Intro to text analysis

Rongbo Jin

2023-04-17

```r
# install packages
# installed.packages(c("quanteda", "ggplot2", "dplyr", "smart", "quanteda.dictionaries", "quanteda.text

# load packages
library(quanteda)
```

```
## Package version: 3.2.1
## Unicode version: 13.0
## ICU version: 69.1

## Parallel computing: 4 of 4 threads used.

## See https://quanteda.io for tutorials and examples.
```

```r
library(ggplot2)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(quanteda.dictionaries)
library(quanteda.textplots)
library(topicmodels)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2
## --

## v tibble  3.2.1     v purrr   0.3.4
## v tidyr   1.2.0     v stringr 1.5.0
## v readr   2.1.2     v forcats 0.5.1

## Warning: package 'tibble' was built under R version 4.2.3

## Warning: package 'stringr' was built under R version 4.2.3

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(tidyr)
library(tidytext)
```

```r
# load data
df <- read.csv("platforms_1928-2020.csv")
table(df$years)
```

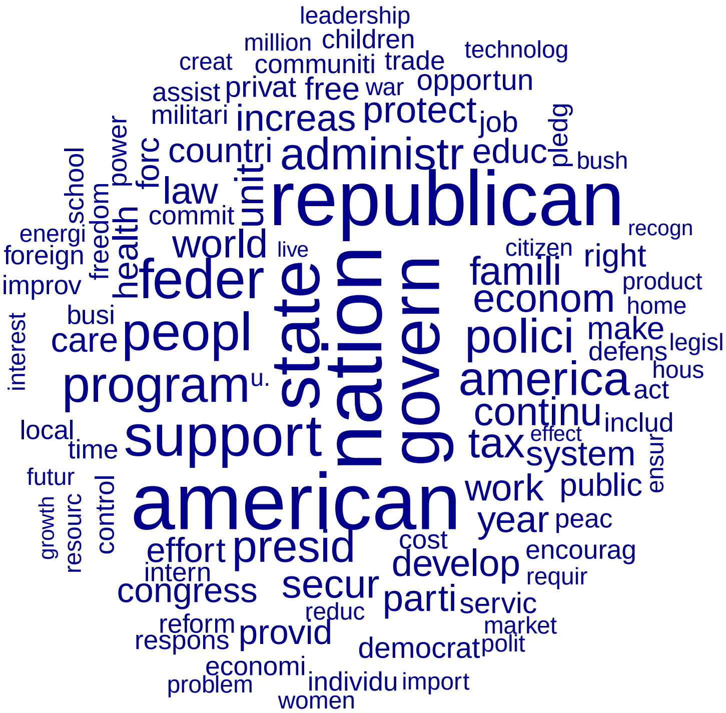
```
##
## 1928 1932 1936 1940 1944 1948 1952 1956 1960 1964 1968 1972 1976 1980 1984 1988
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 1992 1996 2000 2004 2008 2012 2016
##    1    1    1    1    1    1    1
```

```r
# create corpus and preprocessing
corpus1 <- corpus(df,
                  text_field = "textR") %>%
  tokens(remove_punct = TRUE, # remove punctuation
         remove_numbers = TRUE, # remove numbers
         remove_symbols = TRUE, # remove special symbols
         padding = TRUE) %>%
  tokens_remove(c(quanteda::stopwords(language = "en", source = "smart")), padding = TRUE) %>% # remove
  tokens_tolower() %>% # make all words into lower cases
  tokens_wordstem()
```

```r
# create DFM
# trim DFM, delete words with counts less than 10 or more than 1000
dfm1 <- dfm(corpus1) %>%
  dfm_trim(min_termfreq = 1, max_termfreq = 10000)
nfeat(dfm1)
```

```
## [1] 10134
```

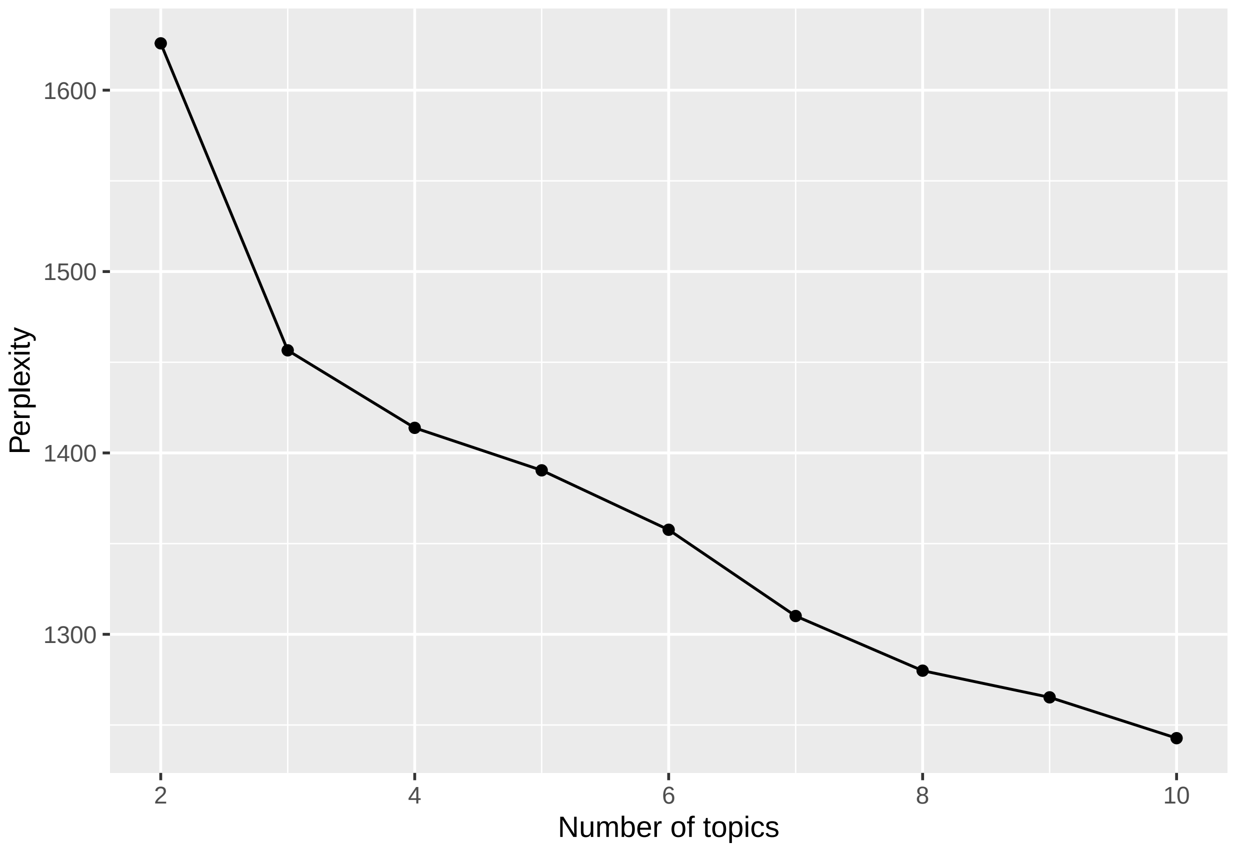
```r
# make a word cloud
set.seed(132)
textplot_wordcloud(dfm1, max_words = 100)
```



```r
## LDA Model ##
dtm1 <- quanteda::convert(dfm1, to = "topicmodels")
```

```
## as(<dgCMatrix>, "dgTMatrix") is deprecated since Matrix 1.5-0; do as(., "TsparseMatrix") instead
```

```r
# fitting an LDA model is determining the size of k.
n_topics <- c(2, 3, 4, 5, 6, 7, 8, 9, 10)
lda_compare1 <- n_topics %>%
  map(LDA, x = dfm1, control = list(seed = 1109))
tibble(k = n_topics,
       perplex = map_dbl(lda_compare1, perplexity)) %>%
  ggplot(aes(k, perplex)) +
  geom_point() +
  geom_line() +
  labs(# title = "Evaluating LDA topic models",
    # subtitle = "Optimal number of topics (smaller is better)",
    x = "Number of topics",
    y = "Perplexity")
```



```r
### K=7
set.seed(122)
lda1 <- LDA(dtm1, method = "Gibbs", k = 7, control = list(alpha = 0.1))
terms(lda1, 10)
```

```
##       Topic 1      Topic 2      Topic 3     Topic 4      Topic 5
##  [1,] "soviet"     "nation"     "famili"    "nation"     "presid"
##  [2,] "polici"     "administr"  "ensur"     "american"   "bush"
##  [3,] "republican" "govern"     "american"  "state"      "republican"
##  [4,] "percent"    "parti"      "care"      "govern"     "america"
##  [5,] "democrat"   "polici"     "reform"    "support"    "health"
##  [6,] "famili"     "favor"      "america"   "feder"      "congress"
##  [7,] "carter"     "pledg"      "u."        "republican" "support"
##  [8,] "u."         "public"     "technolog" "peopl"      "terror"
##  [9,] "reagan"     "republican" "protect"   "tax"        "applaud"
## [10,] "inflat"     "agricultur" "law"       "program"    "terrorist"
##       Topic 6     Topic 7
##  [1,] "program"   "current"
##  [2,] "continu"   "constitut"
##  [3,] "pledg"     "call"
##  [4,] "develop"   "advanc"
##  [5,] "area"      "healthcar"
##  [6,] "peac"      "energi"
##  [7,] "communist" "regul"
##  [8,] "improv"    "govern"
##  [9,] "vigor"     "right"
## [10,] "assur"     "u."
```

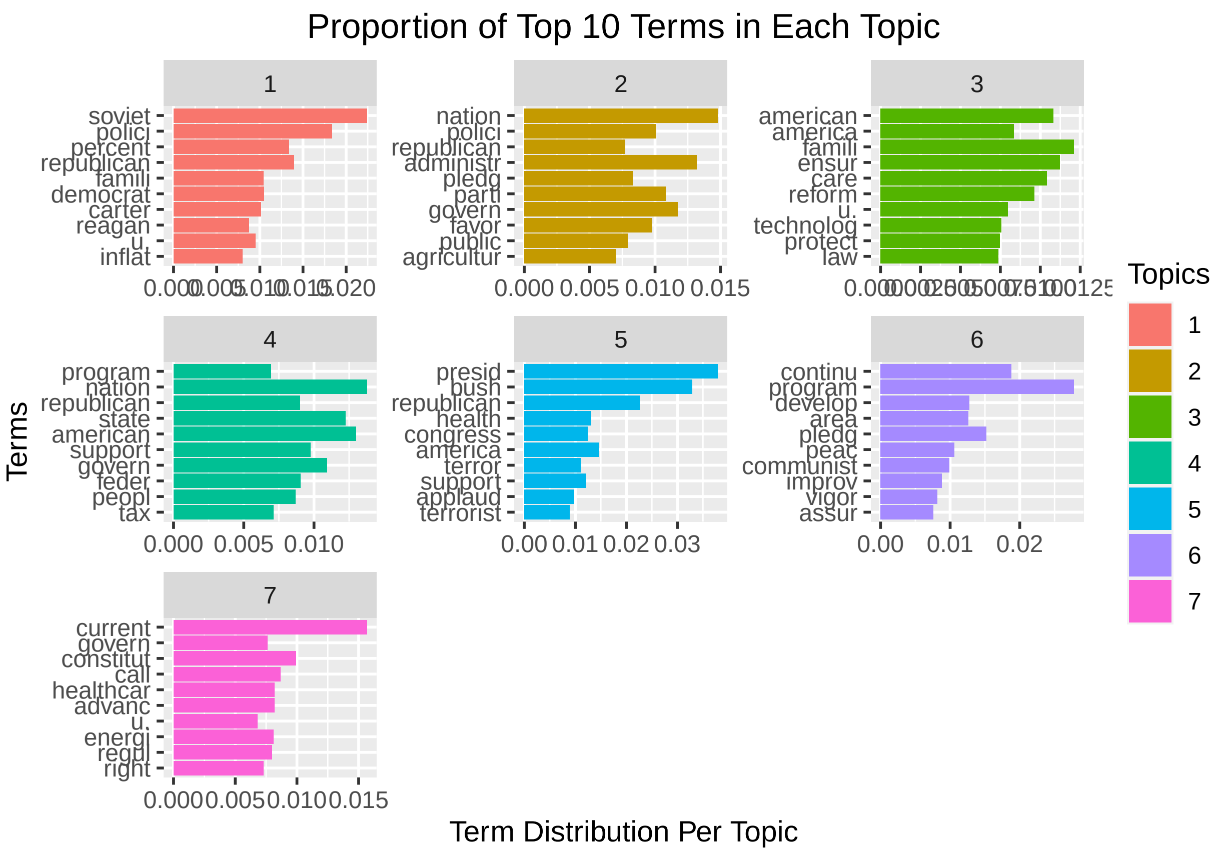
```
topics1 <- tidy(lda1, matrix="beta")
head(topics1, 10)
```

```
## # A tibble: 10 x 3
##    topic term             beta
##    <int> <chr>           <dbl>
##  1     1 republican 0.0139
##  2     2 republican 0.00770
##  3     3 republican 0.00620
##  4     4 republican 0.00898
##  5     5 republican 0.0226
##  6     6 republican 0.00000957
##  7     7 republican 0.0000115
##  8     1 parti      0.00136
##  9     2 parti      0.0108
## 10     3 parti      0.00000360
```

```
# Reshape this by grouping values by topic
top_terms1 <- topics1 %>%
  group_by(topic) %>%
  top_n(10, beta) %>% # We just keep top 20 words
  ungroup() %>%
  arrange(topic, -beta) # We arrange them by descending beta values
# Let's see what this looks like
head(top_terms1, 25)
```

```
## # A tibble: 25 x 3
##    topic term         beta
##    <int> <chr>       <dbl>
##  1     1 soviet     0.0224
##  2     1 polici     0.0184
##  3     1 republican 0.0139
##  4     1 percent    0.0134
##  5     1 democrat   0.0105
##  6     1 famili     0.0104
##  7     1 carter     0.0101
##  8     1 u.         0.00946
##  9     1 reagan     0.00875
## 10     1 inflat     0.00797
## # i 15 more rows
```

```
top_terms1 %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = T) +
  labs(title = "Proportion of Top 10 Terms in Each Topic",
       x = "Terms", y = "Term Distribution Per Topic") +
  theme(plot.title = element_text(hjust = 0.5)) +
  guides(fill = guide_legend(title = "Topics", title.position = "top")) +
  # scale_fill_discrete(limits=c("1", "2", "3"), labels=c("")) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip()
```


Proportion of Top 10 Terms in Each Topic
1
soviet
polici
percent
republican
famili
democrat
carter
reagan
u.
inflat
0.000 0.005 0.010 0.015 0.020
2
nation
polici
republican
administr
pledg
parti
govern
favor
public
agricultur
0.000 0.005 0.010 0.015
3
american
america
famili
ensur
care
reform
u.
technolog
protect
law
0.0000 0.0025 0.0050 0.0075 0.0100 0.0125
4
program
nation
republican
state
american
support
govern
feder
peopl
tax
0.000 0.005 0.010
5
presid
bush
republican
health
congress
america
terror
support
applaud
terrorist
0.00 0.01 0.02 0.03
6
continu
program
develop
area
pledg
peac
communist
improv
vigor
assur
0.00 0.01 0.02
7
current
govern
constitut
call
healthcar
advanc
u.
energi
regul
right
0.000 0.005 0.010 0.015
Terms
Term Distribution Per Topic
Topics
1
2
3
4
5
6
7