Topic 6: Topic Analysis

Joe DeCesaro

2022-05-10

```
library(here)
library(pdftools)
library(quanteda)
library(tm)
library(topicmodels)
library(ldatuning)
library(tidyverse)
library(tidytext)
library(reshape2)
```

Load the data

```
##Topic 6 .Rmd here:https://raw.githubusercontent.com/MaRo406/EDS_231-text-sentiment/main/topic_6.Rmd
#grab data here:
comments_df<-read_csv("https://raw.githubusercontent.com/MaRo406/EDS_231-text-sentiment/main/dat/commen

#comments_df <- read_csv(here("dat", "comments_df.csv")) #if reading from local
```

Now we'll build and clean the corpus

```
epa_corp <- corpus(x = comments_df, text_field = "text")
epa_corp.stats <- summary(epa_corp)
head(epa_corp.stats, n = 25)
```

```
##      Text Types Tokens Sentences
## 1   text1  1196   3973       178
## 2   text2   830   2509       111
## 3   text3   279    571        31
## 4   text4  1745   6904       251
## 5   text5   581   1534        49
## 6   text6   469   1187        53
## 7   text7   424    903        38
## 8   text8  3622  22270       655
## 9   text9   373    717        25
## 10 text10   404    971        42
## 11 text11   710   2190        77
## 12 text12   636   1896        82
## 13 text13   146    206         3
## 14 text14  1124   3197        86
## 15 text15   914   2943        90
```

```
## 16 text16    13     45         1
## 17 text17  1043   3190       103
## 18 text18   313    601        24
## 19 text19   152    229         6
## 20 text20   341    786        35
## 21 text21   211    403        15
## 22 text22   186    322        12
## 23 text23   211    398        14
## 24 text24   325    696        33
## 25 text25  1749   5382       115
##                                                 Document
## 1                                     1_Air Alliance.pdf
## 2                                         10_Bus NEJ.pdf
## 3                                    11_Carlton Ginny.pdf
## 4                                     15_City Project.pdf
## 5                                    16_Corporate EEC.pdf
## 6                                  17_Detriot Sierra Club.pdf
## 7                                        18_District DOE.pdf
## 8                                    19_Earth Justice.pdf
## 9                                         2_Alex Kidd.pdf
## 10                                  20_Elizabeth Mooney.pdf
## 11                                           21_Env COS.pdf
## 12                                      22_Env Def Fund.pdf
## 13                                    23_Env Health Watch.pdf
## 14 24_Env Justice Leadership Forum on Climate Change.pdf
## 15                                    25_Env Law at Duke.pdf
## 16                                     26_Farm worker AF.pdf
## 17                                  27_Farm Worker Justice.pdf
## 18                                       28_Faulker County.pdf
## 19                                        29_First Peoples.pdf
## 20                                   3_Alliance for Metro.pdf
## 21                                          30_Gage Blasi.pdf
## 22                                           31_Gull Leon.pdf
## 23                                       32_Hilary Kramer.pdf
## 24                                    33_Housing Land Advoc.pdf
## 25                                          34_Human rights.pdf
```

```
toks <- tokens(epa_corp, remove_punct = TRUE, remove_numbers = TRUE)
#I added some project-specific stop words here
add_stops <- c(stopwords("en"),"environmental", "justice", "ej", "epa", "public", "comment")
toks1 <- tokens_select(toks, pattern = add_stops, selection = "remove")
```

And now convert to a document-feature matrix

```
dfm_comm<- dfm(toks1, tolower = TRUE)
dfm <- dfm_wordstem(dfm_comm)
dfm <- dfm_trim(dfm, min_docfreq = 2) #remove terms only appearing in one doc (min_termfreq = 10)

print(head(dfm))
```

```
## Document-feature matrix of: 6 documents, 2,781 features (82.75% sparse) and 1 docvar.
##        features
## docs    charl lee deputi associ assist administr usepa offic 2201-a
```

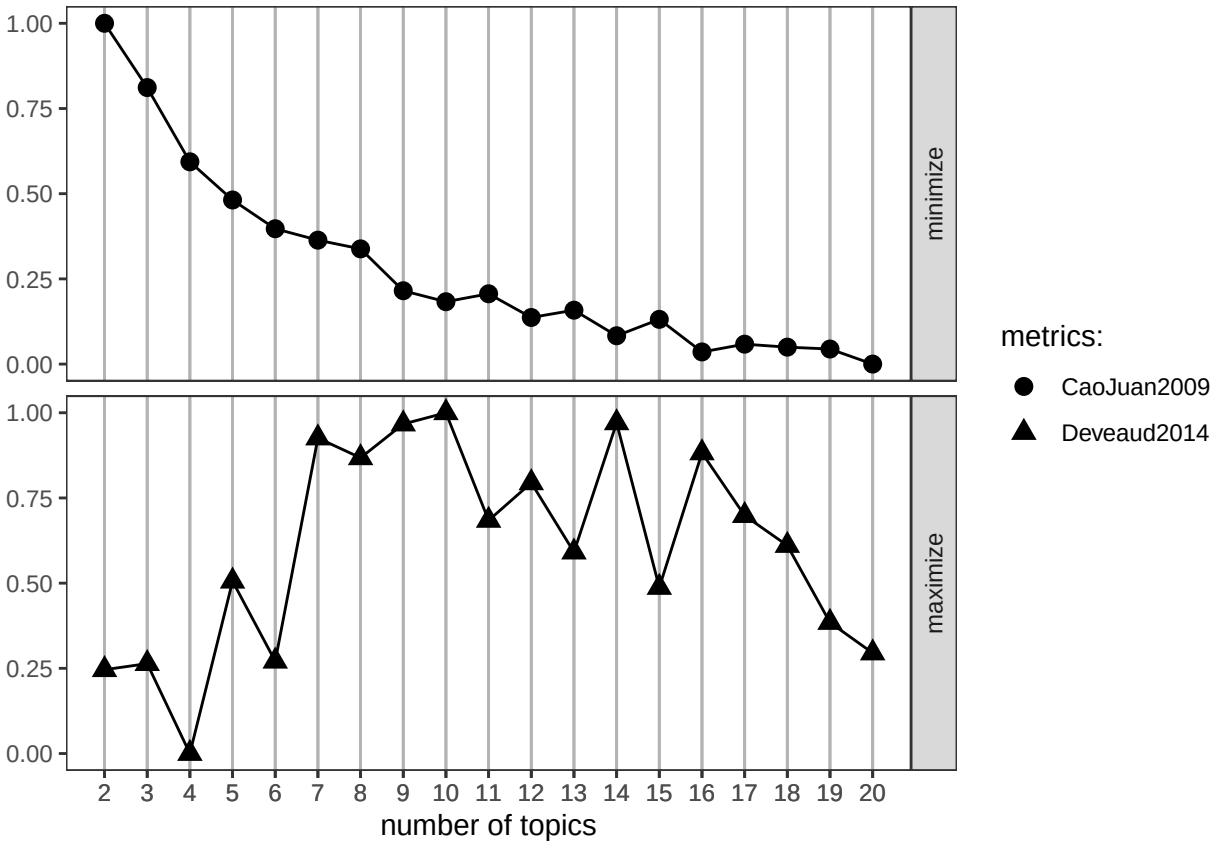
```
##   text1     1   2      1      1      6         6     1     7      1
##   text2     1   1      1      4      3         1     0     5      0
##   text3     0   0      0      0      1         0     0     2      0
##   text4     0   0      0      0      1         9     0     1      0
##   text5     4   5      1      1      1         1     0     1      1
##   text6     1   1      1      3      1         3     0     4      0
##        features
## docs    pennsylvania
##   text1            1
##   text2            0
##   text3            0
##   text4            0
##   text5            1
##   text6            0
## [ reached max_nfeat ... 2,771 more features ]
```

```
#remove rows (docs) with all zeros
sel_idx <- slam::row_sums(dfm) > 0
dfm <- dfm[sel_idx, ]
#comments_df <- dfm[sel_idx, ]
```

```
#
result <- FindTopicsNumber(
  dfm,
  topics = seq(from = 2, to = 20, by = 1),
  metrics = c("CaoJuan2009",  "Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 77),
  verbose = TRUE
)
```

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

```
FindTopicsNumber_plot(result)
```



Assignment:

Run three more models and select the overall best value for k (the number of topics) - include some justification for your selection: theory, FindTopicsNumber() optimization metrics, interpretability, LDAvis

Model 1 (k = 10)

```
k <- 10

topicModel_k10 <- LDA(dfm, k,
                      method = "Gibbs",
                      control = list(iter = 500, verbose = 25))
```

```
## K = 10; V = 2781; M = 77
## Sampling 500 iterations!
## Iteration 25 ...
## Iteration 50 ...
## Iteration 75 ...
## Iteration 100 ...
## Iteration 125 ...
## Iteration 150 ...
## Iteration 175 ...
## Iteration 200 ...
## Iteration 225 ...
```

```
## Iteration 250 ...
## Iteration 275 ...
## Iteration 300 ...
## Iteration 325 ...
## Iteration 350 ...
## Iteration 375 ...
## Iteration 400 ...
## Iteration 425 ...
## Iteration 450 ...
## Iteration 475 ...
## Iteration 500 ...
## Gibbs sampling completed!
```

```
tmResult <- posterior(topicModel_k10)
terms(topicModel_k10, 10)
```

```
##       Topic 1    Topic 2   Topic 3    Topic 4     Topic 5    Topic 6
##  [1,] "state"    "health"  "permit"   "communiti" "prison"   "framework"
##  [2,] "rule"     "peopl"   "state"    "local"     "facil"    "draft"
##  [3,] "popul"    "right"   "consid"   "plan"      "popul"    "effort"
##  [4,] "ejscreen" "park"    "use"      "comment"   "industri" "agenc"
##  [5,] "also"     "citi"    "organ"    "particip"  "energi"   "state"
##  [6,] "asthma"   "includ"  "feder"    "govern"    "project"  "epa"
##  [7,] "health"   "project" "meet"     "collabor"  "center"   "develop"
##  [8,] "air"      "climat"  "air"      "juli"      "sourc"    "water"
##  [9,] "avail"    "law"     "polici"   "develop"   "site"     "comment"
## [10,] "must"     "green"   "carolina" "agenda"    "gas"      "tool"
##       Topic 7     Topic 8    Topic 9   Topic 10
##  [1,] "communiti" "work"     "program" "enforc"
##  [2,] "pollut"    "make"     "agenc"   "includ"
##  [3,] "impact"    "need"     "state"   "action"
##  [4,] "clean"     "help"     "feder"   "monitor"
##  [5,] "also"      "subject"  "titl"    "communiti"
##  [6,] "protect"   "year"     "vi"      "permit"
##  [7,] "air"       "strategi" "issu"    "comment"
##  [8,] "health"    "thank"    "right"   "report"
##  [9,] "plan"      "can"      "act"     "data"
## [10,] "comment"   "action"   "civil"   "assess"
```

```
theta <- tmResult$topics
beta <- tmResult$terms # probability of each term in each topic
vocab <- (colnames(beta))
```

```
comment_topics <- tidy(topicModel_k10, matrix = "beta")

top_terms <- comment_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
top_terms
```

```
## # A tibble: 102 x 3
```

```
##    topic term        beta
##    <int> <chr>      <dbl>
##  1     1 state    0.0141
##  2     1 rule     0.0139
##  3     1 popul    0.0126
##  4     1 ejscreen 0.0110
##  5     1 also     0.0108
##  6     1 asthma   0.0107
##  7     1 health   0.0104
##  8     1 air      0.00942
##  9     1 avail    0.00942
## 10     1 must     0.00920
## # ... with 92 more rows
```

```
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()
```


```
top5termsPerTopic <- terms(topicModel_k10, 5)
topicNames_k10 <- apply(top5termsPerTopic, 2, paste, collapse=" ")
```

Model 2 (k = 5)

```
k <- 5

topicModel_k5 <- LDA(dfm, k,
                     method = "Gibbs",
                     control = list(iter = 500, verbose = 25))
```

```
## K = 5; V = 2781; M = 77
## Sampling 500 iterations!
## Iteration 25 ...
## Iteration 50 ...
## Iteration 75 ...
## Iteration 100 ...
## Iteration 125 ...
## Iteration 150 ...
## Iteration 175 ...
## Iteration 200 ...
## Iteration 225 ...
## Iteration 250 ...
## Iteration 275 ...
## Iteration 300 ...
## Iteration 325 ...
## Iteration 350 ...
## Iteration 375 ...
## Iteration 400 ...
## Iteration 425 ...
## Iteration 450 ...
## Iteration 475 ...
## Iteration 500 ...
## Gibbs sampling completed!
```

```
tmResult <- posterior(topicModel_k5)
terms(topicModel_k5, 10)
```

```
##       Topic 1    Topic 2     Topic 3   Topic 4     Topic 5
##  [1,] "state"    "communiti" "right"   "state"     "communiti"
##  [2,] "prison"   "enforc"    "civil"   "pollut"    "framework"
##  [3,] "permit"   "includ"    "health"  "impact"    "draft"
##  [4,] "comment"  "action"    "peopl"   "popul"     "state"
##  [5,] "like"     "comment"   "vi"      "health"    "effort"
##  [6,] "consid"   "permit"    "feder"   "rule"      "agenc"
##  [7,] "use"      "requir"    "agenc"   "communiti" "agenda"
##  [8,] "make"     "monitor"   "titl"    "also"      "action"
##  [9,] "grant"    "air"       "citi"    "guidanc"   "develop"
## [10,] "carolina" "region"    "address" "provid"    "will"
```

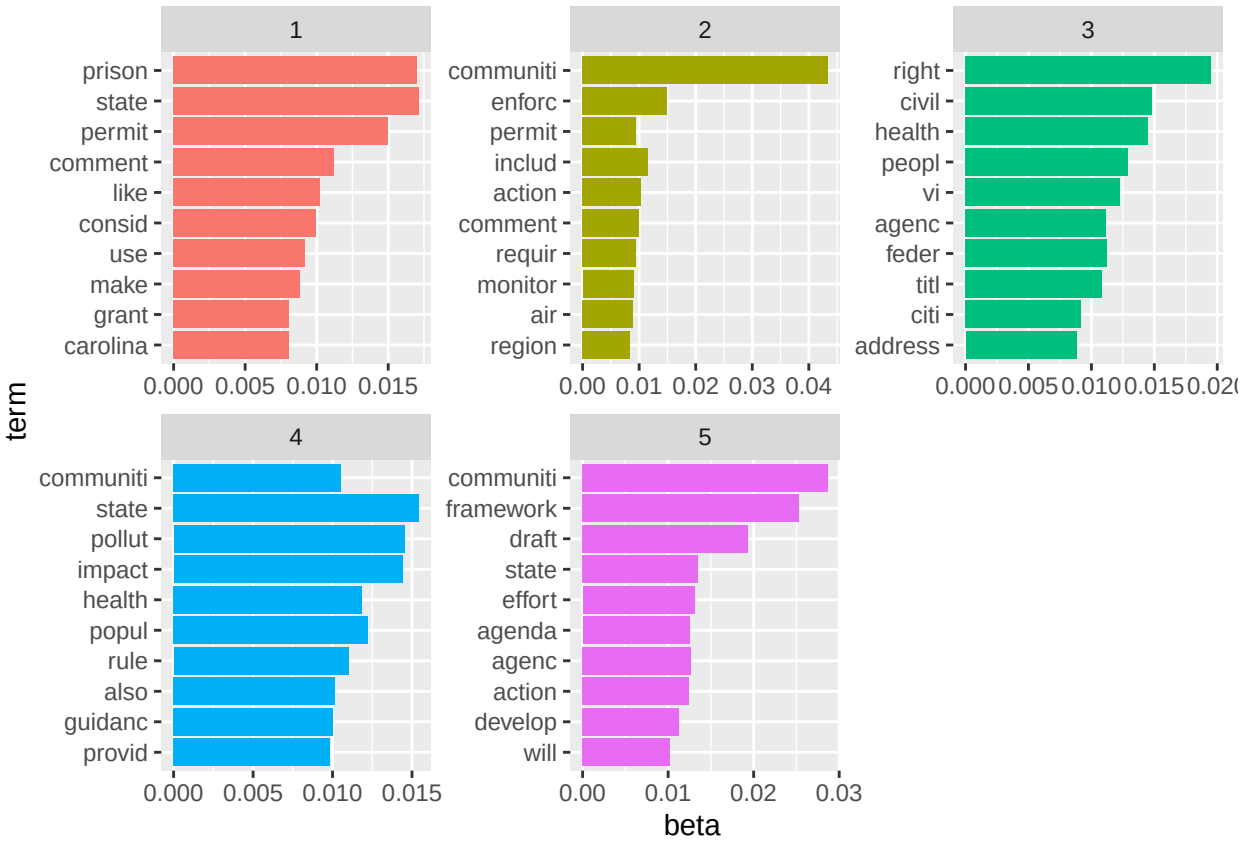
```
theta <- tmResult$topics
beta <- tmResult$terms # probability of each term in each topic
vocab <- (colnames(beta))
```

```r
comment_topics <- tidy(topicModel_k5, matrix = "beta")

top_terms <- comment_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
top_terms
```

```
## # A tibble: 50 x 3
##    topic term        beta
##    <int> <chr>      <dbl>
##  1     1 state    0.0171
##  2     1 prison   0.0170
##  3     1 permit   0.0150
##  4     1 comment  0.0112
##  5     1 like     0.0102
##  6     1 consid   0.00996
##  7     1 use      0.00921
##  8     1 make     0.00883
##  9     1 grant    0.00807
## 10     1 carolina 0.00807
## # ... with 40 more rows
```

```r
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()
```



Model 3 (k = 14)

```
k <- 14

topicModel_k14 <- LDA(dfm, k,
                      method = "Gibbs",
                      control = list(iter = 500, verbose = 25))
```

```
## K = 14; V = 2781; M = 77
## Sampling 500 iterations!
## Iteration 25 ...
## Iteration 50 ...
## Iteration 75 ...
## Iteration 100 ...
## Iteration 125 ...
## Iteration 150 ...
## Iteration 175 ...
## Iteration 200 ...
## Iteration 225 ...
## Iteration 250 ...
## Iteration 275 ...
## Iteration 300 ...
## Iteration 325 ...
## Iteration 350 ...
## Iteration 375 ...
```

```
## Iteration 400 ...
## Iteration 425 ...
## Iteration 450 ...
## Iteration 475 ...
## Iteration 500 ...
## Gibbs sampling completed!
```

```
tmResult <- posterior(topicModel_k14)
terms(topicModel_k14, 10)
```

```
##       Topic 1     Topic 2    Topic 3     Topic 4     Topic 5     Topic 6
##  [1,] "communiti" "rule"     "agenc"     "communiti" "framework" "communiti"
##  [2,] "comment"   "impact"   "right"     "comment"   "agenc"     "enforc"
##  [3,] "pollut"    "health"   "civil"     "also"      "draft"     "monitor"
##  [4,] "polici"    "pollut"   "issu"      "process"   "action"    "action"
##  [5,] "energi"    "state"    "titl"      "use"       "state"     "complianc"
##  [6,] "will"      "ejscreen" "vi"        "provid"    "develop"   "includ"
##  [7,] "new"       "asthma"   "plan"      "impact"    "epa"       "air"
##  [8,] "clean"     "popul"    "act"       "exampl"    "effort"    "pollut"
##  [9,] "emiss"     "agenc"    "communiti" "mani"      "advanc"    "assess"
## [10,] "reduct"    "also"     "state"     "will"      "within"    "report"
##       Topic 7   Topic 8     Topic 9    Topic 10   Topic 11  Topic 12
##  [1,] "program" "farmwork"  "permit"   "plan"     "health"  "can"
##  [2,] "feder"   "work"      "state"    "agenda"   "park"    "need"
##  [3,] "state"   "water"     "consid"   "action"   "right"   "air"
##  [4,] "regul"   "pesticid"  "carolina" "work"     "citi"    "environ"
##  [5,] "requir"  "use"       "grant"    "strategi" "law"     "peopl"
##  [6,] "epa"     "health"    "air"      "way"      "peopl"   "qualiti"
##  [7,] "includ"  "exposur"   "opportun" "will"     "project" "communiti"
##  [8,] "polici"  "econom"    "feder"    "comment"  "green"   "area"
##  [9,] "draft"   "individu"  "comment"  "subject"  "civil"   "scienc"
## [10,] "comment" "implement" "use"      "like"     "climat"  "help"
##       Topic 13    Topic 14
##  [1,] "communiti" "prison"
##  [2,] "local"     "center"
##  [3,] "govern"    "popul"
##  [4,] "plan"      "facil"
##  [5,] "land"      "sourc"
##  [6,] "resourc"   "project"
##  [7,] "use"       "water"
##  [8,] "can"       "report"
##  [9,] "develop"   "site"
## [10,] "engag"     "peopl"
```

```
theta <- tmResult$topics
beta <- tmResult$terms # probability of each term in each topic
vocab <- (colnames(beta))
```

```
comment_topics <- tidy(topicModel_k14, matrix = "beta")

top_terms <- comment_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
```

```
  ungroup() %>%
  arrange(topic, -beta)
top_terms
```

```
## # A tibble: 148 x 3
##    topic term        beta
##    <int> <chr>      <dbl>
##  1     1 communiti 0.0382
##  2     1 comment   0.0230
##  3     1 pollut    0.0224
##  4     1 polici    0.0207
##  5     1 energi    0.0165
##  6     1 will      0.0151
##  7     1 new       0.0145
##  8     1 clean     0.0145
##  9     1 emiss     0.0125
## 10     1 reduct    0.0115
## # ... with 138 more rows
```

```
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()
```

Best k-value

The Deveaud method shows noticeable spikes at 5, 10, and 14 topics so I chose to explore these beyond the 7 and 9 topics we did in class. I based my choices on the Deveaud method because it does not just improve with the more topics you choose.

The model where k=14 seems to be over fit as there are low beta values and multiple overlapping top words like communiti and state. For k=5 4/5 of the topics have the top word as communiti with low distinctiveness. When k=10 there is still low beta values however there only 4/10 categories have the same top word. To me this seems to mean we get more distinctive groups without spliting things up too much so I would go with k=10 as the best metric of the three k values I chose.