PS5

Yaoxi Shi

20 November 2019

```
# Load data
platforms <- read_csv("~/Uchicago/courses/Unsupervised machine learning/HW2/Problem-Set-5/Party Platfor
platforms[1,2]
```

```
## # A tibble: 1 x 1
##   platform
##   <chr>
## 1 "In 2016, Democrats meet in Philadelphia with the same basic belief that~
```

```
doc_democrat=VCorpus(VectorSource(platforms[1,2]))
doc_republican=VCorpus(VectorSource(platforms[2,2]))

# Preprocessing
doc_democrat=tm_map(doc_democrat, tolower)
doc_democrat=tm_map(doc_democrat, removePunctuation)
doc_democrat=tm_map(doc_democrat, removeNumbers)
doc_democrat=tm_map(doc_democrat, removeWords, stopwords("english"))
doc_democrat=tm_map(doc_democrat, stripWhitespace)
doc_democrat=tm_map(doc_democrat, PlainTextDocument)
for(j in seq(doc_democrat)){
  doc_democrat[[j]] = gsub("will", "", doc_democrat[[j]])
}
doc_democrat=tm_map(doc_democrat, PlainTextDocument)

writeLines(as.character(doc_democrat[1]))
```

```
## list(platform = list(content = " democrats meet philadelphia basic belief animated continental congr
##     meta = list(author = character(0), datetimestamp = list(sec = 51.8200509548187, min = 3, hour = 
## list()
## list()
```

```
doc_republican=tm_map(doc_republican, tolower)
doc_republican=tm_map(doc_republican, removePunctuation)
doc_republican=tm_map(doc_republican, removeNumbers)
doc_republican=tm_map(doc_republican, removeWords, stopwords("english"))
doc_republican=tm_map(doc_republican, stripWhitespace)
for (j in seq(doc_republican)) {
  doc_republican[[j]] <- gsub("-", "", doc_republican[[j]])
  gsub("will", "", doc_democrat[[j]])}
doc_republican=tm_map(doc_republican, PlainTextDocument)

writeLines(as.character(doc_republican[1]))
```

```
## list(platform = list(content = " platform republican party reaffirm principles unite us common purpo
##     meta = list(author = character(0), datetimestamp = list(sec = 52.1228289604187, min = 3, hour = 
## list()
## list()
```

```r
# WordCloud

set.seed(1234)
wordcloud(doc_democrat,max.words = 100,colors = brewer.pal(8, "Dark2"))
```


```r
set.seed(1111)
wordcloud(doc_republican, scale = c(3,.5), max.words = 100,colors = brewer.pal(8, "Dark2"))
```

From the wordcloud for each of the party, we can find that the democratic party uses words “wokers”, “jobs”, “believe”, “people”, “housing” etc more often, suggesting it pays more attention to the labor market. Repablican uses words “federal”, “business”, “growth”, “economy”, “market” etc more often, which shows this party pays more attention to free market and economic growth.

```
# Sentiment Analysis
# Democrat Bing
corpus_demo=doc_democrat %>% tidy()
demo_df <- corpus_demo %>%
  unnest_tokens(word, text) %>%
  select(word)

democrat_sent_bing=demo_df %>% inner_join(get_sentiments("bing")) %>%
  count(word, sort=TRUE)
```

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

```
democrat_sent_bing
```

```
## # A tibble: 202 x 2
##    word           n
##    <chr>      <int>
##  1 support       24
##  2 work          21
##  3 better        11
##  4 right         10
##  5 affordable     9
##  6 innovation     9
```

```
##   7 stronger       8
##   8 top            8
##   9 benefits       7
##  10 fair           6
## # ... with 192 more rows
```

```
democrat_sent_bing2=demo_df %>% inner_join(get_sentiments("bing")) %>%
  count(sentiment, sort=TRUE) %>%
  spread(sentiment, n, fill=0) %>%
  mutate(positive/nrow(demo_df),negative/nrow(demo_df), sentiment=positive-negative)
```

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

```
democrat_sent_bing2
```

```
## # A tibble: 1 x 5
##   negative positive `positive/nrow(demo_d~ `negative/nrow(demo_d~ sentiment
##      <dbl>    <dbl>                  <dbl>                  <dbl>     <dbl>
## 1      118      289                 0.0964                 0.0394       171
```

```
# Republican Bing
corpus_repu=doc_republican %>% tidy()
repu_df= corpus_repu %>%
  unnest_tokens(word, text) %>%
  select(word)

republican_sent_bing=repu_df %>% inner_join(get_sentiments("bing"))%>%
  count(word, sort=TRUE)
```

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

```
republican_sent_bing
```

```
## # A tibble: 252 x 2
##    word            n
##    <chr>       <int>
##  1 innovation     10
##  2 freedom         7
##  3 prosperity      7
##  4 reform          7
##  5 support         7
##  6 free            5
##  7 best            4
##  8 fair            4
##  9 interests       4
## 10 led             4
## # ... with 242 more rows
```

```
republican_sent_bing2=repu_df %>% inner_join(get_sentiments("bing")) %>%
  count(sentiment, sort=TRUE) %>%
  spread(sentiment, n, fill=0) %>%
  mutate(positive/nrow(repu_df),negative/nrow(repu_df), sentiment=positive-negative)
```

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

```
republican_sent_bing2
```

```
## # A tibble: 1 x 5
##   negative positive `positive/nrow(repu_d~ `negative/nrow(repu_d~ sentiment
```

```
##      <dbl>    <dbl>                    <dbl>                    <dbl>      <dbl>
## 1      148      231                   0.0802                   0.0514         83
```

```
# Democrat Afinn
democrat_sent_afinn=demo_df %>% inner_join(get_sentiments("afinn")) %>%
  summarise(sentiment=sum(value))
```

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

```
democrat_sent_afinn
```

```
## # A tibble: 1 x 1
##   sentiment
##       <dbl>
## 1       386
```

```
democrat_sent_afinn2=demo_df %>% inner_join(get_sentiments("afinn")) %>%
  count(value)
```

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

```
democrat_sent_afinn2
```

```
## # A tibble: 7 x 2
##   value     n
##   <dbl> <int>
## 1    -3    13
## 2    -2    37
## 3    -1    57
## 4     1   112
## 5     2   175
## 6     3    30
## 7     4     1
```

```
# Republican Afinn
republican_sent_afinn=repu_df %>% inner_join(get_sentiments("afinn"))%>%
  summarise(sentiment=sum(value))
```

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

```
republican_sent_afinn
```

```
## # A tibble: 1 x 1
##   sentiment
##       <dbl>
## 1       210
```

```
republican_sent_afinn2=repu_df %>% inner_join(get_sentiments("afinn"))%>%
  count(value)
```

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

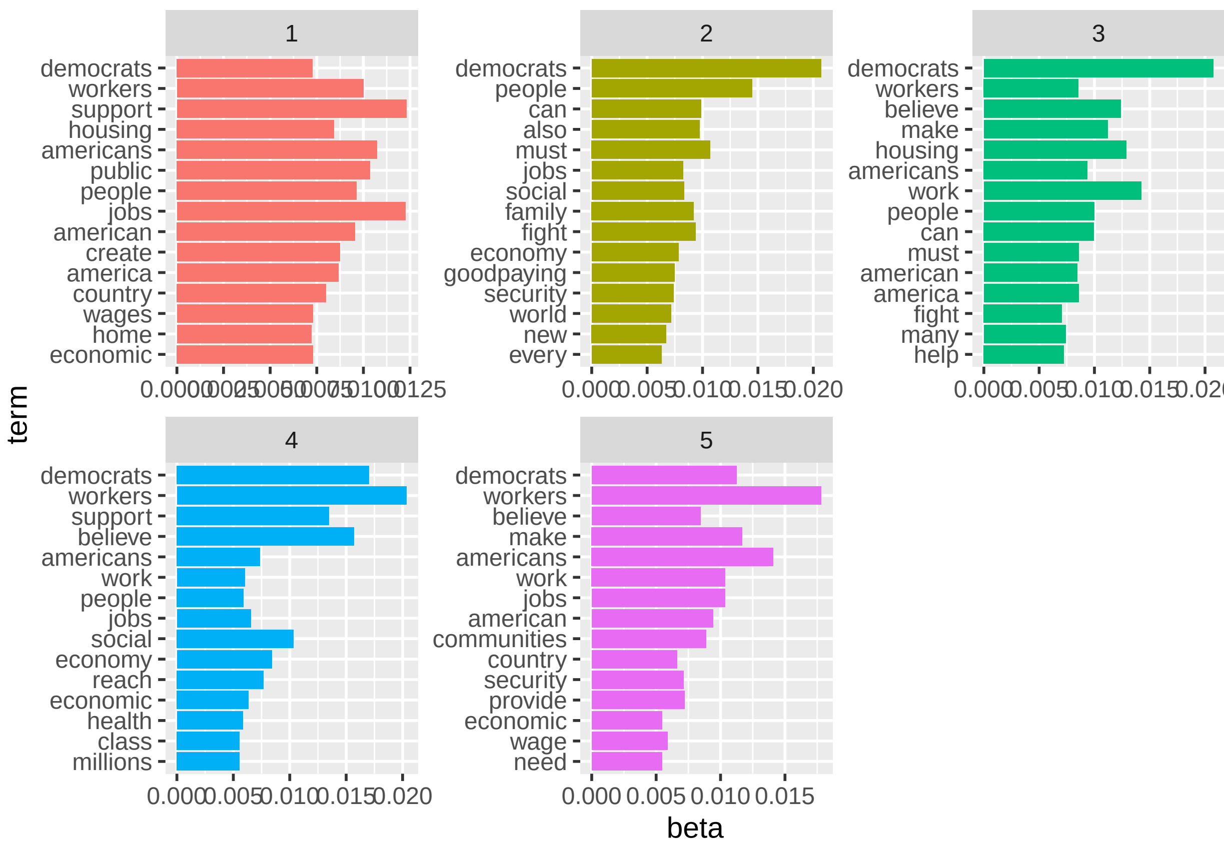
```
republican_sent_afinn2
```

```
## # A tibble: 7 x 2
##   value     n
##   <dbl> <int>
## 1    -3    18
## 2    -2    53
## 3    -1    31
## 4     1   104
```

```
## 5      2   127
## 6      3    13
## 7      4     1
```

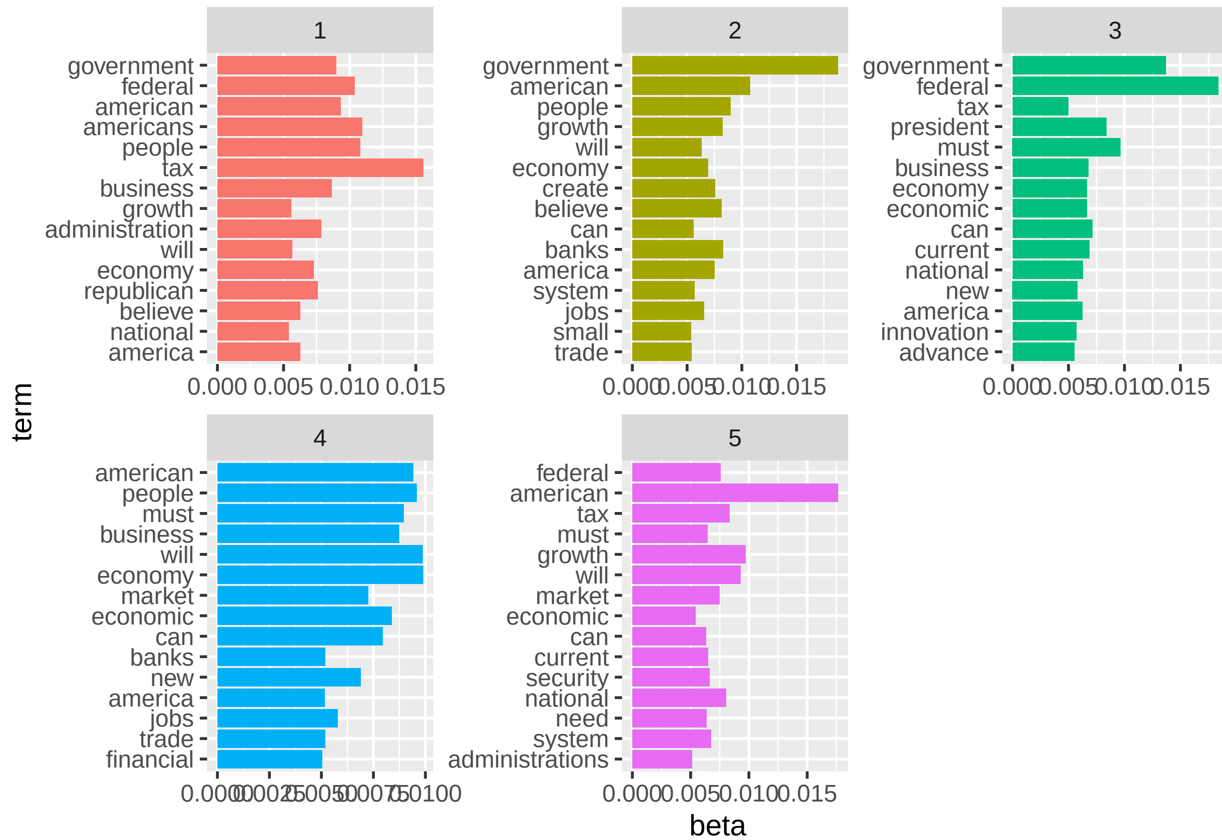
5. Democrat has a larger propotion of positive sentiments compared to republican, and smaller propotion of negative words, also, based on dictionary Afinn, democrat has higher value than republican, so democrat is more positive. This is comport with my perception.

```r
# Democrat
dtm_demo=DocumentTermMatrix(doc_democrat)
demo_lda=LDA(dtm_demo, k=5, control=list(seed=123))
demo_topics=tidy(demo_lda, matrix="beta")

demo_top_terms <- demo_topics %>%
  group_by(topic) %>%
  top_n(15, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
demo_top_terms %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip()
```




```r
# Republican
dtm_repu=DocumentTermMatrix(doc_republican)
repu_lda=LDA(dtm_repu, k=5, control=list(seed=124))
```

```r
repu_topics=tidy(repu_lda, matrix="beta")

repu_top_terms <- repu_topics %>%
  group_by(topic) %>%
  top_n(15, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
repu_top_terms %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip()
```



7. Both of the parties focus on economy, But there are general differences between the two parties, topics like workers, wages, jobs, housing, support, community appear more in democrat, while topics such as business, growth, banks, federal, govenment, market.

```r
# Democrat
#k=10
demo_lda_10=LDA(dtm_demo, k=10, control=list(seed=125))
demo_topics_10=tidy(demo_lda_10, matrix="beta")

demo_top_terms_10 <- demo_topics_10 %>%
  group_by(topic) %>%
  top_n(15, beta) %>%
  ungroup() %>%
```

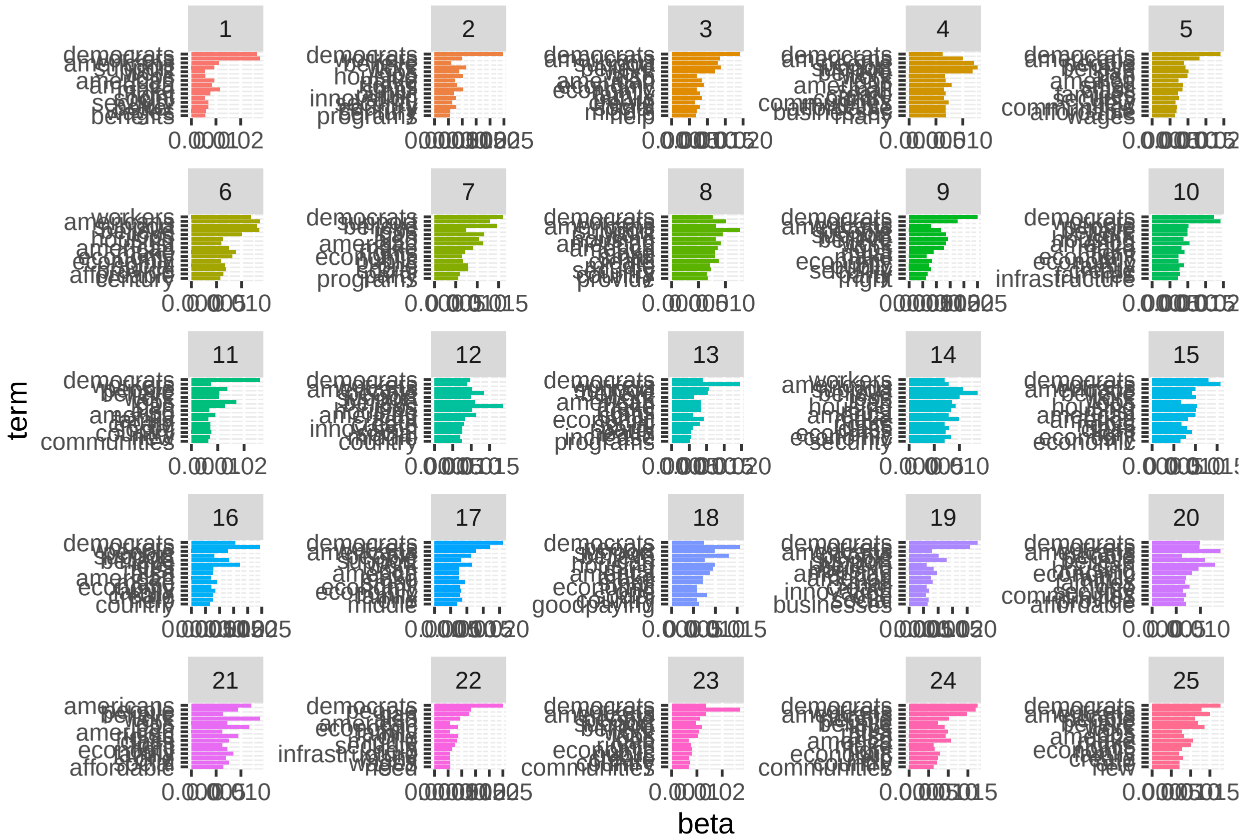
```r
  arrange(topic, -beta)
terms(demo_lda_10, 3)
```

```
##      Topic 1   Topic 2     Topic 3   Topic 4     Topic 5     Topic 6
## [1,] "make"    "democrats" "workers" "democrats" "workers"   "workers"
## [2,] "support" "believe"   "believe" "support"   "democrats" "democrats"
## [3,] "people"  "workers"   "jobs"    "believe"   "american"  "people"
##      Topic 7     Topic 8     Topic 9     Topic 10
## [1,] "democrats" "americans" "democrats" "americans"
## [2,] "believe"   "democrats" "workers"   "believe"
## [3,] "america"   "american"  "america"   "economic"
```

```r
# k=25
demo_lda_25=LDA(dtm_demo, k=25, control=list(seed=163))
demo_topics_25=tidy(demo_lda_25, matrix="beta")

demo_top_terms_25 <- demo_topics_25 %>%
  group_by(topic) %>%
  top_n(15, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
terms(demo_lda_25, 3)
```

```
##      Topic 1     Topic 2     Topic 3     Topic 4   Topic 5     Topic 6
## [1,] "workers"   "democrats" "democrats" "support" "democrats" "americans"
## [2,] "democrats" "work"      "believe"   "people"  "americans" "believe"
## [3,] "make"      "small"     "americans" "believe" "can"       "support"
##      Topic 7     Topic 8   Topic 9     Topic 10    Topic 11    Topic 12
## [1,] "democrats" "support" "democrats" "workers"   "democrats" "jobs"
## [2,] "believe"   "workers" "workers"   "democrats" "jobs"      "people"
## [3,] "support"   "work"    "believe"   "america"   "people"    "american"
##      Topic 13  Topic 14  Topic 15    Topic 16    Topic 17    Topic 18
## [1,] "workers" "believe" "workers"   "workers"   "democrats" "people"
## [2,] "support" "support" "democrats" "jobs"      "workers"   "believe"
## [3,] "believe" "jobs"    "housing"   "democrats" "americans" "support"
##      Topic 19    Topic 20    Topic 21    Topic 22    Topic 23
## [1,] "democrats" "americans" "work"      "democrats" "workers"
## [2,] "workers"   "can"       "americans" "people"    "democrats"
## [3,] "support"   "believe"   "can"       "can"       "americans"
##      Topic 24    Topic 25
## [1,] "democrats" "democrats"
## [2,] "workers"   "americans"
## [3,] "americans" "work"
```

```r
demo_top_terms_25 %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip()
```




```r
# Republican
# k=10
repu_lda_10=LDA(dtm_repu, k=10, control=list(seed=173))
repu_topics_10=tidy(repu_lda_10, matrix="beta")

repu_top_terms_10 <- repu_topics_10 %>%
  group_by(topic) %>%
  top_n(15, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
terms(repu_lda_10, 3)
```

```
##      Topic 1      Topic 2    Topic 3     Topic 4      Topic 5
## [1,] "american"   "american" "americans" "government" "economy"
## [2,] "federal"    "people"   "economic"  "must"       "government"
## [3,] "government" "federal"  "can"       "people"     "tax"
##      Topic 6    Topic 7    Topic 8      Topic 9    Topic 10
## [1,] "will"     "american" "american"   "national" "must"
## [2,] "business" "federal"  "government" "people"   "federal"
## [3,] "american" "will"     "federal"    "federal"  "tax"
```

```r
# k=25
repu_lda_25=LDA(dtm_repu, k=25, control=list(seed=193))
repu_topics_25=tidy(repu_lda_25, matrix="beta")

repu_top_terms_25 <- repu_topics_25 %>%
  group_by(topic) %>%
```

```r
  top_n(15, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
terms(repu_lda_25, 3)
```

```
##      Topic 1      Topic 2    Topic 3      Topic 4      Topic 5
## [1,] "federal"    "american" "american"   "government" "american"
## [2,] "can"        "economy"  "government" "will"       "government"
## [3,] "government" "must"     "growth"     "can"        "tax"
##      Topic 6      Topic 7    Topic 8   Topic 9      Topic 10     Topic 11
## [1,] "federal"    "american" "growth"  "can"        "government" "people"
## [2,] "government" "economy"  "current" "government" "will"       "can"
## [3,] "american"   "people"   "people"  "america"    "can"        "america"
##      Topic 12  Topic 13   Topic 14  Topic 15     Topic 16   Topic 17
## [1,] "federal" "federal"  "federal" "federal"    "growth"   "american"
## [2,] "must"    "people"   "tax"     "will"       "economy"  "federal"
## [3,] "tax"     "business" "must"    "government" "american" "people"
##      Topic 18  Topic 19   Topic 20   Topic 21    Topic 22   Topic 23
## [1,] "federal" "must"     "federal"  "economy"   "federal"  "american"
## [2,] "will"    "american" "american" "will"      "american" "economy"
## [3,] "growth"  "will"     "people"   "president" "tax"      "tax"
##      Topic 24     Topic 25
## [1,] "government" "america"
## [2,] "will"       "government"
## [3,] "people"     "can"
```

```r
repu_top_terms_25 %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip()
```

```r
# Perplexity
perplexity(demo_lda)
```

```
## [1] 814.2498
```

```r
perplexity(demo_lda_10)
```

```
## [1] 814.4056
```

```r
perplexity(demo_lda_25)
```

```
## [1] 814.8643
```

```r
perplexity(repu_lda)
```

```
## [1] 1020.33
```

```r
perplexity(repu_lda_10)
```

```
## [1] 1020.607
```

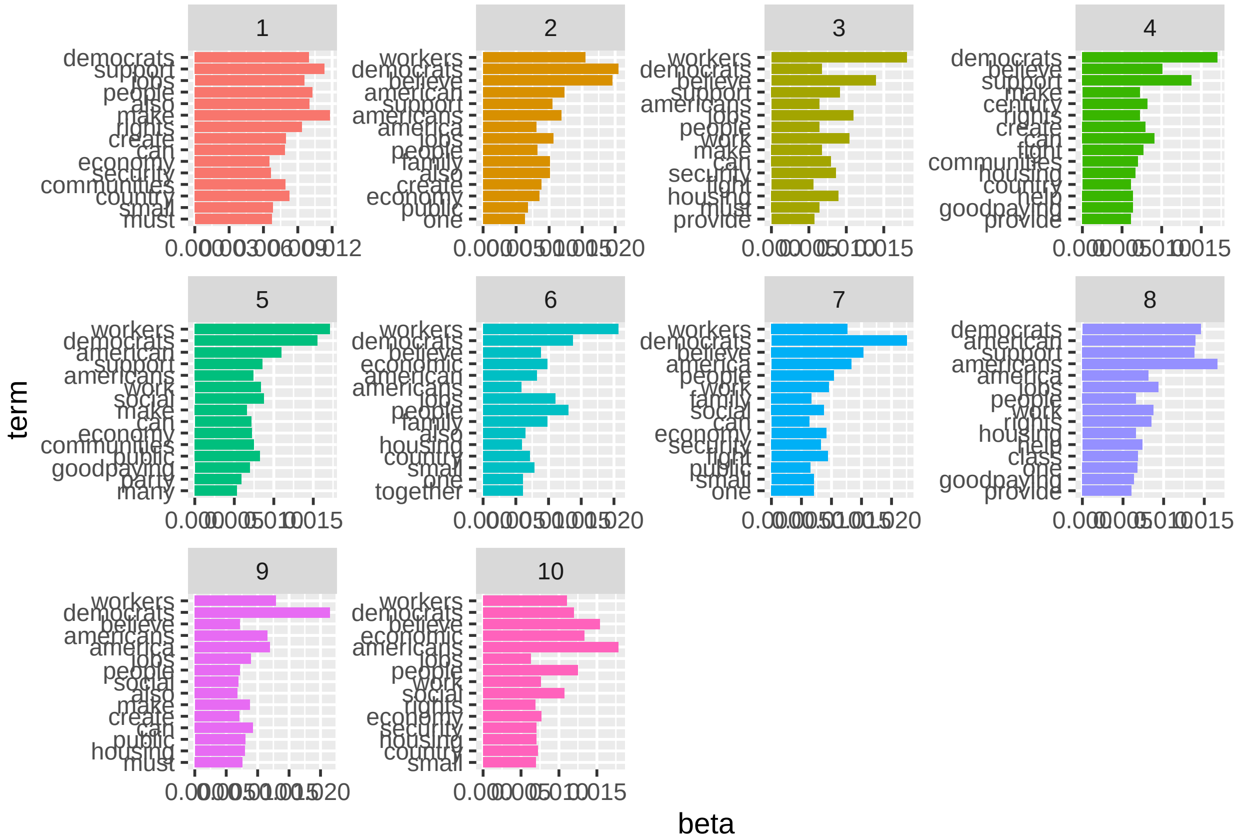
```r
perplexity(repu_lda_25)
```

```
## [1] 1021.021
```

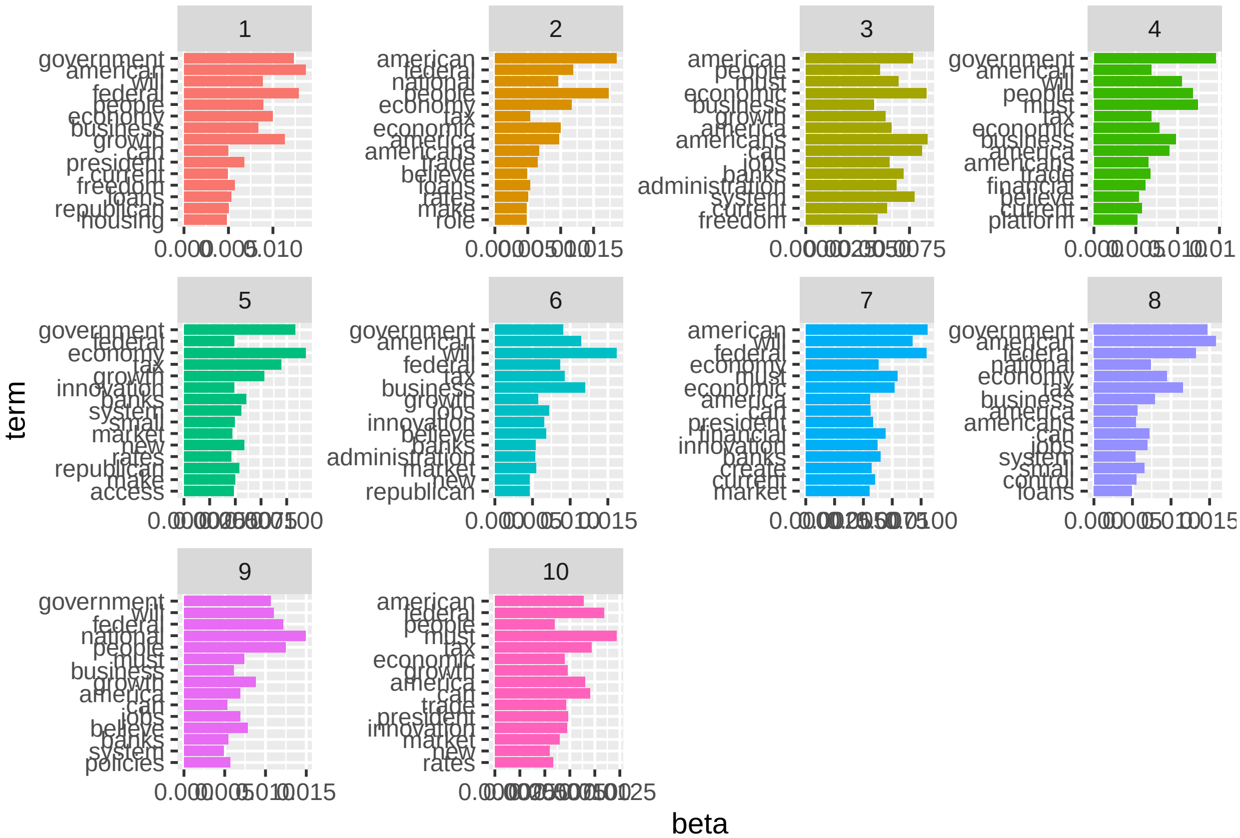
Technically, for democrate: k=25 fits best, for republican, k=5 fits best.

```r
# Democrat
# k=10
demo_top_terms_10 %>%
  mutate(term = reorder(term, beta)) %>%
```

```
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip()
```




```
# Republican
# k=10
repu_top_terms_10 %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip()
```



There are common topics from both of the parties, especially "economy" as a word appears frequently in models in each party. Also, some models in republican seems similar to the models in democrat, for example, model 5, 8, 9 of republican include "housing", "jobs", and models 6, 8, 10 in democrat also contain these words. I don't think k=10 is a very efficient way to see the difference bwteen partie, bucause we can see some over-clustering in the results here, some topic models in the results are very similar, have the same highest beta-loading word, some of these models should be combined together to offer a more concise set of models.

Conclution

I would support democrat in 2020 based on the results from the analysis above. Because from the democrat uses more postive words, suggesting they are more optimistic compared to republican, and I believe a more optimistic party would employ policies that are more active and positive to the society. Another reason is that from the topic modelling, democrat seems to pay more attention to communities, support, family, workers, these words show that it cares more about ordinary people in the country, which gives me a sence that electing democrat would be helpful to shrink the increasing gap between the rich and the poor. Instead, republican uses more words like business, market, financial, loans etc, it seems they try to stimulate the national economy by increasing business and trade, which is also good, but personally I prefer the overall tones of demorat.