Problem-set-5

Daejin Kim

11/27/2019

Problem Set 5: Text Mining

For this final problem set, suppose you are an undecided American voter. You are interested in exploring both major political parties in America and how they brand themselves and things they choose to focus on. As such, you get the most recent formalization of these things you can find, which are the 2016 party platforms for each of the two major parties (also called "manifestos" in other countries).

You decide to employ your computational skills to get explore that which these parties have to offer. At the end of the analysis, you must make a decision. I.e., answer the following question: *Based **only on your analysis here**, NOT your current political biases, which party would you support in 2020?*

Note: In this assignment, I am asking you to lay aside any feelings toward either party or political actor in America, whether good or bad. Rather, approach the question as unbiased as possible and try hard to let your results inform your response. Regardless of your current political proclivities, I promise I won't (care or) broadcast what your hypothetical response is in this problem set; its just a fun exercise in "data-drive decision making". Good luck!

PREPROCESSING & (light) EDA

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.

```
#setting a directory
directory<-setwd("/Users/daejin/Desktop/Fall Quarter/MACS 40800/MACS problem sets/Problem-Set-5")

#loading "d16.txt" and creating a corpus
democrat<-file.path(directory, "2d16.txt")
doc_d<-read.delim(file=democrat, header=TRUE)
d<-c(doc_d)
Dcorpus<-VCorpus(VectorSource(d))

#loading "r16.txt" and creating a corpus
republican<-file.path(directory, "3r16.txt")
doc_r<-read.delim(file=republican, header=TRUE)
r<-c(doc_r)
Rcorpus<-VCorpus(VectorSource(r))
```

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
```

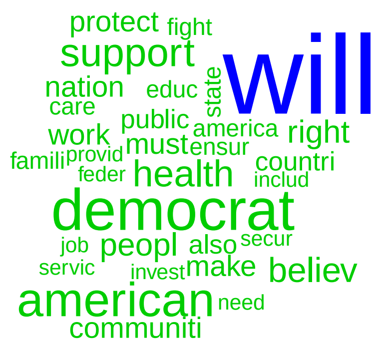
```
* Remove all punctuation
* Remove the whitespace
```

```r
#https://www.rdocumentation.org/packages/tm/versions/0.7-6/topics/Corpus
library(SnowballC)
#preprocessing Dcorpus, a corpus for Democratic party platform.
Dcorpus=tm_map(Dcorpus, content_transformer(tolower)) #changing it to lower case
Dcorpus = tm_map(Dcorpus, removeNumbers) #remove numbers
Dcorpus = tm_map(Dcorpus, removePunctuation) #remove punctuation
Dcorpus = tm_map(Dcorpus, removeWords, stopwords()) #remove stopwords
Dcorpus = tm_map(Dcorpus, stemDocument) #stem words
Dcorpus = tm_map(Dcorpus, stripWhitespace) #removing the whitespace
#Creating a document-term matrix
Ddtm = DocumentTermMatrix(Dcorpus)
Ddtm = removeSparseTerms(Ddtm, 0.999)


#Preprocessing Rcorpus, a corpus for Republican party platform.
Rcorpus=tm_map(Rcorpus, content_transformer(tolower))
Rcorpus = tm_map(Rcorpus, removeNumbers)
Rcorpus = tm_map(Rcorpus, removePunctuation)
Rcorpus = tm_map(Rcorpus, removeWords, stopwords())
Rcorpus = tm_map(Rcorpus, stemDocument)
Rcorpus = tm_map(Rcorpus, stripWhitespace)
#Creating a document-term matrix
Rdtm = DocumentTermMatrix(Rcorpus)
Rdtm = removeSparseTerms(Rdtm, 0.999)
```

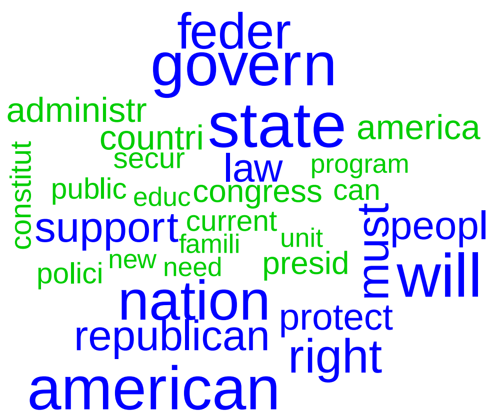
3. Visually inspect your cleaned documents by creating a wordcloud for each major party's platform. Based on this naive visualization, offer a few-sentence-description of general patterns you see (e.g., What are commonly used words? What are less commonly used words? Can you get a sense of differences between the parties at this early stage?

Both parties demonstrate common words, such as "american", "nation", "right", "must", "protect", "peopl", support","countri, which are not surprising to see in political party platform. However, the Democratic platform shows a high frequency of "work", "communiti", and "health" compared to the Republican platform. These words may suggest that the Democratic party focuses on how to "work" to improve the "health" of "communities." On the other hand, the Republican party platform mostly uses political terms including "govern", "state", "feder(al)", "law", "congress," which are not meant for specific policy or agenda.

```r
library(wordcloud)
#Wordcloud for Democratic party platform
Ddataset = as.matrix(Ddtm)
Dv = sort(colSums(Ddataset),decreasing=TRUE)
DmyNames = names(Dv)
Dd = data.frame(word=DmyNames,freq=Dv)
set.seed(123)
wordcloud(Dd$word, colors=c(3,4),random.color=FALSE, Dd$freq, min.freq=60, scale=c(3.5,0.25))
```




```
#Wordcloud for Republican party platform
Rdataset = as.matrix(Rdtm)
Rv = sort(colSums(Rdataset),decreasing=TRUE)
RmyNames = names(Rv)
Rd = data.frame(word=RmyNames,freq=Rv)
set.seed(123)
wordcloud(Rd$word, colors=c(3,4),random.color=FALSE, Rd$freq, min.freq=60, scale=c(2,0.25))
```



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).

```r
#https://cran.r-project.org/web/packages/tidytext/vignettes/tidying_casting.html
library(tidytext)
library(tidyverse)
library(glue)
```

```
##
## Attaching package: 'glue'

## The following object is masked from 'package:dplyr':
##
##     collapse
```

```r
library(stringr)
library(dplyr)
library(tidytext)
library(textdata)

#Sentiment for Democratic party platform
```

```r
td_Ddtm <- tidy(Ddtm)
D_bing_sentiment <- td_Ddtm %>%
  inner_join(get_sentiments("bing"), by = c(term = "word"))
D_afinn_sentiment <- td_Ddtm %>%
  inner_join(get_sentiments("afinn"), by = c(term = "word"))
D_bing_sentiment
```

```
## # A tibble: 258 x 4
##    document term       count sentiment
##    <chr>    <chr>      <dbl> <chr>
##  1 1        abolish        2 negative
##  2 1        abort          5 negative
##  3 1        accomplish     1 positive
##  4 1        addict         7 negative
##  5 1        affirm         2 positive
##  6 1        afford        30 positive
##  7 1        ail            1 negative
##  8 1        ailment        1 negative
##  9 1        appeal         2 positive
## 10 1        applaud        5 positive
## # ... with 248 more rows
```

```r
D_afinn_sentiment
```

```
## # A tibble: 204 x 4
##    document term       count value
##    <chr>    <chr>      <dbl> <dbl>
##  1 1        abandon        6    -2
##  2 1        accept         1     1
##  3 1        accomplish     1     2
##  4 1        adopt          3     1
##  5 1        agreement     10     1
##  6 1        allow         13     1
##  7 1        applaud        5     2
##  8 1        arrest         2    -2
##  9 1        asset          3     2
## 10 1        attack        10    -1
## # ... with 194 more rows
```

```r
D_afinn_sum<-sum(D_afinn_sentiment$value)

#Getting the sum of the sentiment of Democratic party platform
D_bing_sentiment$sentiment [D_bing_sentiment$sentiment == "positive"] <- 1
D_bing_sentiment$sentiment [D_bing_sentiment$sentiment == "negative"] <- -1
D_bing_sum<-as.numeric(D_bing_sentiment$sentiment)%>%
  sum()


#Sentiment for Republican party platforom
td_Rdtm <- tidy(Rdtm)
R_bing_sentiment <- td_Rdtm %>%
  inner_join(get_sentiments("bing"), by = c(term = "word"))
```

```r
R_afinn_sentiment <- td_Rdtm %>%
  inner_join(get_sentiments("afinn"), by = c(term = "word"))
R_bing_sentiment
```

```
## # A tibble: 360 x 4
##    document term       count sentiment
##    <chr>    <chr>      <dbl> <chr>
##  1 1        abolish        4 negative
##  2 1        abort         33 negative
##  3 1        abrupt         2 negative
##  4 1        absurd         2 negative
##  5 1        accomplish     3 positive
##  6 1        addict         2 negative
##  7 1        affirm        26 positive
##  8 1        afflict        2 negative
##  9 1        afford        10 positive
## 10 1        affront        2 negative
## # ... with 350 more rows
```

```r
R_afinn_sentiment
```

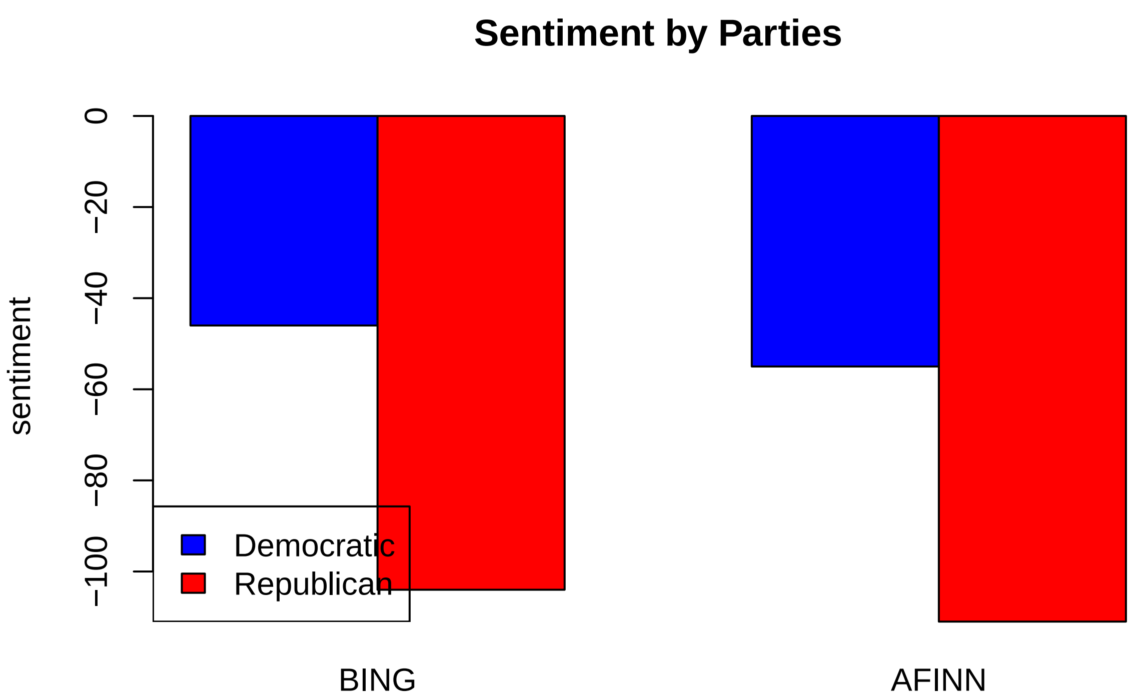
```
## # A tibble: 252 x 4
##    document term       count value
##    <chr>    <chr>      <dbl> <dbl>
##  1 1        abandon        6    -2
##  2 1        abhor          1    -3
##  3 1        accept        10     1
##  4 1        accomplish     3     2
##  5 1        admit          3    -1
##  6 1        adopt         11     1
##  7 1        agreement     26     1
##  8 1        alarm          1    -2
##  9 1        allow         37     1
## 10 1        anger          2    -3
## # ... with 242 more rows
```

```r
R_afinn_sum<-sum(R_afinn_sentiment$value)

#Getting the sum of the sentiment of Democratic party platform
R_bing_sentiment$sentiment [R_bing_sentiment$sentiment == "positive"] <- 1
R_bing_sentiment$sentiment [R_bing_sentiment$sentiment == "negative"] <- -1
R_bing_sum<-as.numeric(R_bing_sentiment$sentiment)%>%
  sum()

#Comparing sentiments between parties
a<-c(D_bing_sum,R_bing_sum,D_afinn_sum,R_afinn_sum)
a<-matrix(a, byrow=FALSE, ncol=2)
rownames(a)<-c("Democratic","Republican")
colnames(a)<-c("BING","AFINN")
a
```

```
##            BING AFINN
```

```
## Democratic  -46   -55
## Republican -104  -111
```

```r
barplot(a,
main = "Sentiment by Parties",
ylab = "sentiment",
col = c("blue","red"), beside=TRUE
)
legend("bottomleft",
c("Democratic","Republican"),
fill = c("blue","red")
)
```



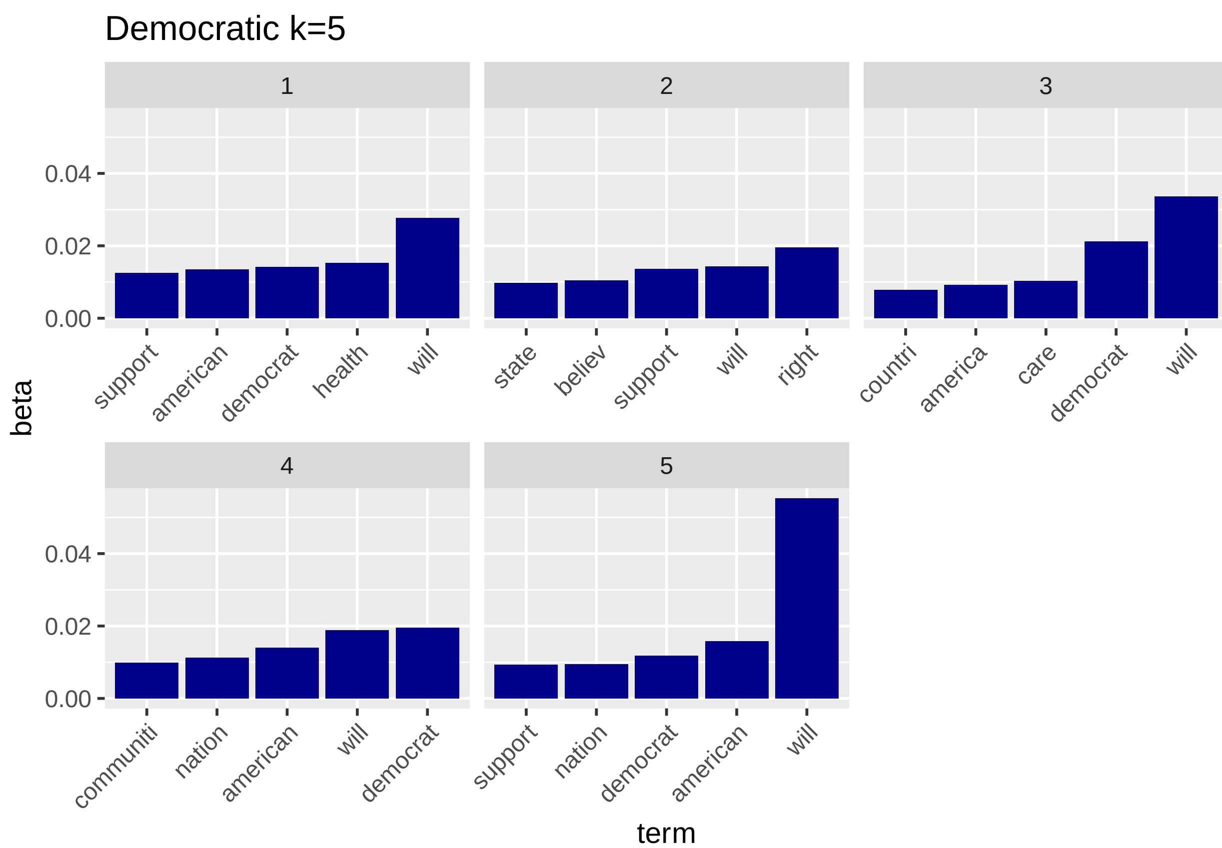
5. Compare and discuss the sentiments of these platforms: which party tends to be more optimistic about the future? Does this comport with your perceptions of the parties?

The sentiments from both "bing" and "afinn" dictionaries indicate that the Democratic party tends to be more optimistic or less pessimistic about the future. The Democratic party uses about half less negative words than the other party. It is surprising to see this as the Republican party or any conservative party is known to use more negative language.

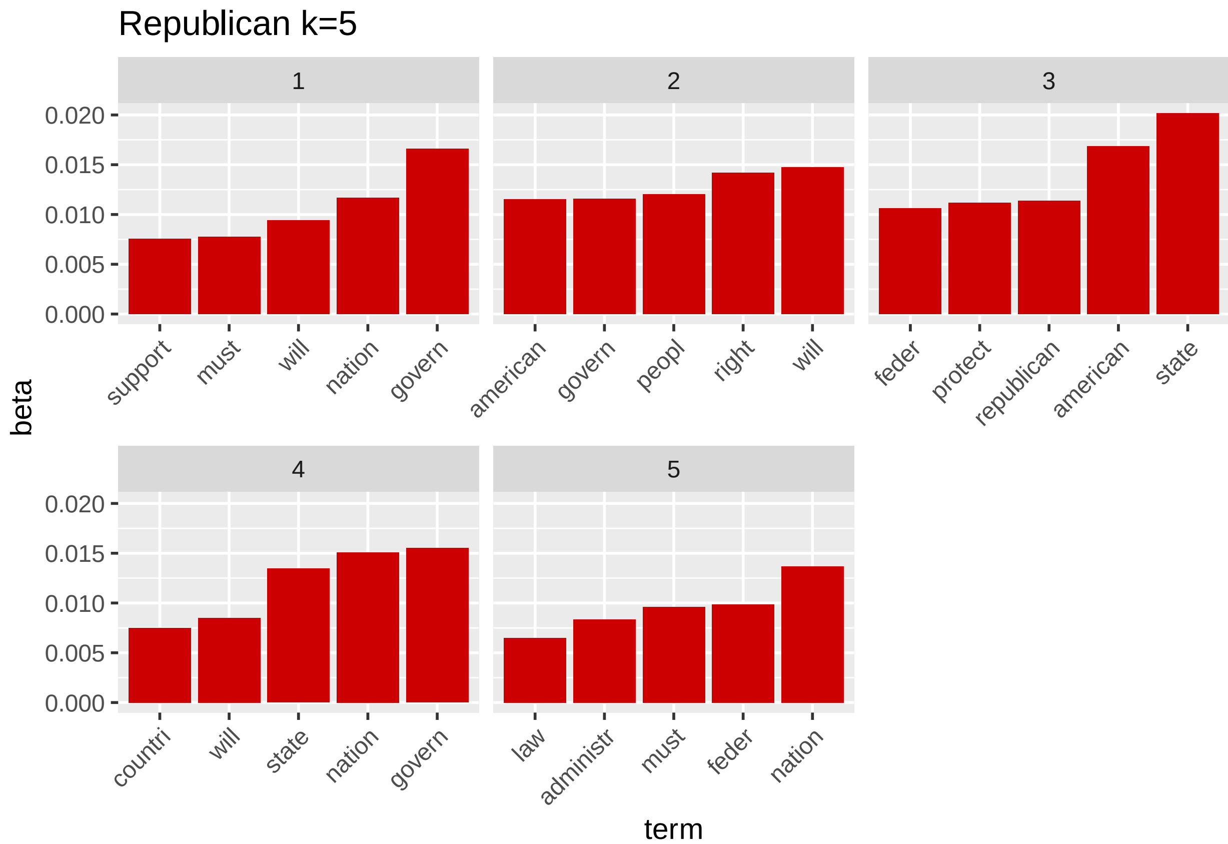
TOPIC MODELS

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).

```r
#https://cran.r-project.org/web/packages/tidytext/vignettes/topic_modeling.html
library(topicmodels)
#Fitting a topic model for Democratic
Ddtm_lda5 <- LDA(Ddtm, k = 5, control = list(seed = 123))
Ddtm_lda5_td <- tidy(Ddtm_lda5)
Dtop_terms <- Ddtm_lda5_td %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
Dtop_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta)) +
  geom_bar(stat = "identity", fill="blue4") +
  scale_x_reordered() +
  facet_wrap(~ topic, scales = "free_x")+
  ggtitle("Democratic k=5") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
#Fitting a topic model for Republican
Rdtm_lda5 <- LDA(Rdtm, k = 5, control = list(seed = 1234))
Rdtm_lda5_td <- tidy(Rdtm_lda5)
Rtop_terms <- Rdtm_lda5_td %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
Rtop_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta)) +
  geom_bar(stat = "identity", fill="red3") +
  scale_x_reordered() +
  facet_wrap(~ topic, scales = "free_x")+
  ggtitle("Republican k=5") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



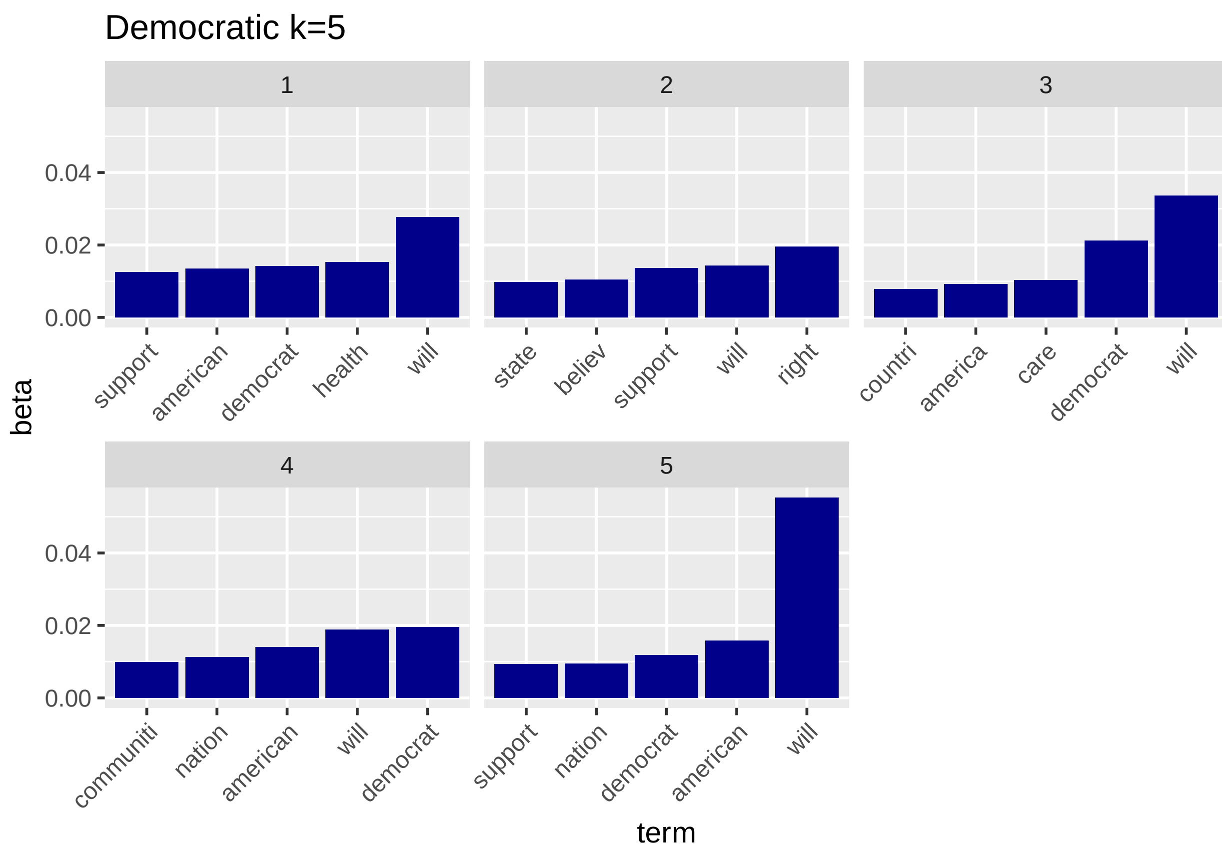
7. Describe the general trends in topics that emerge from this stage. Are the parties focusing on similar or different topics, generally?

All the topics regardless of the party capture similar words. It seems that both parties talk about how the state or nation should support and protect American people. Therefore, the topic models with k=5 only show some general topics across parties.

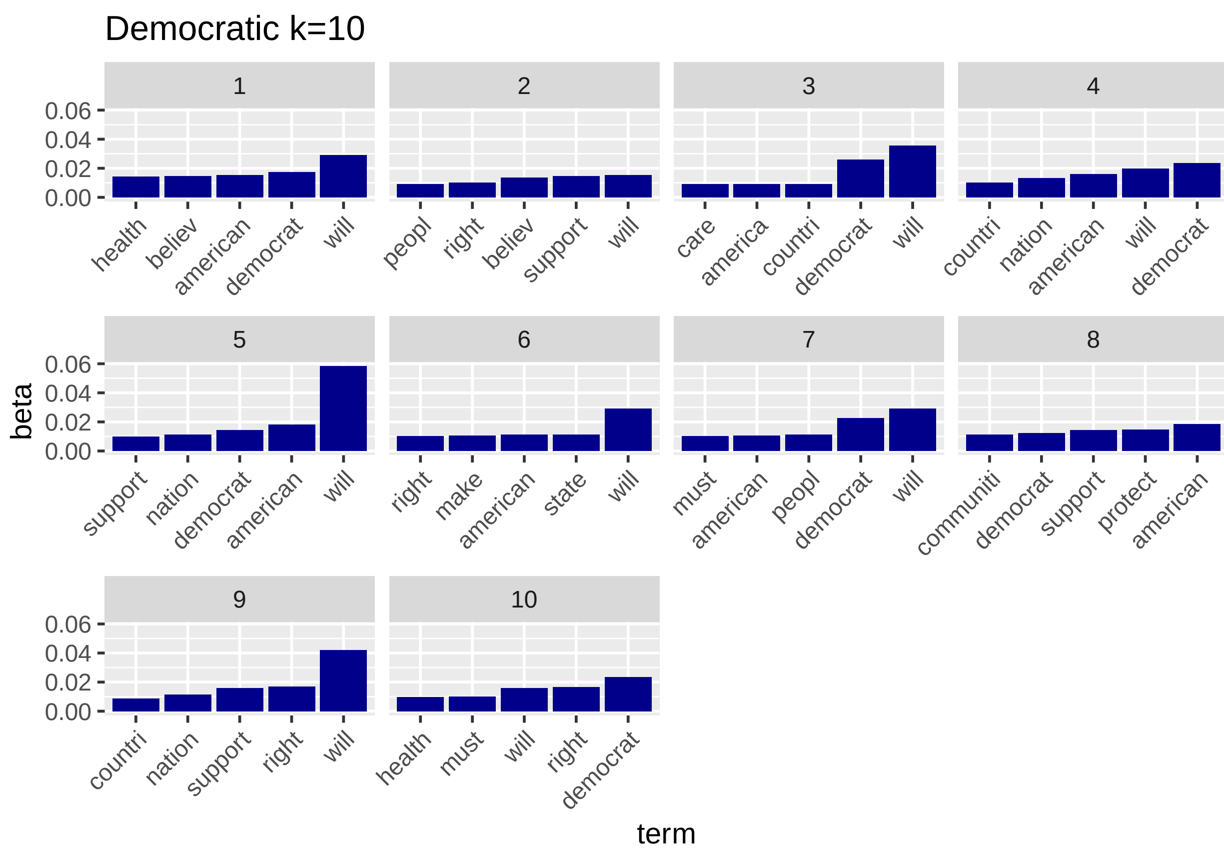
8. Fit 6 more topic models at the follow levels of `k` for each party: `5, 10, 25`. Present the results however you'd like (e.g., visually and/or numerically).

```
#Democratic, k=5
Ddtm_lda5 <- LDA(Ddtm, k = 5, control = list(seed = 123))
Ddtm_lda5_td <- tidy(Ddtm_lda5)
Dtop_terms <- Ddtm_lda5_td %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
Dtop_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta)) +
  geom_bar(stat = "identity", fill="blue4") +
  scale_x_reordered() +
```

```
  facet_wrap(~ topic, scales = "free_x")+
  ggtitle("Democratic k=5") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
#Democratic, k=10
Ddtm_lda10 <- LDA(Ddtm, k = 10, control = list(seed = 123))
Ddtm_lda10_td <- tidy(Ddtm_lda10)
Dtop_terms <- Ddtm_lda10_td %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
Dtop_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta)) +
  geom_bar(stat = "identity", fill="blue4") +
  scale_x_reordered() +
  facet_wrap(~ topic, scales = "free_x")+
  ggtitle("Democratic k=10") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
#Democratic, k=25
Ddtm_lda25 <- LDA(Ddtm, k = 25, control = list(seed = 123))
Ddtm_lda25_td <- tidy(Ddtm_lda25)
Dtop_terms <- Ddtm_lda25_td %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
Dtop_terms %>%
  mutate(term = reorder_within(term, beta, topic))
```

```
## # A tibble: 125 x 3
##    topic term            beta
##    <int> <fct>          <dbl>
##  1     1 will___1      0.0245
##  2     1 democrat___1  0.0154
##  3     1 health___1    0.0136
##  4     1 american___1  0.0133
##  5     1 believ___1    0.0117
##  6     2 will___2      0.0129
##  7     2 support___2   0.0115
##  8     2 believ___2    0.0108
##  9     2 peopl___2     0.00901
## 10     2 must___2      0.00836
## # ... with 115 more rows
```

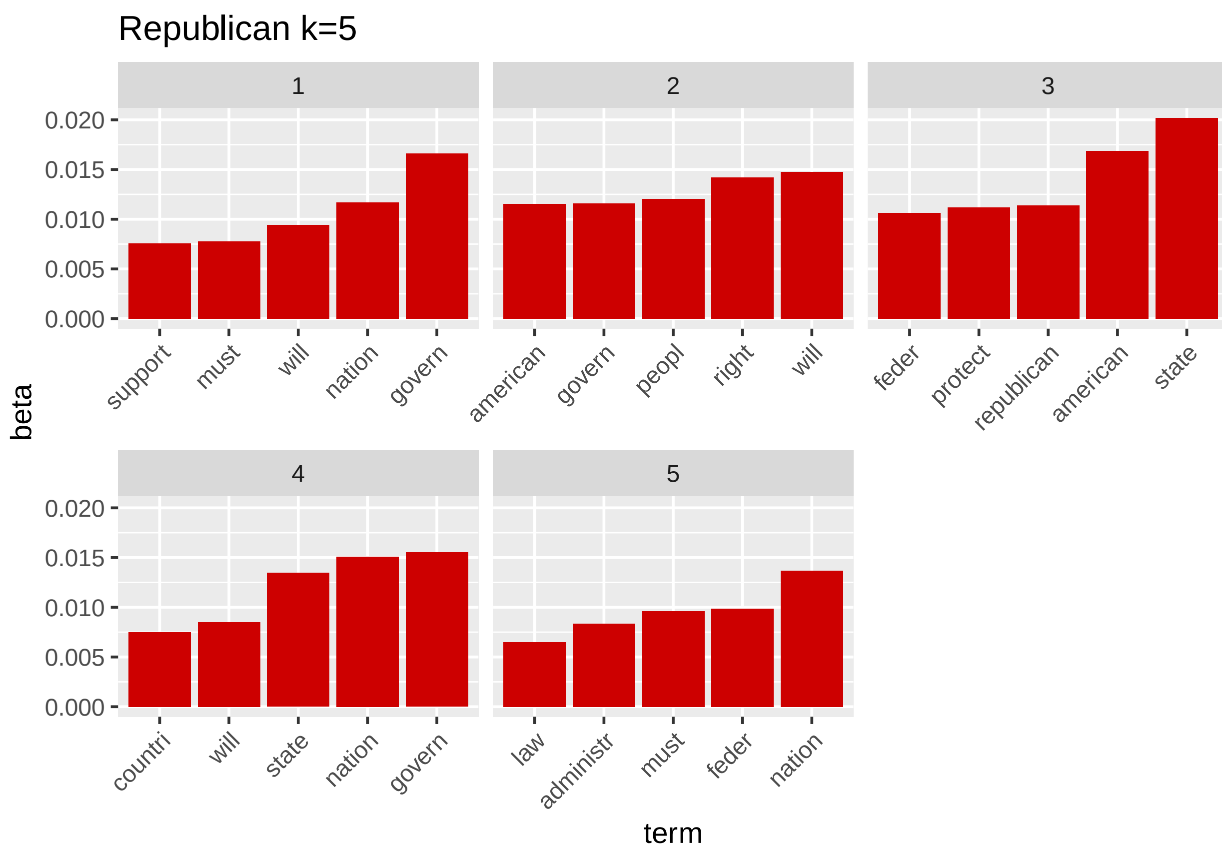
```r
print(as_tibble(Dtop_terms), n = 125)
```

```
## # A tibble: 125 x 3
##     topic term         beta
##     <int> <chr>       <dbl>
##   1     1 will      0.0245
##   2     1 democrat  0.0154
##   3     1 health    0.0136
##   4     1 american  0.0133
##   5     1 believ    0.0117
##   6     2 will      0.0129
##   7     2 support   0.0115
##   8     2 believ    0.0108
##   9     2 peopl     0.00901
##  10     2 must      0.00836
##  11     3 will      0.0299
##  12     3 democrat  0.0232
##  13     3 america   0.00912
##  14     3 also      0.00820
##  15     3 countri   0.00778
##  16     4 democrat  0.0213
##  17     4 will      0.0167
##  18     4 american  0.0140
##  19     4 nation    0.0126
##  20     4 communiti 0.0101
##  21     5 will      0.0494
##  22     5 american  0.0160
##  23     5 democrat  0.0132
##  24     5 nation    0.0107
##  25     5 support   0.00777
##  26     6 will      0.0244
##  27     6 communiti 0.0111
##  28     6 make      0.0102
##  29     6 american  0.00968
##  30     6 peopl     0.00944
##  31     7 will      0.0243
##  32     7 democrat  0.0201
##  33     7 peopl     0.0109
##  34     7 must      0.00967
##  35     7 american  0.00924
##  36     8 american  0.0162
##  37     8 communiti 0.0131
##  38     8 support   0.0113
##  39     8 democrat  0.0112
##  40     8 work      0.00978
##  41     9 will      0.0354
##  42     9 right     0.0137
##  43     9 support   0.0124
##  44     9 nation    0.0109
##  45     9 communiti 0.00876
##  46    10 democrat  0.0211
##  47    10 right     0.0135
##  48    10 will      0.0134
```

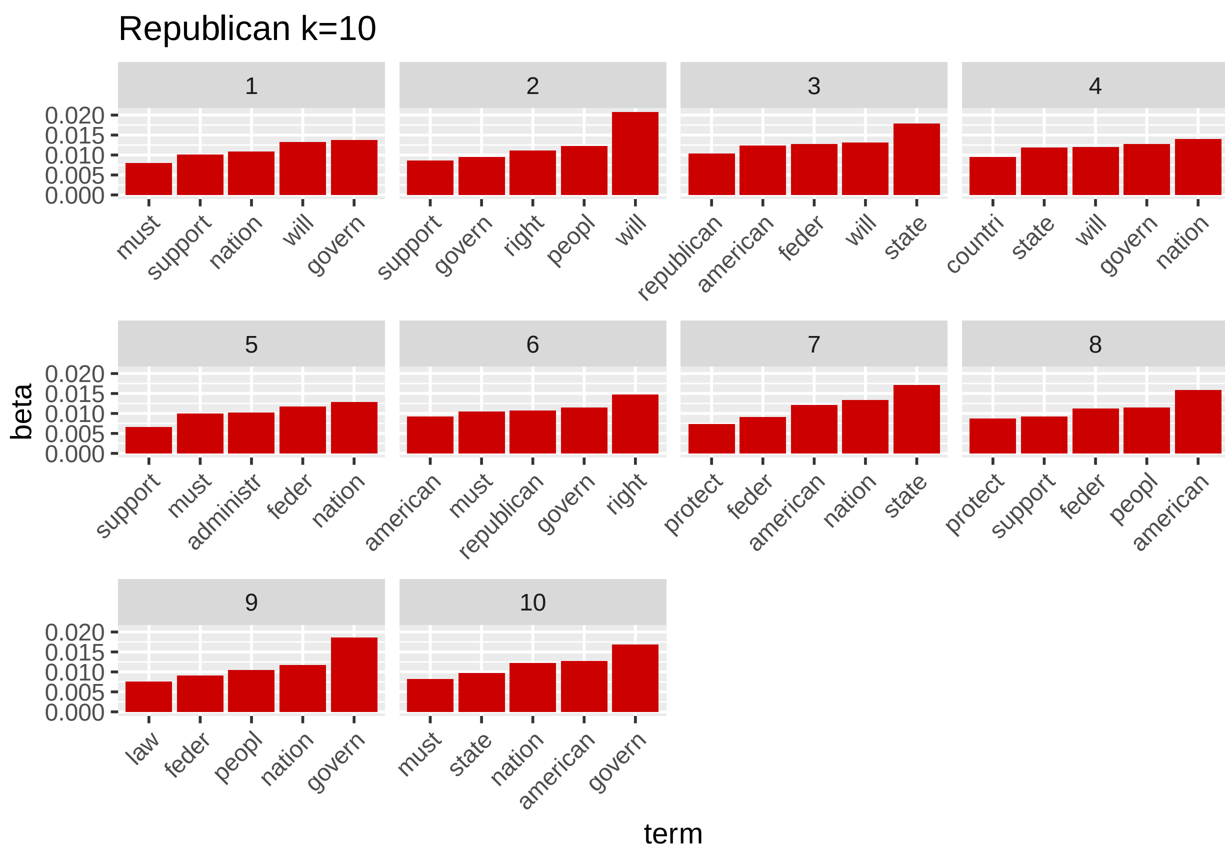
```
## 49     10 must      0.00951
## 50     10 health    0.00934
## 51     11 will      0.0296
## 52     11 democrat  0.0168
## 53     11 health    0.0124
## 54     11 america   0.0107
## 55     11 make      0.0102
## 56     12 will      0.0429
## 57     12 health    0.0158
## 58     12 believ    0.0112
## 59     12 democrat  0.0104
## 60     12 job       0.00881
## 61     13 will      0.0168
## 62     13 communiti 0.0136
## 63     13 right     0.0131
## 64     13 support   0.0120
## 65     13 america   0.0108
## 66     14 will      0.0502
## 67     14 american  0.0176
## 68     14 democrat  0.0136
## 69     14 communiti 0.0113
## 70     14 support   0.0106
## 71     15 will      0.0333
## 72     15 democrat  0.0219
## 73     15 believ    0.0119
## 74     15 right     0.0113
## 75     15 make      0.0108
## 76     16 health    0.0162
## 77     16 will      0.0139
## 78     16 right     0.0136
## 79     16 communiti 0.0132
## 80     16 american  0.0128
## 81     17 democrat  0.0202
## 82     17 believ    0.0135
## 83     17 will      0.0129
## 84     17 nation    0.0116
## 85     17 right     0.0112
## 86     18 will      0.0414
## 87     18 right     0.0132
## 88     18 democrat  0.0108
## 89     18 peopl     0.0104
## 90     18 american  0.00926
## 91     19 will      0.0475
## 92     19 american  0.0151
## 93     19 believ    0.0120
## 94     19 support   0.0101
## 95     19 care      0.00934
## 96     20 american  0.0142
## 97     20 right     0.0119
## 98     20 communiti 0.0105
## 99     20 health    0.00990
## 100    20 care      0.00902
## 101    21 democrat  0.0154
## 102    21 support   0.0118
```

```
## 103    21 believ    0.0116
## 104    21 peopl     0.0116
## 105    21 american  0.00972
## 106    22 will      0.0260
## 107    22 american  0.0151
## 108    22 support   0.00970
## 109    22 work      0.00820
## 110    22 ensur     0.00812
## 111    23 will      0.0469
## 112    23 democrat  0.0239
## 113    23 health    0.0132
## 114    23 support   0.0116
## 115    23 american  0.0113
## 116    24 american  0.0124
## 117    24 will      0.0119
## 118    24 believ    0.0114
## 119    24 america   0.0106
## 120    24 make      0.0102
## 121    25 will      0.0193
## 122    25 american  0.0186
## 123    25 believ    0.0118
## 124    25 support   0.0111
## 125    25 democrat  0.00994
```

```r
#Republican, k=5
Rdtm_lda5 <- LDA(Rdtm, k = 5, control = list(seed = 1234))
Rdtm_lda5_td <- tidy(Rdtm_lda5)
Rtop_terms <- Rdtm_lda5_td %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
Rtop_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta)) +
  geom_bar(stat = "identity", fill="red3") +
  scale_x_reordered() +
  facet_wrap(~ topic, scales = "free_x")+
  ggtitle("Republican k=5") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
#Republican, k=10
Rdtm_lda10 <- LDA(Rdtm, k = 10, control = list(seed = 1234))
Rdtm_lda10_td <- tidy(Rdtm_lda10)
Rtop_terms <- Rdtm_lda10_td %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
Rtop_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta)) +
  geom_bar(stat = "identity", fill="red3") +
  scale_x_reordered() +
  facet_wrap(~ topic, scales = "free_x")+
  ggtitle("Republican k=10") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
#Remocratic, k=25
Rdtm_lda25 <- LDA(Rdtm, k = 25, control = list(seed = 1234))
Rdtm_lda25_td <- tidy(Rdtm_lda25)
Rtop_terms <- Rdtm_lda25_td %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
Rtop_terms %>%
  mutate(term = reorder_within(term, beta, topic))
```

```
## # A tibble: 125 x 3
##    topic term          beta
##    <int> <fct>        <dbl>
##  1     1 govern___1  0.0129
##  2     1 support___1 0.0117
##  3     1 nation___1  0.0107
##  4     1 will___1    0.00971
##  5     1 must___1    0.00850
##  6     2 will___2    0.0152
##  7     2 peopl___2   0.0105
##  8     2 right___2   0.0104
##  9     2 support___2 0.0101
## 10     2 govern___2  0.00900
## # ... with 115 more rows
```

```r
print(as_tibble(Rtop_terms), n = 125)
```

```
## # A tibble: 125 x 3
##     topic term          beta
##     <int> <chr>        <dbl>
##   1     1 govern     0.0129
##   2     1 support    0.0117
##   3     1 nation     0.0107
##   4     1 will       0.00971
##   5     1 must       0.00850
##   6     2 will       0.0152
##   7     2 peopl      0.0105
##   8     2 right      0.0104
##   9     2 support    0.0101
##  10     2 govern     0.00900
##  11     3 state      0.0168
##  12     3 feder      0.0133
##  13     3 republican 0.0124
##  14     3 american   0.0105
##  15     3 will       0.00961
##  16     4 nation     0.0138
##  17     4 govern     0.0121
##  18     4 state      0.0112
##  19     4 support    0.00886
##  20     4 will       0.00874
##  21     5 nation     0.0125
##  22     5 feder      0.0122
##  23     5 must       0.0105
##  24     5 administr  0.0101
##  25     5 support    0.00763
##  26     6 right      0.0137
##  27     6 republican 0.0128
##  28     6 must       0.0111
##  29     6 govern     0.0108
##  30     6 american   0.00777
##  31     7 state      0.0160
##  32     7 nation     0.0130
##  33     7 american   0.0103
##  34     7 feder      0.00950
##  35     7 republican 0.00752
##  36     8 american   0.0137
##  37     8 feder      0.0119
##  38     8 support    0.0107
##  39     8 peopl      0.0100
##  40     8 state      0.00825
##  41     9 govern     0.0176
##  42     9 nation     0.0114
##  43     9 feder      0.00954
##  44     9 peopl      0.00911
##  45     9 law        0.00836
##  46    10 govern     0.0159
##  47    10 nation     0.0120
##  48    10 american   0.0108
```

```
## 49    10 state      0.00907
## 50    10 must       0.00871
## 51    11 will       0.0153
## 52    11 state      0.0140
## 53    11 nation     0.00929
## 54    11 peopl      0.00905
## 55    11 support    0.00880
## 56    12 support    0.0108
## 57    12 nation     0.0107
## 58    12 peopl      0.0106
## 59    12 administr  0.00854
## 60    12 govern     0.00797
## 61    13 right      0.0140
## 62    13 nation     0.0136
## 63    13 state      0.0132
## 64    13 american   0.0103
## 65    13 peopl      0.00843
## 66    14 american   0.0164
## 67    14 right      0.0147
## 68    14 govern     0.0142
## 69    14 must       0.0140
## 70    14 republican 0.0126
## 71    15 must       0.0126
## 72    15 administr  0.00899
## 73    15 nation     0.00722
## 74    15 will       0.00660
## 75    15 state      0.00653
## 76    16 american   0.0135
## 77    16 will       0.0133
## 78    16 must       0.0119
## 79    16 law        0.00844
## 80    16 administr  0.00775
## 81    17 govern     0.0131
## 82    17 nation     0.0128
## 83    17 state      0.0126
## 84    17 right      0.0108
## 85    17 presid     0.00784
## 86    18 will       0.0159
## 87    18 peopl      0.0108
## 88    18 nation     0.00950
## 89    18 govern     0.00923
## 90    18 current    0.00670
## 91    19 govern     0.0148
## 92    19 american   0.0112
## 93    19 law        0.00984
## 94    19 protect    0.00739
## 95    19 nation     0.00632
## 96    20 will       0.0155
## 97    20 american   0.0133
## 98    20 state      0.0122
## 99    20 feder      0.00949
## 100   20 peopl      0.00933
## 101   21 govern     0.0159
## 102   21 nation     0.0131
```

```
## 103    21 will       0.0105
## 104    21 law        0.00773
## 105    21 administr  0.00750
## 106    22 state      0.0149
## 107    22 must       0.0141
## 108    22 govern     0.0122
## 109    22 will       0.0115
## 110    22 american   0.0101
## 111    23 right      0.0120
## 112    23 peopl      0.0104
## 113    23 american   0.00923
## 114    23 support    0.00840
## 115    23 law        0.00714
## 116    24 will       0.0165
## 117    24 american   0.0138
## 118    24 feder      0.0130
## 119    24 right      0.0109
## 120    24 must       0.00873
## 121    25 state      0.0139
## 122    25 feder      0.0135
## 123    25 nation     0.0117
## 124    25 govern     0.00971
## 125    25 administr  0.00924
```

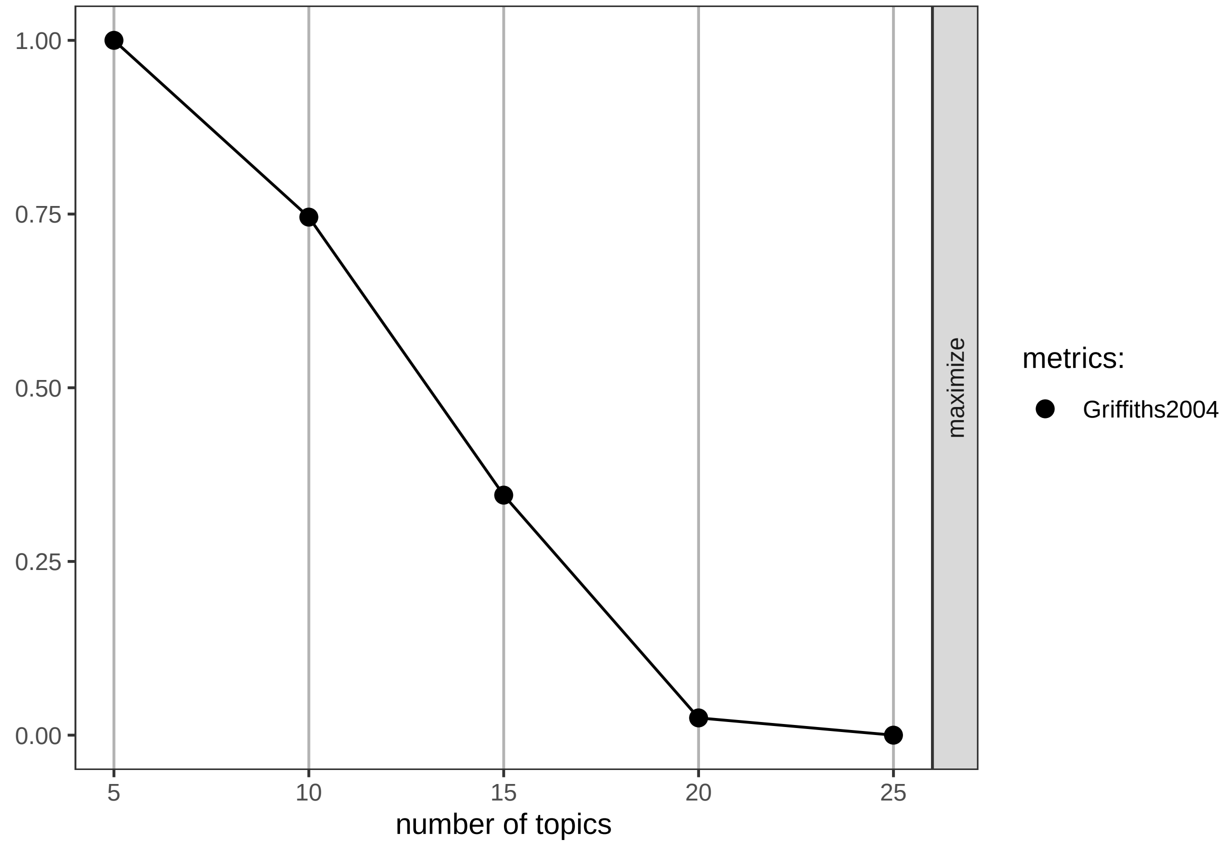
9. Calculate the perplexity of each model iteration and describe which technically fits best.

The perplexity indicates that the model would fit the best with 25 topics for both Democratic and Republican platforms. As shown in the graphs below, the perplexity has the smallest value when the number of the topic is 25.

```r
library(topicmodels)
library(ldatuning)
Dresult <- FindTopicsNumber(
  Ddtm,
  topics = seq(from = 5, to = 25, by = 5),
  metrics = "Griffiths2004",
  method = "Gibbs",
  control = list(seed = 77),
  mc.cores = 2L,
  verbose = TRUE
)
```

```
## fit models... done.
## calculate metrics:
##   Griffiths2004... done.
```

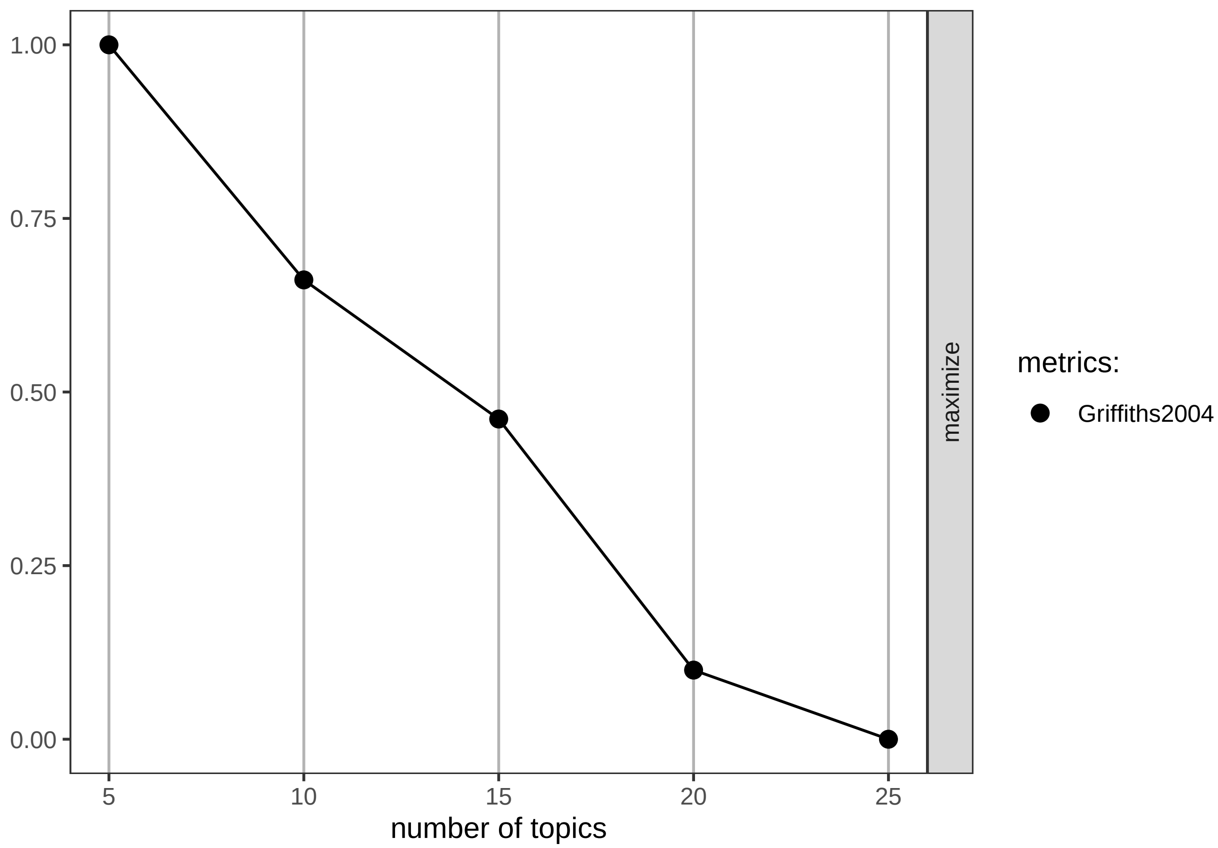
```r
FindTopicsNumber_plot(Dresult)
```




```r
Rresult <- FindTopicsNumber(
  Rdtm,
  topics = seq(from = 5, to = 25, by = 5),
  metrics = "Griffiths2004",
  method = "Gibbs",
  control = list(seed = 77),
  mc.cores = 2L,
  verbose = TRUE
)
```

```
## fit models... done.
## calculate metrics:
##   Griffiths2004... done.
```

```r
FindTopicsNumber_plot(Rresult)
```

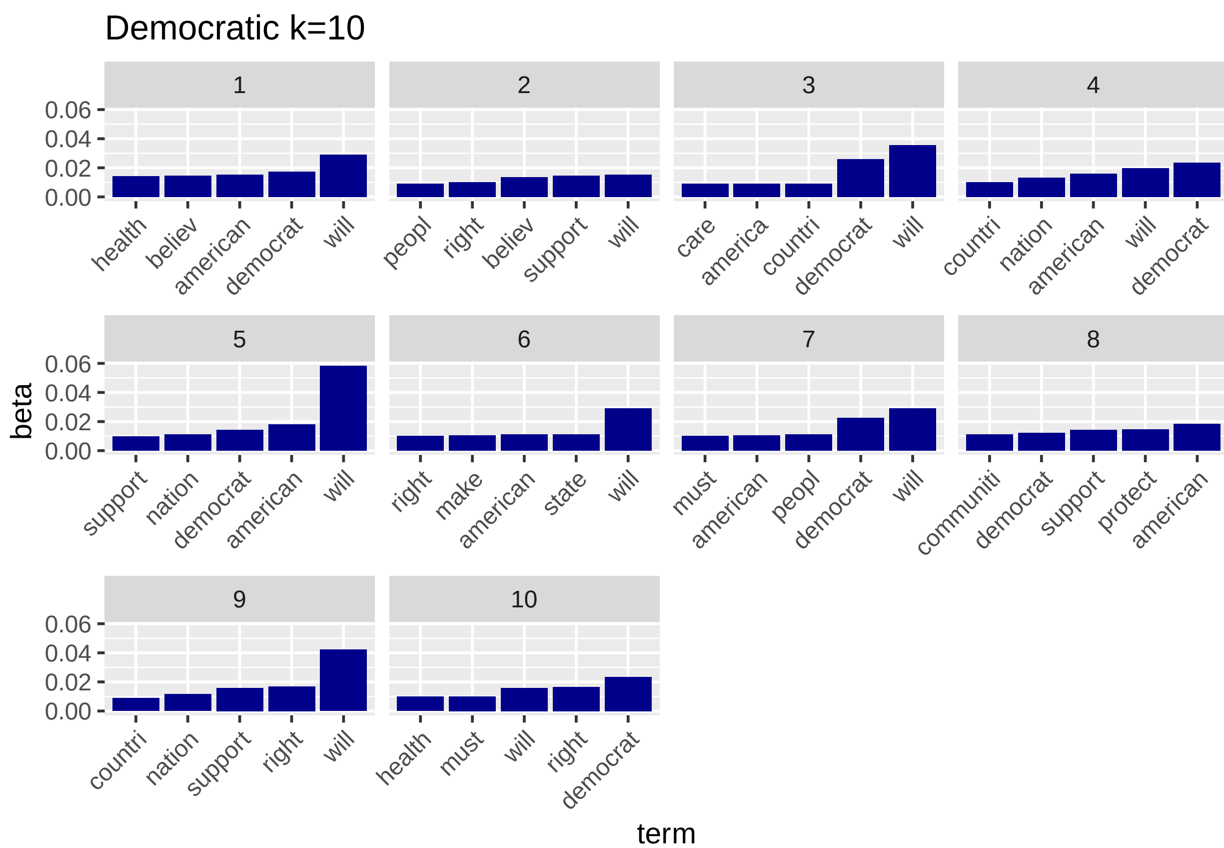


10. Building on the previous question, display a barplot of the `k = 10` model for each party, and offer some general inferences as to the main trends that emerge. Are there similar themes between the parties? Do you think `k = 10` likely picks up differences more efficiently? Why or why not?

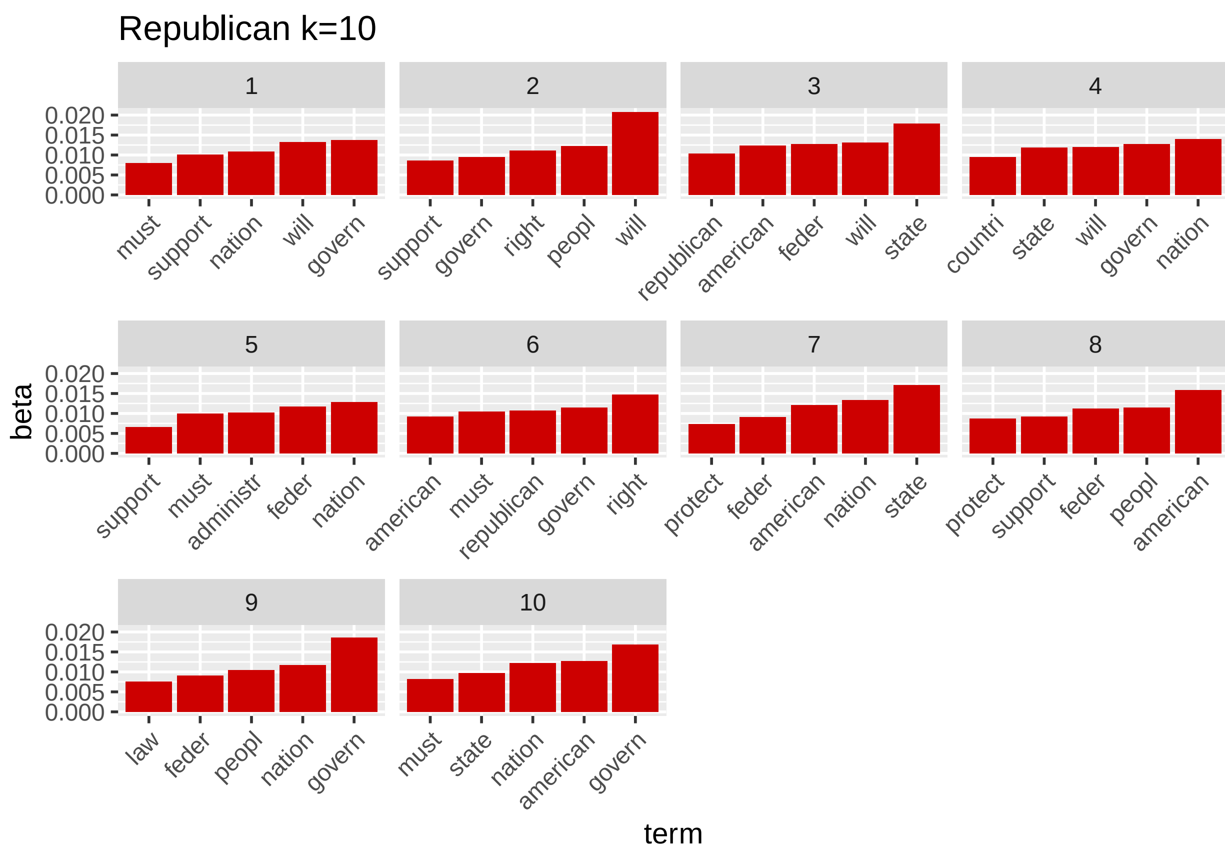
I do not think increasing the k value from five to ten makes the model capture differences more efficiently. As demonstrated in the barplots, the top five words of each topic are still similar within each party and do not differ much between parties. Increasing the number of topics only distributed common words a little wider without meaningful findings. This might have been because some stopwords had not been removed. However, my question lies in how to decide which stopwords should be removed. I understand that stopwords are usually decided from one's prior knowledge on the topic. However, if we have to deal with a new or unfamiliar topic, can we remove words that have the high frequency for all or most of the topics?

```
#Democratic, k=10
Ddtm_lda10 <- LDA(Ddtm, k = 10, control = list(seed = 123))
Ddtm_lda10_td <- tidy(Ddtm_lda10)
Dtop_terms <- Ddtm_lda10_td %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
Dtop_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta)) +
  geom_bar(stat = "identity", fill="blue4") +
```

```
  scale_x_reordered() +
  facet_wrap(~ topic, scales = "free_x")+
  ggtitle("Democratic k=10") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
#Republican, k=10
Rdtm_lda10 <- LDA(Rdtm, k = 10, control = list(seed = 1234))
Rdtm_lda10_td <- tidy(Rdtm_lda10)
Rtop_terms <- Rdtm_lda10_td %>%
  group_by(topic) %>%
  top_n(5, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
Rtop_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta)) +
  geom_bar(stat = "identity", fill="red3") +
  scale_x_reordered() +
  facet_wrap(~ topic, scales = "free_x")+
  ggtitle("Republican k=10") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



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)?

From the topic analyses, the Democratic and Republican parties do not seem to differ in what they talk about in their platforms as I could not find any significant dissimilarities in the topic. However, the sentiment analysis shows that the Democratic party tends to be more negative about the future. Also, the word cloud results indicate that the Democratic party might discuss specific issues such as health policy. Combined with sentiment analysis, I would support the Democratic party since it seems to be more concerned about the future and focused on more specific problems.