10
support
people
federal
american
new
united−states
law
president
freedom
national
11
federal
american
people
president
congress
state
national
can
support
states
12
american
republican
people
president
law
federal
act
can
right
new
13
people
state
support
current
can
new
public
military
rights
federal
14
government
support
national
president
current
republican
people
american
congress
new
15
federal
law
states
people
military
united−states
national
current
economic
state
16
federal
american
congress
people
education
private
security
public
national
administration
17
government
states
current
law
federal
private
support
energy
education
congress
18
republican
state
national
government
education
public
private
rights
united−states
american
19
american
republican
government
rights
support
federal
administration
economy
states
including
20
federal
government
rights
new
american
current
military
every
act
republican
21
people
federal
national
government
law
american
rights
can
country
every
22
federal
government
american
rights
republican
public
every
education
economic
country
23
government
american
support
rights
state
law
congress
republican
can
americans
24
government
rights
congress
support
security
people
federal
law
republican
united−states
25
people
federal
support
congress
rights
national
states
public
republican
government
term
beta

9.

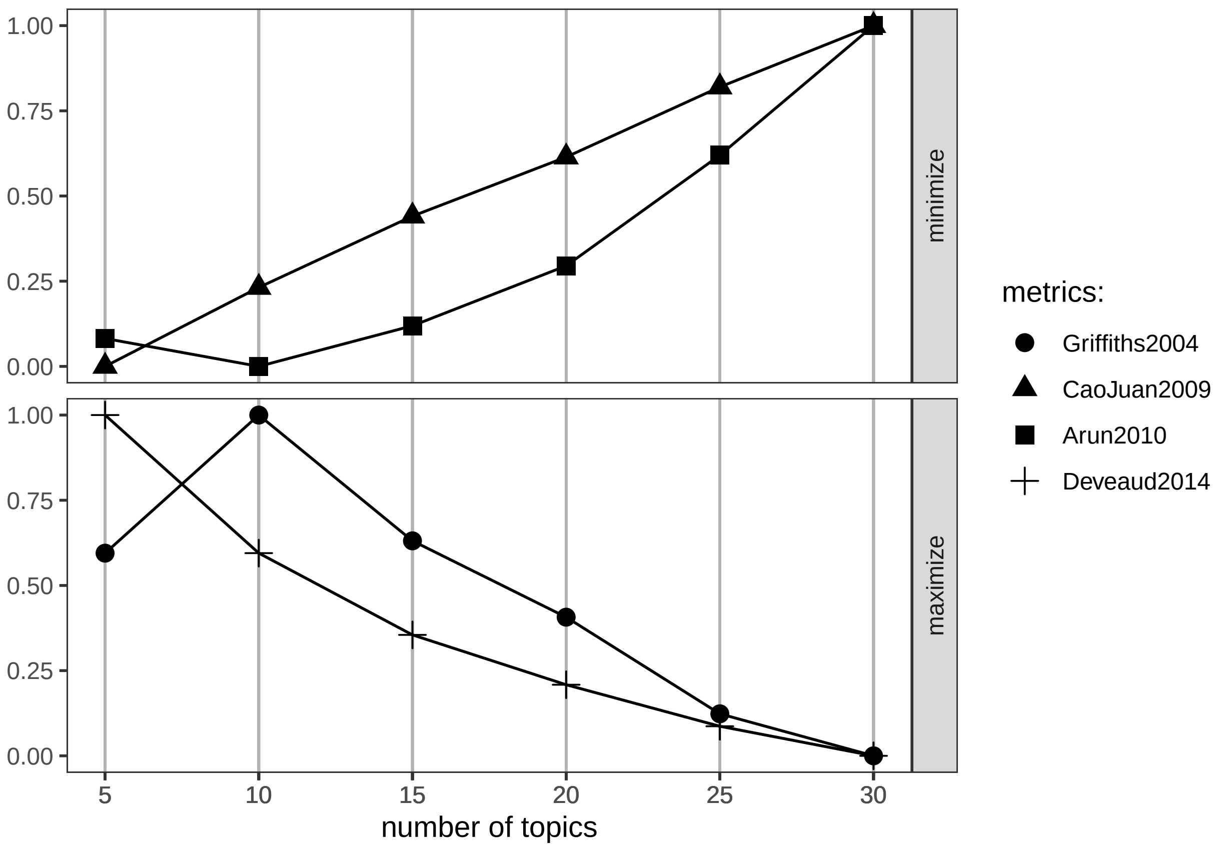
To find the ideal number of topics, I use both the ldatuning library and calculate perplexity.

Calculating and plotting perplexity for democratic corpus:

```
suppressMessages(library('ldatuning'))
result <- FindTopicsNumber(
  dtm,
  topics = seq(from = 5, to = 30, by = 5),
  metrics = c("Griffiths2004", "CaoJuan2009", "Arun2010", "Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 77),
  mc.cores = 2L,
  verbose = TRUE
)
```

```
## fit models... done.
## calculate metrics:
##   Griffiths2004... done.
##   CaoJuan2009... done.
##   Arun2010... done.
##   Deveaud2014... done.
```

```
FindTopicsNumber_plot(result)
```



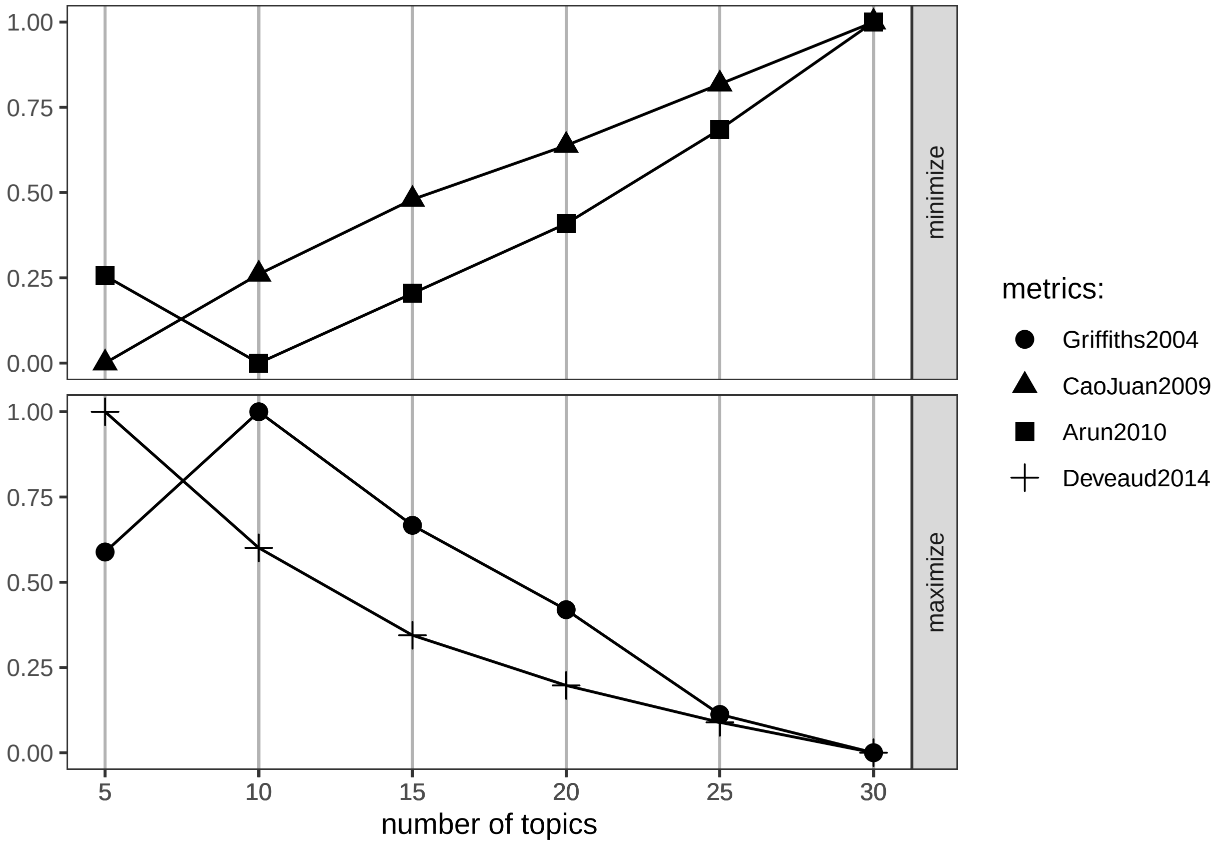
Calculating and plotting perplexity for republican corpus:

```
suppressMessages(library('ldatuning'))
result <- FindTopicsNumber(
```

```
  dtm2,
  topics = seq(from = 5, to = 30, by = 5),
  metrics = c("Griffiths2004", "CaoJuan2009", "Arun2010", "Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 77),
  mc.cores = 2L,
  verbose = TRUE
)
```

```
## fit models... done.
## calculate metrics:
##   Griffiths2004... done.
##   CaoJuan2009... done.
##   Arun2010... done.
##   Deveaud2014... done.
```

```
FindTopicsNumber_plot(result)
```



FindTopicsNumber_plot(result)

Now to calculate perplexity for k=5,10,15 for each party:

```
perplexity(LDA(dtm,5))
```

```
## [1] 1685.308
```

```
perplexity(LDA(dtm,10))
```

```
## [1] 1687.109
```

```r
perplexity(LDA(dtm,25))
```

```
## [1] 1691.685
```

```r
perplexity(LDA(dtm2,5))
```

```
## [1] 2372.86
```

```r
perplexity(LDA(dtm2,10))
```

```
## [1] 2374.779
```

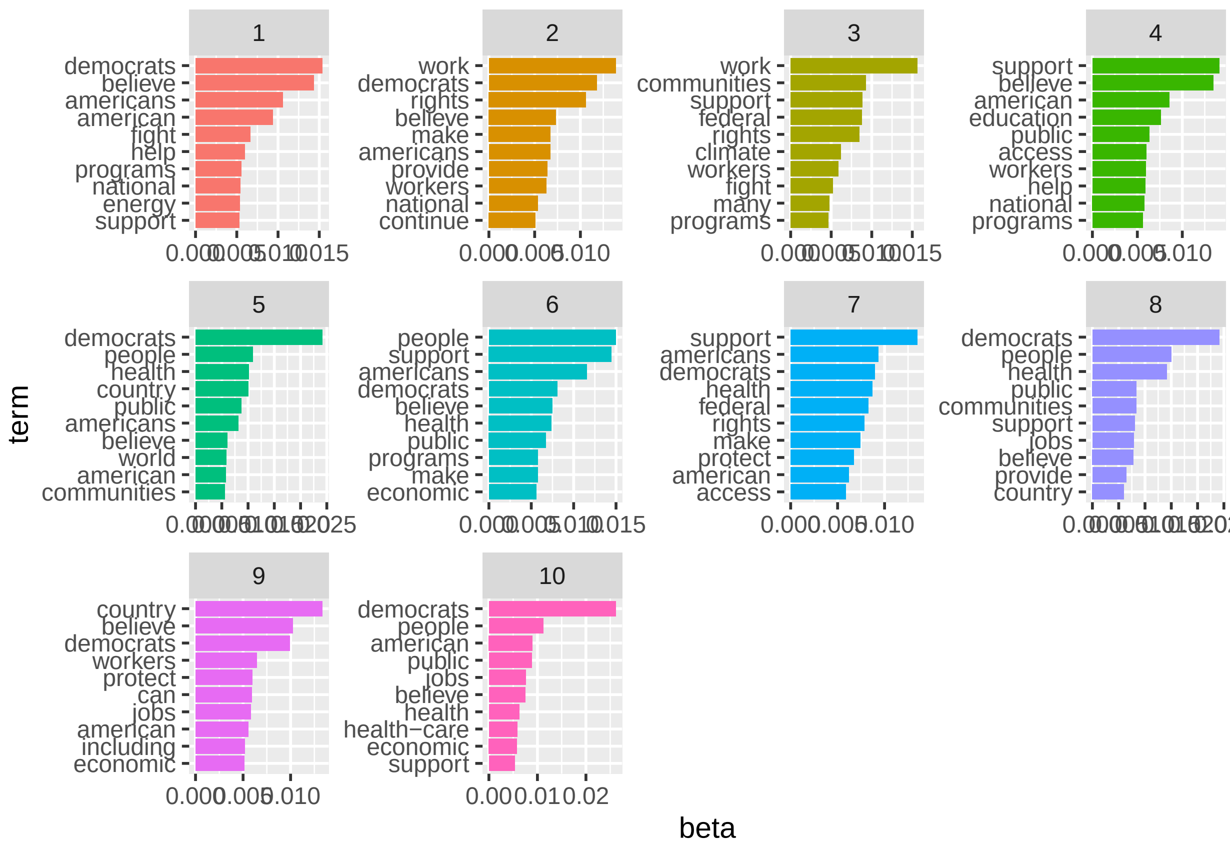
```r
perplexity(LDA(dtm2,25))
```

```
## [1] 2371.918
```

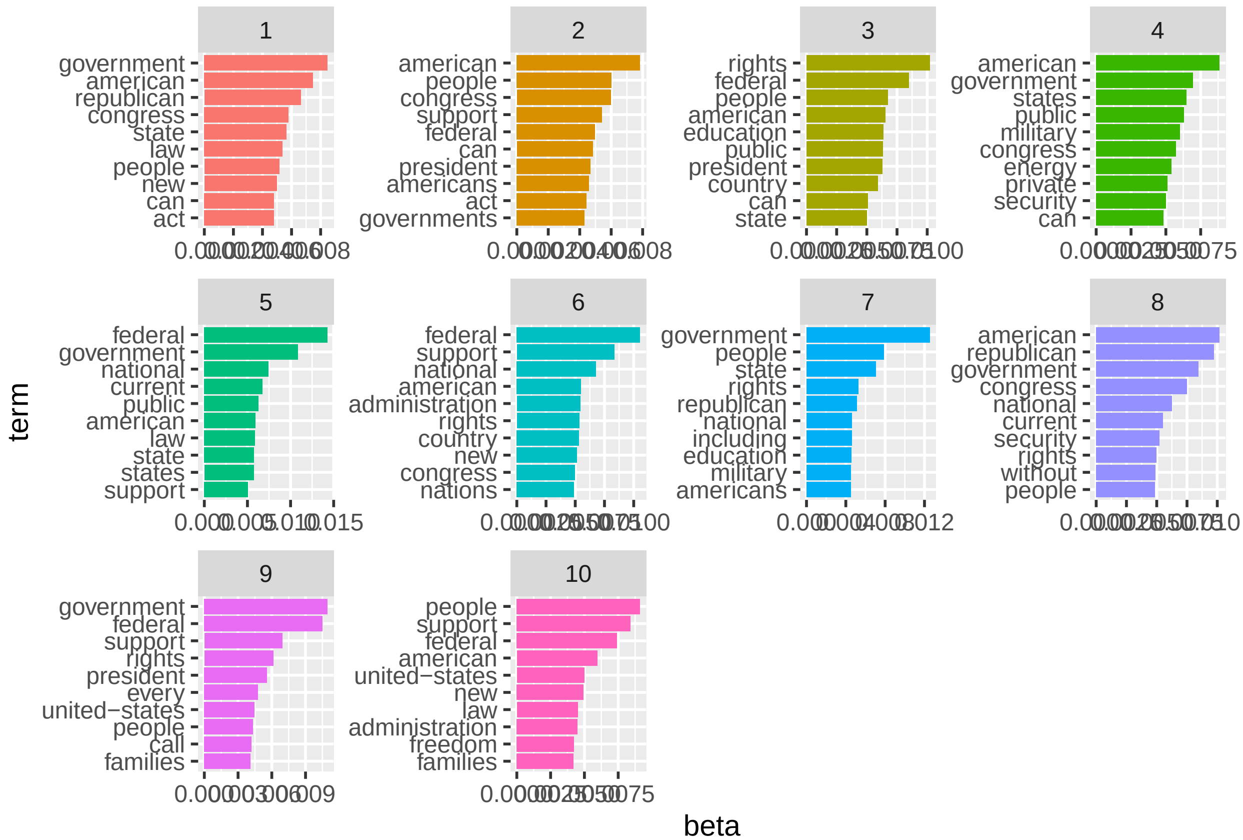
Therefore perplexity is minimized at k= 5 and other ldaturning measures are optimized at k=5 as well, suggesting k=5 is a good fit. It also needs to be mentioned, the lower perplexity of the democratic party's data might point to more coherence in the data generating process relative to the republicans.

10.

```r
#democratic corpus at k=10
plot_func(dtm,10)
```




```r
#republican corpus at k=10
plot_func(dtm2,10)
```



These are the bar plots for the two topic models at k=10. The topics that emerge for the democrats and the republicans seem to be quite different.

For republicans, states and state rights seem to be an recurrent theme across topics. Moreover there is no mention of health care, workers rights, however there is mention of (national) security and the military across topics. Even the topic with education, the word occurs with state, rights and state, suggesting they are talking about the role of states rights in education rather than improving schooling/education access.

For democrats, across topics, there is mention of workers, jobs, health (care), (national) programs (presumably proposed to achieve goals). There is talk of rights(of the kind I mentioned before) across topics, which is missing in the republican corpus, where "rights" is used in a very different way. There is talk of climate change in one of the topics, which is noteworthy since climate change is a pressing issue.

Overall, for both parties, I think k=10 is too many topics, because the topics seem diffuse and not self-contained and coherent. This is seen form the fact that most themes are spread across topics. It is possible that the same words are used in different contexts, but that is not what seems to be happening here.

11.

If I voted, I would vote democratic, because: They even talk about voter rights, civil rights, and civil justice. Additionally, the democrats talk about climate change much more frequently than the republicans. Being that climate change is one of the defining problems of the era, it would be perhaps irresponsible in the long run to vote for a party who doesn't even acknowledge/focus on climate change. Another important aspect of my choice is the fact that democrats about about health care as well as workers rights, two important facets of an ideal society.