CMAP HW3

Wai Laam (Josie) Lui

11/24/2019

Loading Packages

```
library(tidyverse)
library(tm)
library(gridExtra)
library(wordcloud)
library(tidytext)
library(textdata)
library(topicmodels)
```

General NLP/Preprocessing

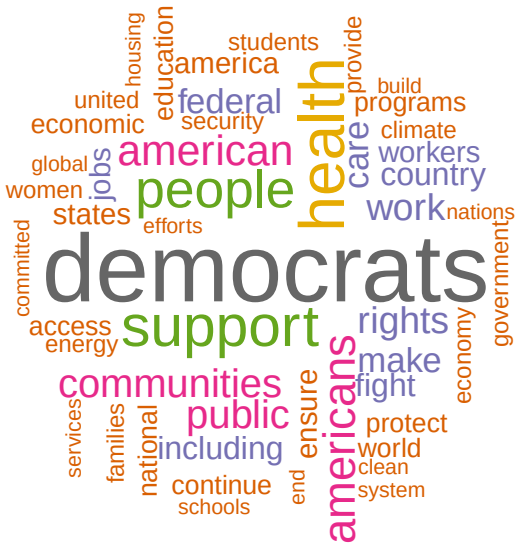
```
platform <- read_csv("Party Platforms Data/platforms.csv")

gop <- VCorpus(VectorSource(read_lines("Party Platforms Data/platforms/r16.txt")))
dem <- VCorpus(VectorSource(read_lines("Party Platforms Data/platforms/d16.txt")))


clean <- function(corp){
  corp <- corp %>% tm_map(removePunctuation) %>%
    tm_map(removeNumbers) %>%
    tm_map(tolower) %>%
    tm_map(removeWords,
           stopwords("SMART")) %>% # consulted w/ stopwords documentation
    # smart seems to be much more useful for topic modeling purposes
    tm_map(stripWhitespace) %>%
    tm_map(PlainTextDocument)
  return(corp)
}

gop <- clean(gop)
dem <- clean(dem)

# remove empty rows for LDA use
gop_dtm <- DocumentTermMatrix(gop)
rowSum <- apply(gop_dtm , 1, sum)
gop_dtm <- gop_dtm[rowSum> 0, ]
dem_dtm <- DocumentTermMatrix(dem)
rowSum <- apply(dem_dtm , 1, sum)
dem_dtm <- dem_dtm[rowSum> 0, ]

gop_freq <- sort(colSums(as.matrix(gop_dtm)),
                 decreasing=TRUE) # add number of times each term is used, and sorting based on frequency

dem_freq <- sort(colSums(as.matrix(dem_dtm)),
                 decreasing=TRUE) # add number of times each term is used, and sorting based on frequency
```

```
gop_wf <- data.frame(word = names(gop_freq),
                     freq = gop_freq)

dem_wf <- data.frame(word = names(dem_freq),
                     freq = dem_freq)
```

```
set.seed(43450) # specifies start/end, making configuration consistent for each plot
wordcloud(names(gop_freq),gop_freq,
          scale = c(3,0.2),
          max.words = 50,
          random.order = FALSE, # centers cloud by frequency, > = center
          rot.per = 0.30, # sets proportion of words oriented horizontally
          main = "Republican Platform",
          colors = brewer.pal(8, "Dark2"))
```

The most common 50 words in the Republican party platform are shown in the wordcloud above. They emphasize the American identity (American and Federal being frequent). In terms of issues such as energy, economy, and healthcare, they are less common than grand yet abstract concepts such as “states”, “congress”, and “rights”.

```
set.seed(902345)
wordcloud(names(dem_freq),dem_freq,
          scale = c(3,0.2),
          max.words = 50,
          random.order = FALSE, # centers cloud by frequency, > = center
          rot.per = 0.30, # sets proportion of words oriented horizontally
          main = "Democratic Platform",
          colors = brewer.pal(8, "Dark2"))
```



As for the Democratic 2016 platform, the word "democrats" happen to be the most frequent word. Since both wordmaps were plotted with the same scaling configurations, the fact that the democratic wordmap occupies less space is because word frequency is perhaps less concentrated than the Republican version. Besides "democrat", the Dem platform has a pretty broad and balanced coverage of many issues: "people", "health", and "communities". Some less mentioned topics include "schools", "economy:, and"climate".

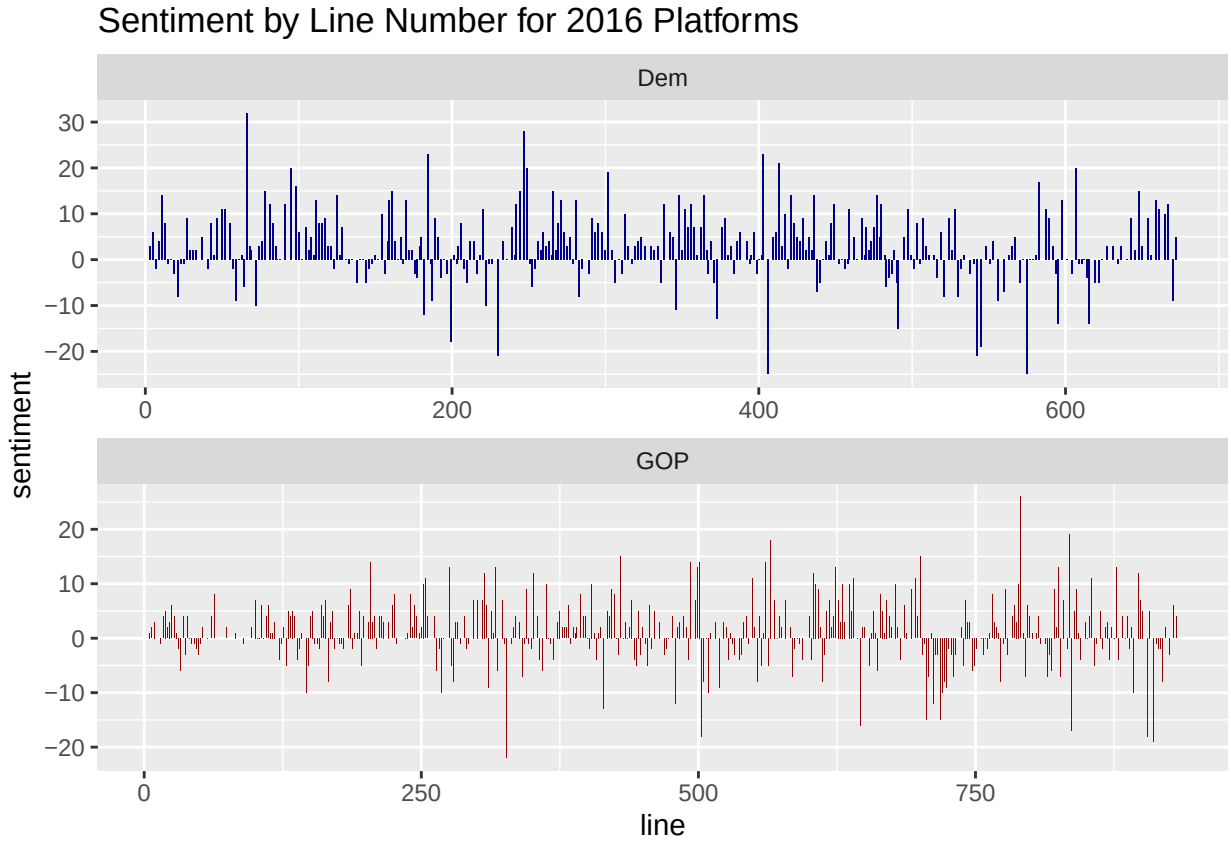
At first glance, in comparison to the Republican platform, the Democratic platform focuses more on issues, especially healthcare, and has a much stronger and approacheable focus on inclusion, while the Republican platform is filled with more "fundamental" issues related to laws and government. The Democratic platform is perhaps for forward-looking in its attention to specific policy initiatives, while the Republican platform still somehow adheres to the conservatism tradition.

Sentiment Analysis (Questions 4-5)

We will conduct sentiment analysis via `tidytext`.

```
gop2 <- read_lines("Party Platforms Data/platforms/r16.txt")
dem2 <- read_lines("Party Platforms Data/platforms/d16.txt")

tidy_clean <- function(chr_vec){
  chr_vec <- chr_vec %>% str_remove_all("[0-9]+") %>% # remove numbers
    str_remove_all("will") %>% str_squish() # remove "will"
  df <- tibble(line = 1:length(chr_vec),text = chr_vec)
  res <- df %>% unnest_tokens(word,text,token = "words") %>% # takes care of to-lower and removes punctuat
    anti_join(stop_words, by = "word") # remove stop word
  return(res)
}

tidy_gop <- tidy_clean(gop2) # tidytext
tidy_dem <- tidy_clean(dem2)

afinn_gop <- tidy_gop %>%
```

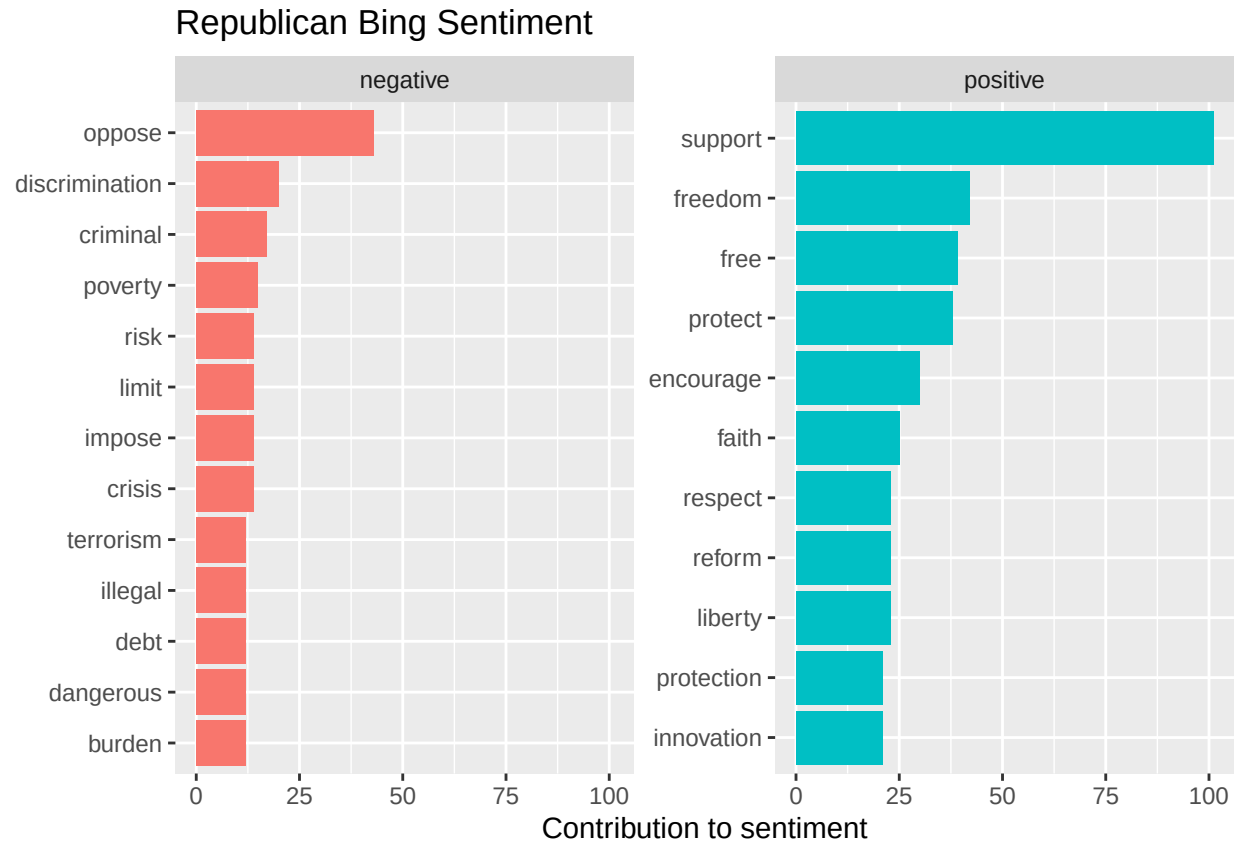
```
  inner_join(get_sentiments("afinn")) %>%
  group_by(line) %>%
  summarise(sentiment = sum(value)) %>%
  mutate(party = "GOP")

afinn_dem <- tidy_dem %>%
  inner_join(get_sentiments("afinn")) %>%
  group_by(line) %>%
  summarise(sentiment = sum(value)) %>%
  mutate(party = "Dem")

bind_rows(afinn_gop,afinn_dem) %>%
  ggplot(aes(line,sentiment,fill=party)) +
  geom_col(show.legend = FALSE) +
  scale_fill_manual(values = c("GOP" = "darkred","Dem" = "darkblue"))+
  labs(title = "Sentiment by Line Number for 2016 Platforms") +
  facet_wrap(~party, ncol = 1, scales = "free")
```



In the above plot, we compute the total AFINN score for each row of text. The two platform differ slightly in length, but they are both in the 600-800 paragraph range. In terms of rhetoric, positive and negative sentiment are interspersed in both platforms, yet both platforms have more positive sentiment than negative. In terms of extremities, the maximum aggregate sentiment, both positive (30) and negative (-25), on the democratic side is slightly larger. Overall the difference in extremity is negligible. In terms of sentiment frequency, the democratic platform has a noticeably higher proportion of positivity throughout the lines than the republican platform. From this we can perhaps conclude that the democratic party is more optimistic about the future.

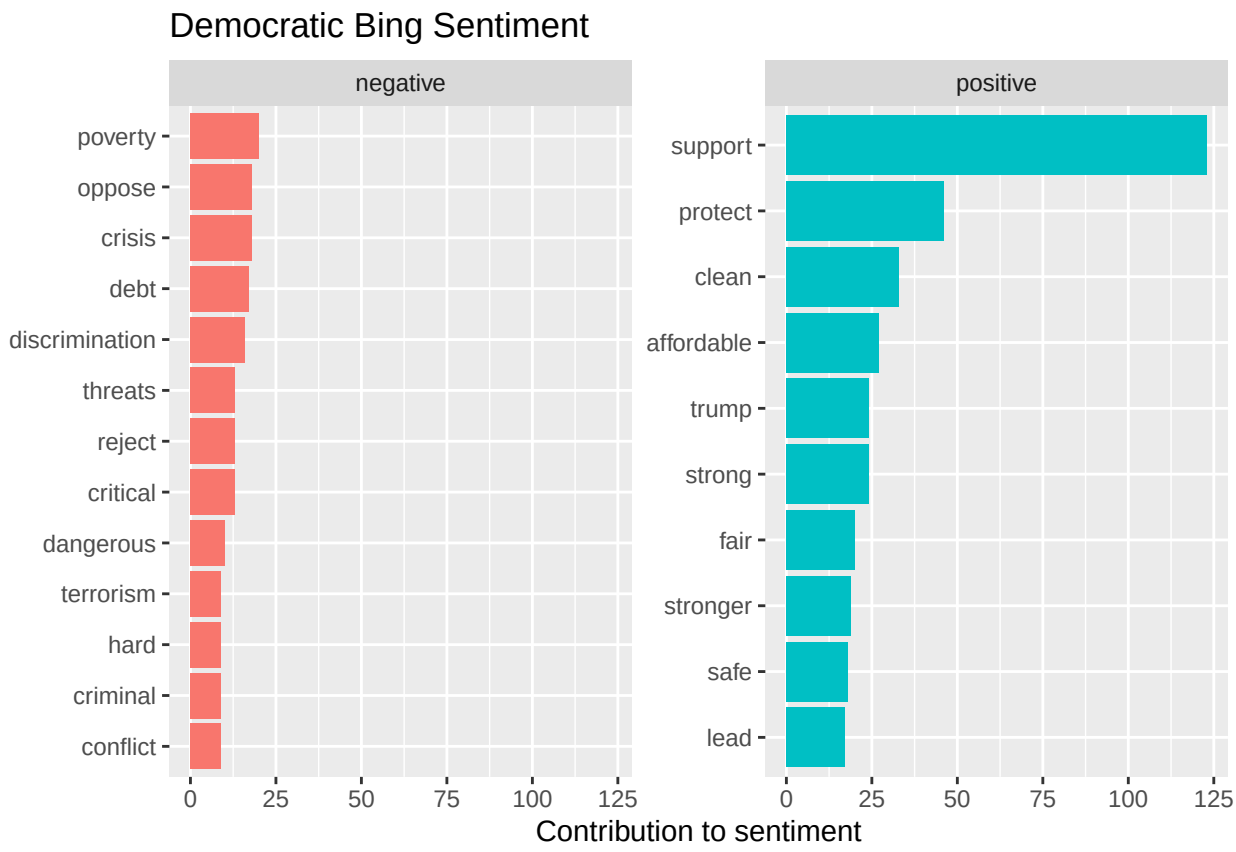
```
bing_gop <- tidy_gop %>%
  inner_join(get_sentiments("bing")) %>%
  count(word,sentiment,sort = TRUE)
```

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

bing_gop %>%
  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(title = "Republican Bing Sentiment",
       y = "Contribution to sentiment",
       x = NULL) +
  coord_flip()
```




```
bing_dem %>%
  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(title = "Democratic Bing Sentiment",
       y = "Contribution to sentiment",
       x = NULL) +
```

```
  coord_flip()
```



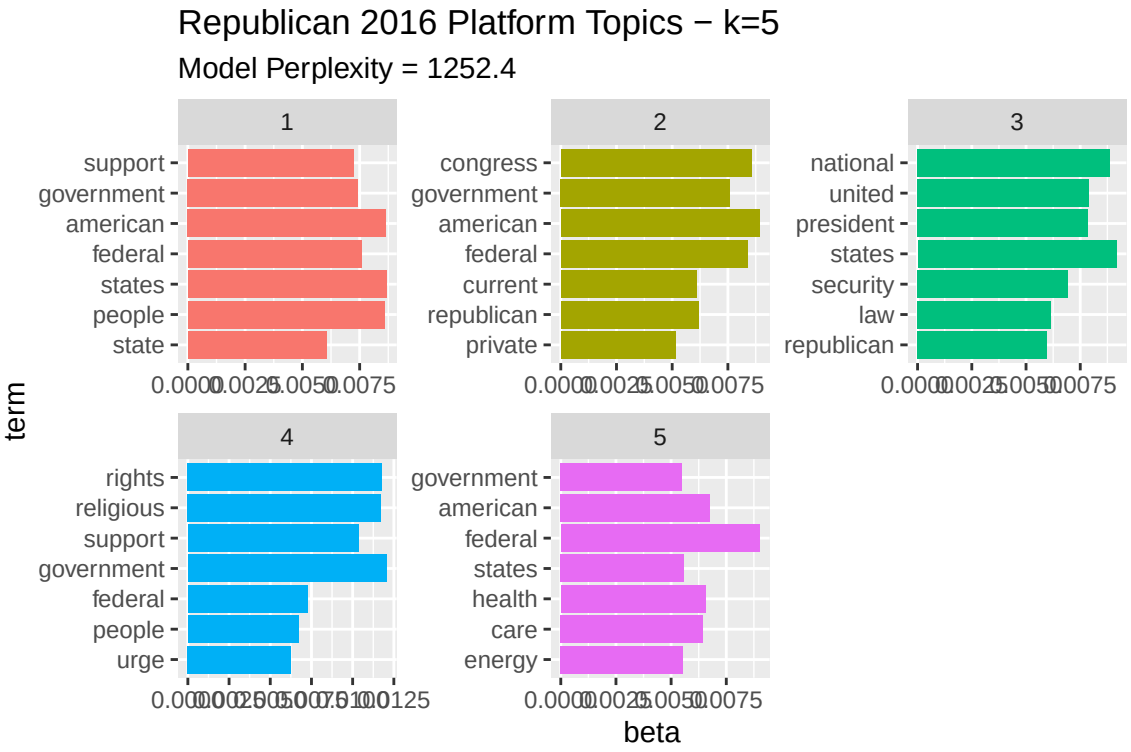
Using the Bing dictionary, we see more clearly that the Democratic party is probably more optimistic about the future, since the occurence of negative words is much less common than the Republican side. In the Republican platform, many positive words relate to the notion of "protection", echoing the GOP tradition of conservatism, while in the Democratic platform, words like "clean", "affordable", and "fair" etc indicate a much more progressive stance towards contemporary issues. Overall, sentiment analysis using word tokens conform with my existing perception about the two parties.

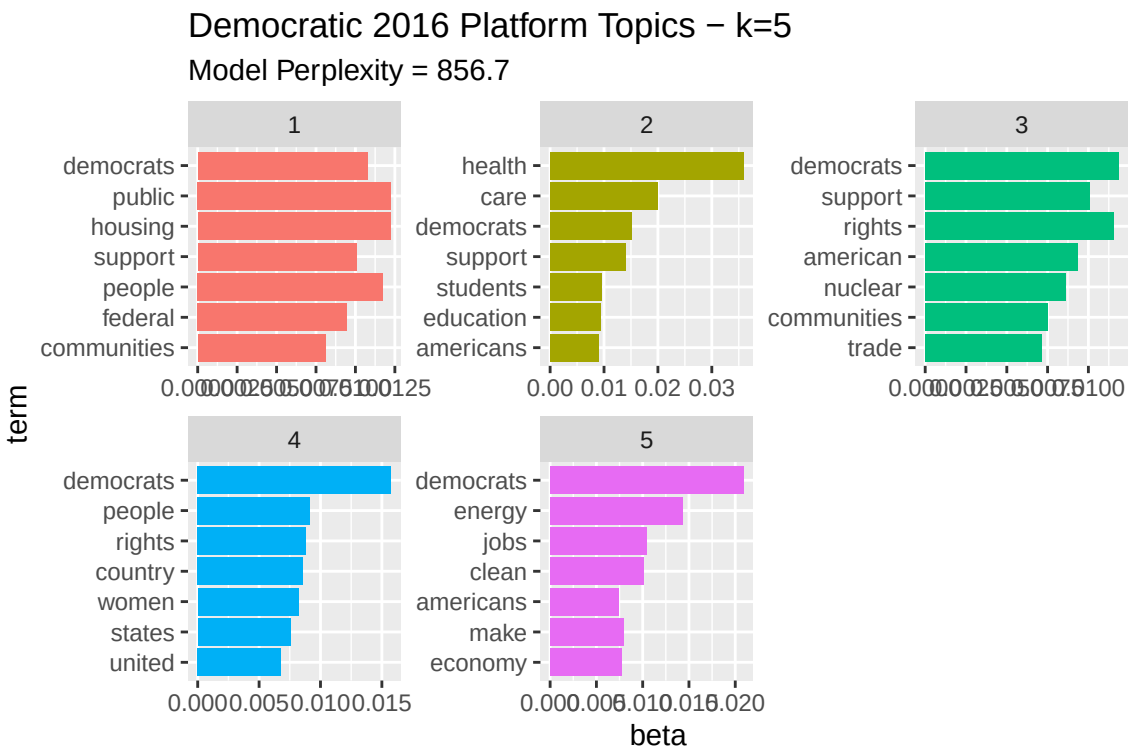
Topic Models

Question 6-7

```
presentLDA <- function(dtm,topic,s,title_text){
  k_lda <- LDA(dtm, k = topic, control = list(seed = s))

  perp <- perplexity(k_lda)

  k_top <- tidy(k_lda,matrix = "beta") %>%
    group_by(topic) %>%
    top_n(7,beta) %>%
    ungroup() %>%
    arrange(topic,-beta)

  if(topic<=12){
    k_top %>%
      mutate(term = reorder(term,beta)) %>%
      filter(topic <= 12) %>%
```

```
      ggplot(aes(term,beta,fill = factor(topic))) +
      labs(title = str_c(title_text," - k=", as.character(topic)),
           subtitle = str_c("Model Perplexity = ",as.character(round(perp,2))))+
      geom_col(show.legend = FALSE) +
      facet_wrap(~ topic, scales = "free") +
      coord_flip()
  }
  else{
    print(str_c("Model Perplexity at k=",as.character(topic)," is ",as.character(perp)))
    for(i in seq(1,topic)){
      header = str_c("Topic ",as.character(i),": ")
      result = str_c(k_top %>% filter(topic == i) %>% pull(term), collapse = ", ")
      print(str_c(header,result))
    }
  }
}
# the 10 LDAs will sequentially use these seeds
set.seed(2498)
seeds = sample(1:100000, size = 10)
```



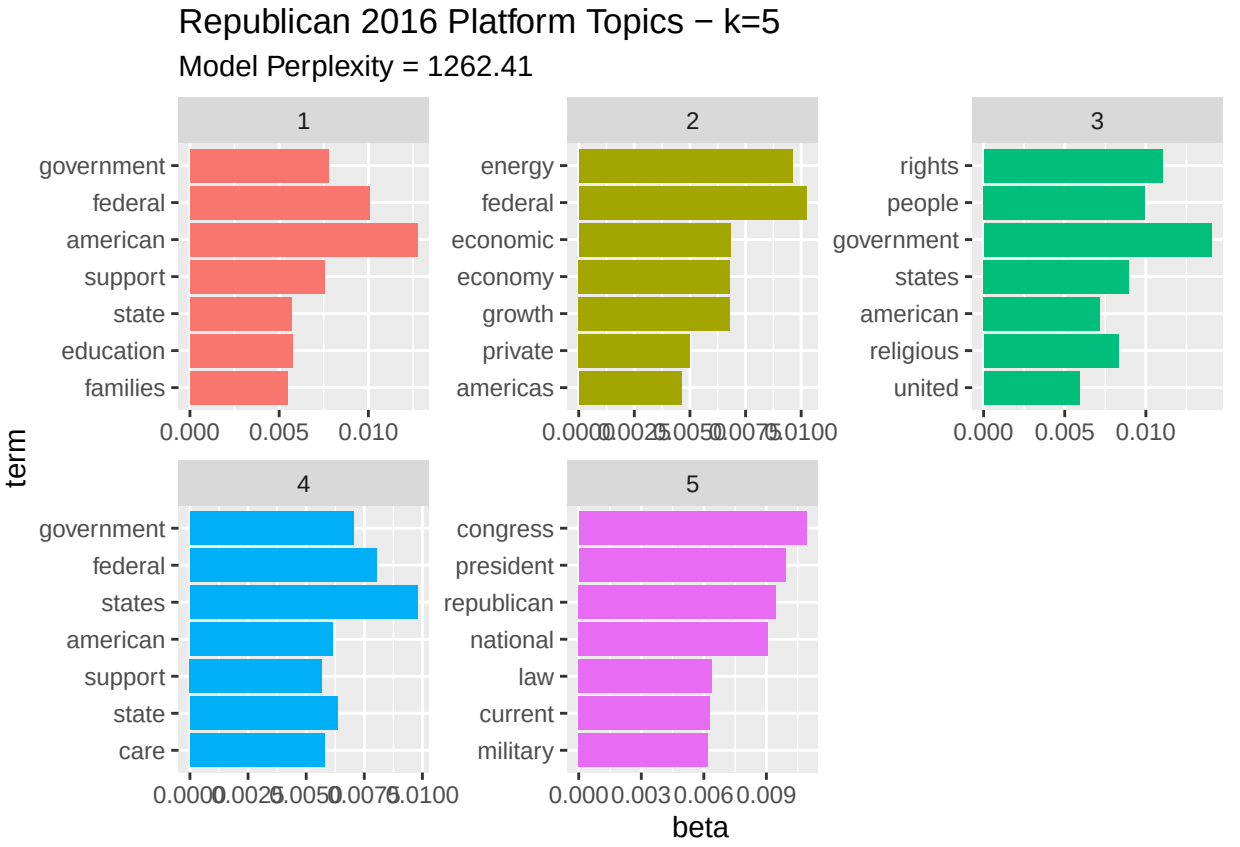


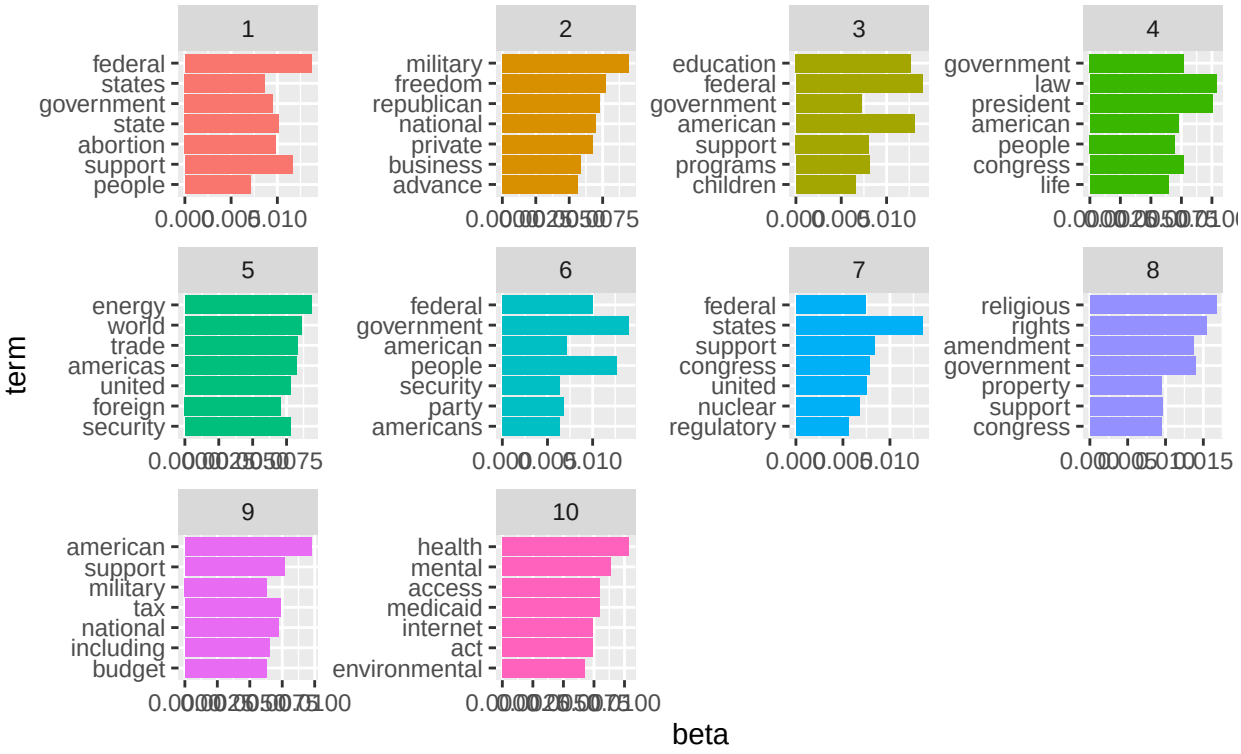
In the Republican LDA plot, we can eyeball the following topics: (1) traditional topic of federal-state relations, (2) traditional issue of federal/congressional relationship with the people/private rights, (3) national security (4) religious freedom/rights, (5) health care and other topics. Each topic seems to be well-separated from one another.

In the Democratic LDA plot, we can infer the following topics: (1) community and public housing, (2) health care and education, (3) community? trade? nuclear? (4) womens' rights, and (5) clean energy and economy. Other than (3), each topic seems to be quite distinct. Separation between topics is also visible.

At the k=5 level, the Republican platform and Democratic platform are very different. Our results are consistent with the conclusions drawn from the wordcloud: Republicans seem more concerned with traditional topics, while Democrats are more concerned with emerging issues.

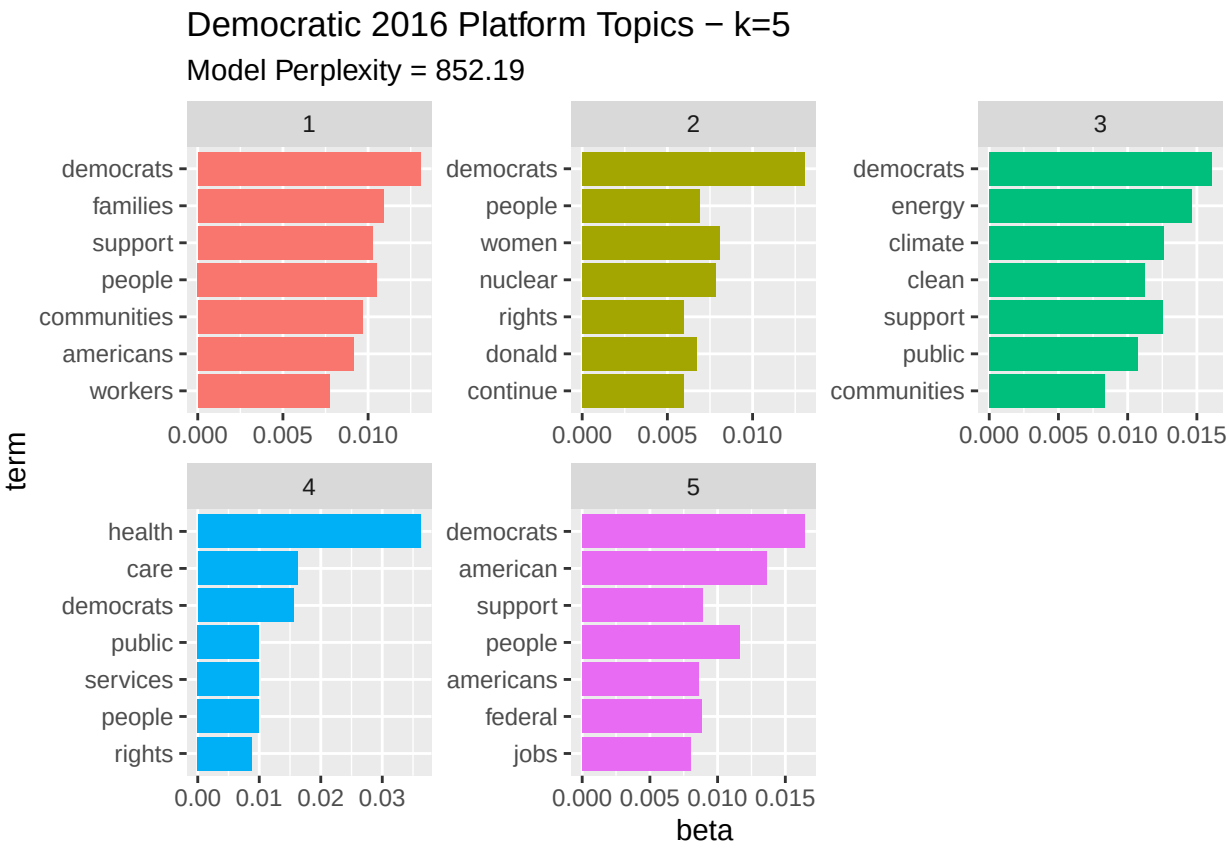
We will now fit six additional topic models - three for each party at topic numbers 5,10,25 (question 8).

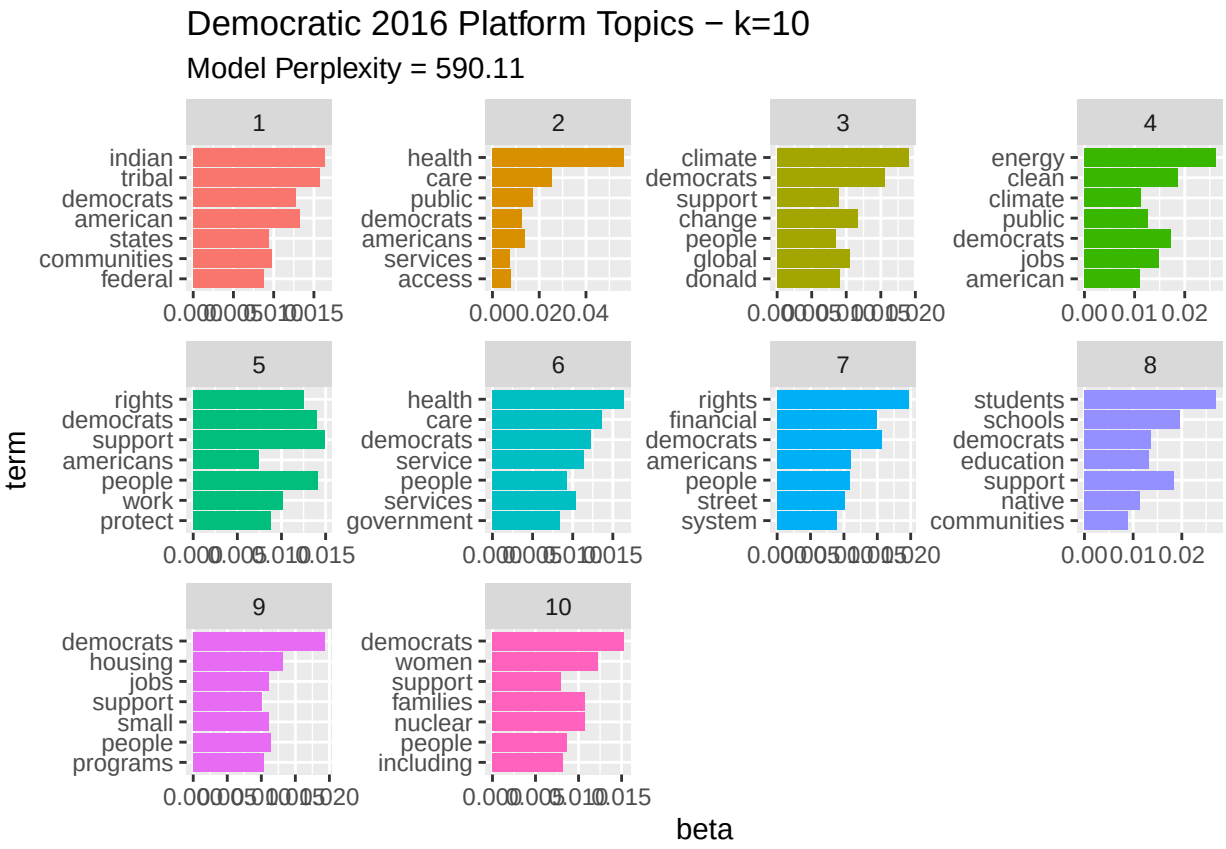





```
## [1] "Model Perplexity at k=25 is 469.716455681809"
```

```
## [1] "Topic 1: americas, private, economic, assistance, foreign, development, prosperity"
## [1] "Topic 2: public, species, fda, resources, health, protection, lands"
## [1] "Topic 3: banks, federal, food, administration, american, tax, current"
## [1] "Topic 4: government, american, people, internet, principles, free, power"
## [1] "Topic 5: states, congress, state, powers, authority, president, laws"
## [1] "Topic 6: research, continue, stem, federal, human, support, people"
## [1] "Topic 7: care, veterans, federal, service, access, make, tribal"
## [1] "Topic 8: environmental, federal, act, national, congress, air, space"
## [1] "Topic 9: religious, taiwan, president, region, agreement, people, relations"
## [1] "Topic 10: energy, support, federal, americas, nuclear, china, future, irs, tax"
## [1] "Topic 11: law, federal, immigration, criminal, illegal, oppose, states"
## [1] "Topic 12: trafficking, law, victims, crime, labor, workers, government, marriage, sexual"
## [1] "Topic 13: poverty, labor, federal, advancing, president, union, workers"
## [1] "Topic 14: government, amendment, private, family, families, life, states"
## [1] "Topic 15: states, united, education, american, government, support, protect"
## [1] "Topic 16: american, rights, government, natural, godgiven, protect, resources"
## [1] "Topic 17: military, national, republican, guard, world, security, america"
## [1] "Topic 18: growth, economy, property, freedom, economic, innovation, intellectual"
## [1] "Topic 19: schools, freedom, education, political, states, control, state"
## [1] "Topic 20: religious, amendment, rights, support, government, congress, tax"
## [1] "Topic 21: federal, private, debt, student, territories, support, advance, college"
## [1] "Topic 22: america, education, federal, american, public, people, government, school"
## [1] "Topic 23: energy, trade, security, markets, current, government, democratic"
## [1] "Topic 24: abortion, support, women, military, states, child, service"
## [1] "Topic 25: health, mental, medicaid, healthcare, insurance, care, coverage, states"
```

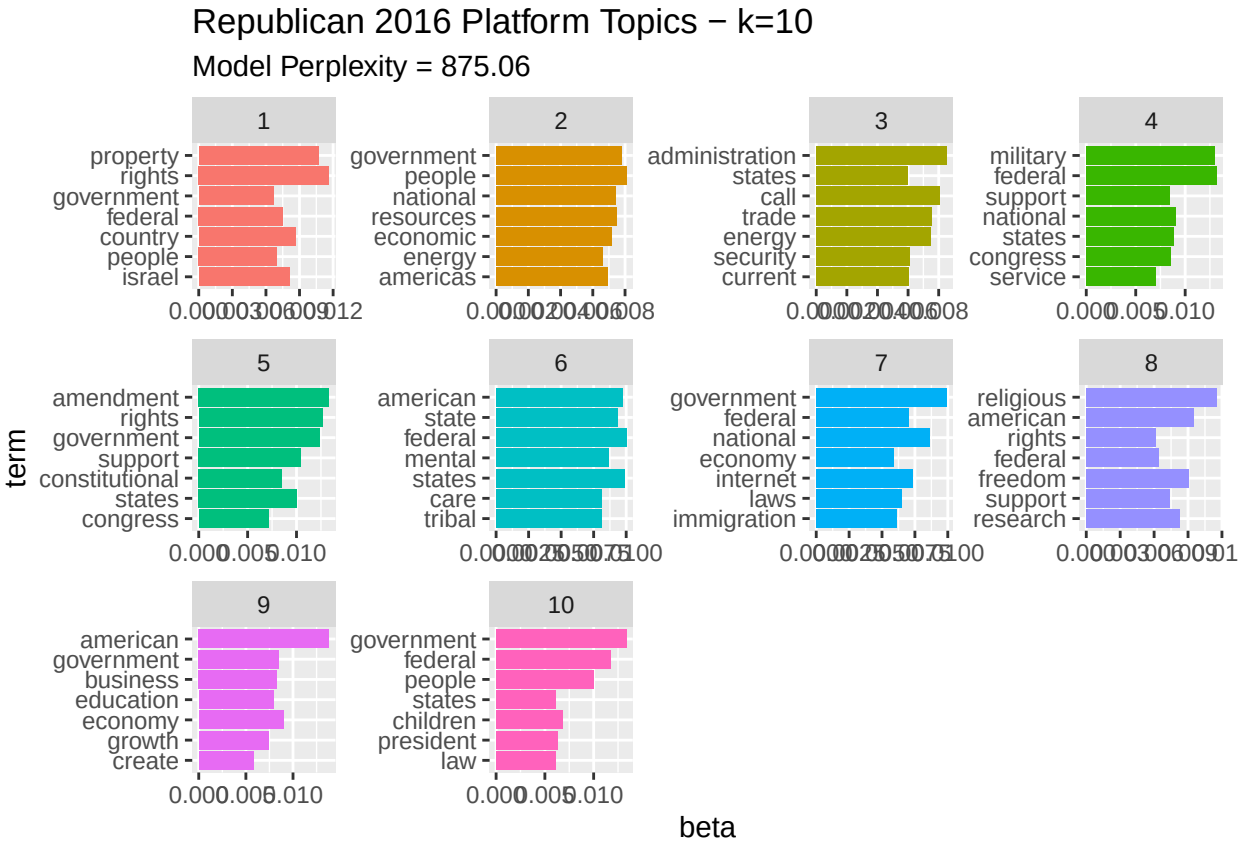




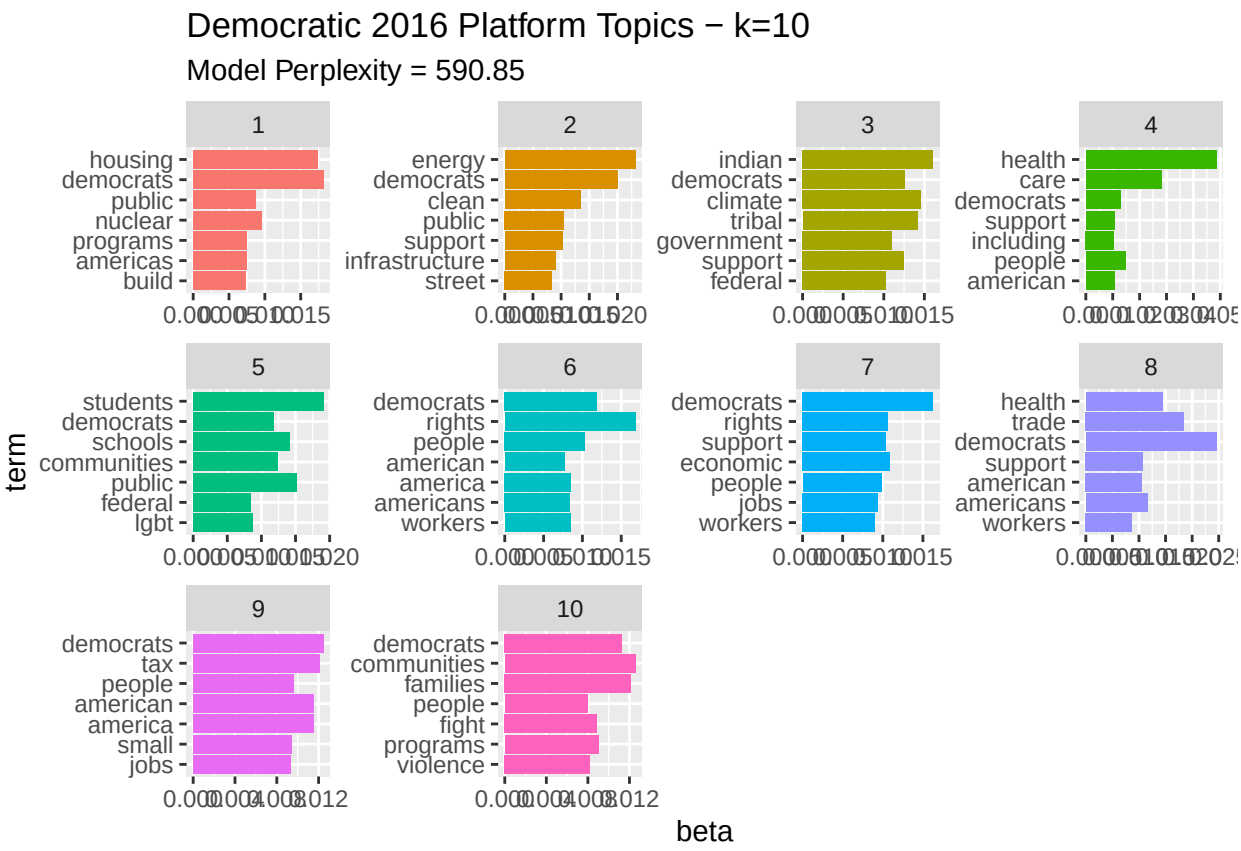

```
## [1] "Model Perplexity at k=25 is 326.115044287847"
## [1] "Topic 1: federal, schools, sexual, democrats, students, veterans, workers"
## [1] "Topic 2: nuclear, democrats, weapons, wage, americans, work, people"
## [1] "Topic 3: housing, services, service, affordable, increase, democrats, postal"
## [1] "Topic 4: global, partners, afghanistan, democrats, allies, nato, threats"
## [1] "Topic 5: students, education, institutions, democrats, racial, americans, federal"
## [1] "Topic 6: rights, americans, people, support, civil, work, workers"
## [1] "Topic 7: voting, people, rights, democrats, system, united, families"
## [1] "Topic 8: climate, change, clean, energy, democrats, national, economy"
## [1] "Topic 9: support, nations, democrats, programs, arts, communities, america"
## [1] "Topic 10: health, public, children, means, education, democrats, ensure"
## [1] "Topic 11: indian, tribal, support, native, recognize, democrats, education, marijuana"
## [1] "Topic 12: democrats, gun, americans, health, violence, communities, families"
## [1] "Topic 13: leave, fight, people, support, defense, paid, democrats"
## [1] "Topic 14: health, care, access, mental, including, hiv, reproductive"
## [1] "Topic 15: people, lgbt, democrats, rights, disabilities, care, americans"
## [1] "Topic 16: trade, american, democrats, workers, world, protect, agreements"
## [1] "Topic 17: small, democrats, government, support, business, communities, federal"
## [1] "Topic 18: economic, communities, democrats, public, economy, extreme, lands, percent"
## [1] "Topic 19: jobs, make, investments, american, workers, goodpaying, support"
## [1] "Topic 20: people, democrats, puerto, american, world, young, development"
## [1] "Topic 21: financial, street, wall, democrats, student, americans, borrowers, loan"
## [1] "Topic 22: energy, clean, democrats, make, government, public, renewable"
## [1] "Topic 23: democrats, iran, american, donald, nuclear, people, russian, terrorism"
## [1] "Topic 24: pay, security, social, democrats, families, health, workers"
## [1] "Topic 25: women, schools, democrats, public, states, rights, united"
```

For Republicans, we notice the perplexity score as follows: 1262 (k=5), 870 (k=10), 470(k=25). For Democrats, we notice the perplexity score as follows: 852 (k=5), 590 (k=10), 326 (k=15). The perplexity score for Democrats is

uniformly lower than that of Republicans. This is expected because the Republican platform is around 40-50% longer. As the number of topics allowed increases, we see consistent decrease in perplexity score, indicating that the model captures term correlations better and better. Technically, we should use k=25 for both parties because it gives the lowest perplexity scores.



For a 10-topic LDA model for the Republican platform, we notice the following topics: property rights, energy, trade, military, constitution, mental care and tribal issues, internet, economy and immigration, religious freedom & research, business and economic growth, and people.



For a 10-topic LDA model for the Democratic platform, we notice the following topics: public housing, clean energy, climate and tribal affairs, health care, communities and education, workers' rights, economy and jobs, workers and trade, taxation, and family.

We see some overlap between the two parties regarding tribal affairs, business and economy. Other than that the platforms are quite divergent. In comparison to k=5, the models pick up more distinct topics, yet there is still cases where one "topic" actually covers more than one issue. Since the inter-topic separation is still quite prominent, we may conclude that k=10 is an efficient configuration.

Conclusions

As of now, I do not see the values & traditions the Republican party is trying to preserve in jeopardy, yet many modern-day issues such as climate change and economic empowerment are of pressing importance. We are in an age of dynamic change, so it is important that we adapt ourselves to imminent change in time and with optimism. As a result, I would support the Democratic platform in 2020.