Topic 6: Topic Analysis

Peter Menzies

5/8/22

```
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)

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

Load the data

```
comments_df <- readRDS(here("data", "comments_df.RDS"))
```

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, ]
```

We somehow have to come up with a value for k,the number of latent topics present in the data. How do we do this? There are multiple methods. Let's use what we already know about the data to inform a prediction. The EPA has 9 priority areas: Rulemaking, Permitting, Compliance and Enforcement, Science, States and Local Governments, Federal Agencies, Community-based Work, Tribes and Indigenous People, National Measures. Maybe the comments correspond to those areas?

```
k <- 9

topicModel_k9 <- LDA(dfm, k, method="Gibbs", control=list(iter = 500, verbose = FALSE))
#nTerms(dfm_comm)

tmResult <- posterior(topicModel_k9)
attributes(tmResult)
```

```
## $names
## [1] "terms"  "topics"
```

```
#nTerms(dfm_comm)
beta <- tmResult$terms   # get beta from results
dim(beta)                # K distributions over nTerms(DTM) terms# lengthOfVocab
```

```
## [1]    9 2781
```

```
terms(topicModel_k9, 10)
```

```
##       Topic 1     Topic 2      Topic 3     Topic 4   Topic 5   Topic 6
##  [1,] "peopl"     "communiti" "impact"    "agenc"   "program" "state"
##  [2,] "health"    "framework" "pollut"    "issu"    "polici"  "permit"
##  [3,] "communiti" "draft"     "state"     "right"   "state"   "use"
##  [4,] "citi"      "action"    "health"    "address" "feder"   "consid"
##  [5,] "park"      "develop"   "rule"      "titl"    "epa"     "feder"
##  [6,] "see"       "effort"    "communiti" "vi"      "requir"  "grant"
##  [7,] "comment"   "agenda"     "popul"     "civil"   "regul"   "implement"
##  [8,] "climat"    "state"      "also"      "includ"  "may"     "opportun"
##  [9,] "access"    "comment"    "ejscreen"  "plan"    "follow"  "organ"
## [10,] "includ"    "overburden" "air"       "commit"  "affect"  "particip"
##       Topic 7         Topic 8     Topic 9
##  [1,] "water"         "communiti" "prison"
##  [2,] "can"           "enforc"    "popul"
##  [3,] "work"          "comment"   "facil"
##  [4,] "farmwork"      "monitor"   "project"
##  [5,] "health"        "action"    "new"
##  [6,] "econom"        "includ"    "subject"
##  [7,] "pesticid"      "permit"    "strategi"
##  [8,] "infrastructur" "complianc" "like"
##  [9,] "clean"         "assess"    "know"
## [10,] "e.g"           "pollut"    "lung"
```

Some of those topics seem related to the cross-cutting and additional topics identified in the EPA's response to the public comments:

1. Title VI of the Civil Rights Act of 1964

2.EJSCREEN

3. climate change, climate adaptation and promoting greenhouse gas reductions co-benefits

4. overburdened communities and other stakeholders to meaningfully, effectively, and transparently participate in aspects of EJ 2020, as well as other agency processes

5. utilize multiple Federal Advisory Committees to better obtain outside environmental justice perspectives

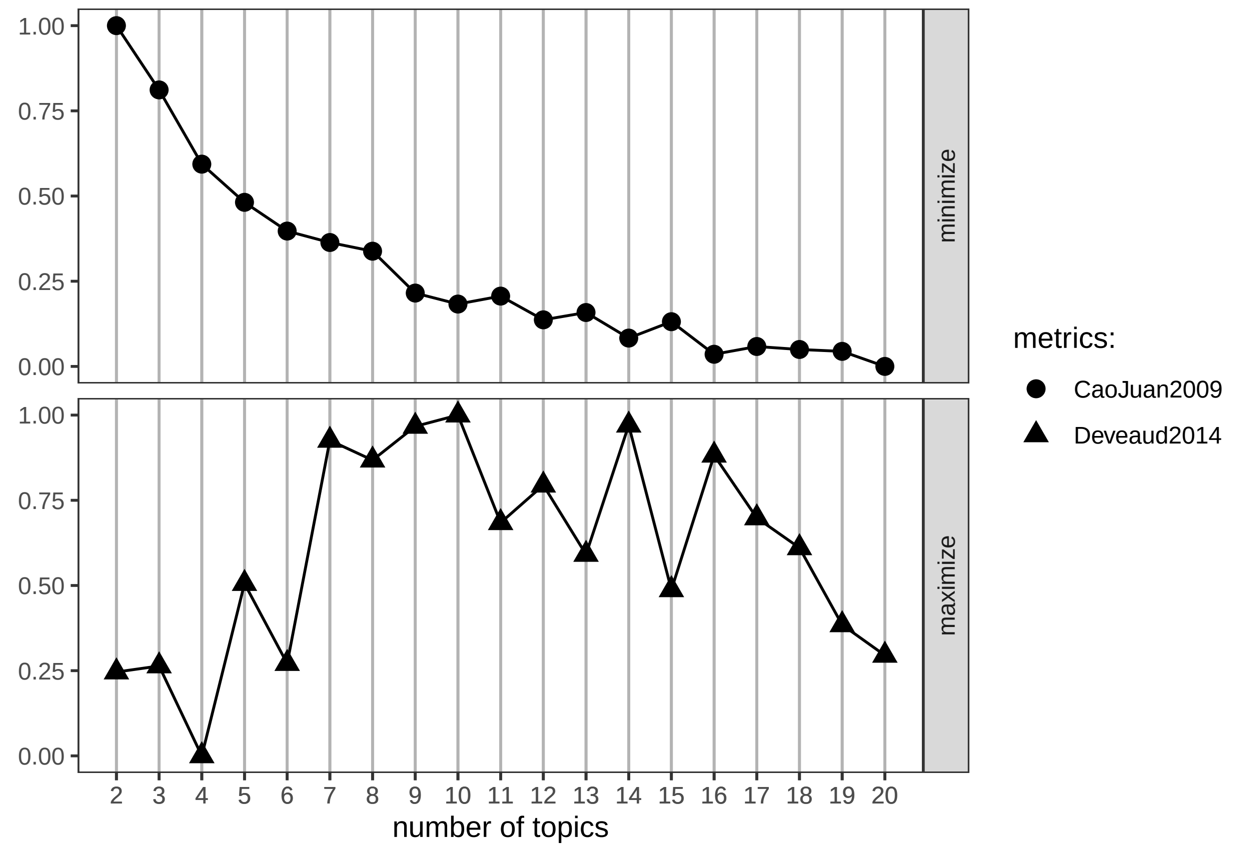
6. environmental justice and area-specific training to EPA staff

7. air quality issues in overburdened communities

So we could guess that there might be a 16 topics (9 priority + 7 additional). Or we could calculate some metrics from the data.

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

FindTopicsNumber_plot(result)
```



```
k <- 7

topicModel_k7 <- LDA(dfm, k, method="Gibbs", control=list(iter = 500, verbose = FALSE))

tmResult <- posterior(topicModel_k7)
terms(topicModel_k7, 10)
```

```
##       Topic 1   Topic 2      Topic 3    Topic 4      Topic 5     Topic 6
##  [1,] "agenc"   "communiti"  "state"    "prison"     "communiti" "pollut"
##  [2,] "program" "framework"  "permit"   "health"     "enforc"    "state"
##  [3,] "state"   "effort"     "consid"   "project"    "includ"    "health"
##  [4,] "epa"     "develop"    "air"      "citi"       "action"    "impact"
##  [5,] "issu"    "action"     "comment"  "peopl"      "monitor"   "popul"
##  [6,] "feder"   "plan"       "feder"    "california" "data"      "communiti"
##  [7,] "right"   "draft"      "organ"    "park"       "air"       "also"
##  [8,] "titl"    "agenda"     "grant"    "nation"     "permit"    "rule"
##  [9,] "work"    "overburden" "carolina" "see"        "need"      "air"
## [10,] "civil"   "process"    "use"      "center"     "comment"   "provid"
##       Topic 7
##  [1,] "comment"
##  [2,] "water"
##  [3,] "work"
##  [4,] "can"
##  [5,] "communiti"
##  [6,] "site"
##  [7,] "make"
##  [8,] "energi"
##  [9,] "area"
## [10,] "need"
```

```r
theta <- tmResult$topics
beta <- tmResult$terms
vocab <- (colnames(beta))
```

There are multiple proposed methods for how to measure the best k value. You can go down the rabbit hole here

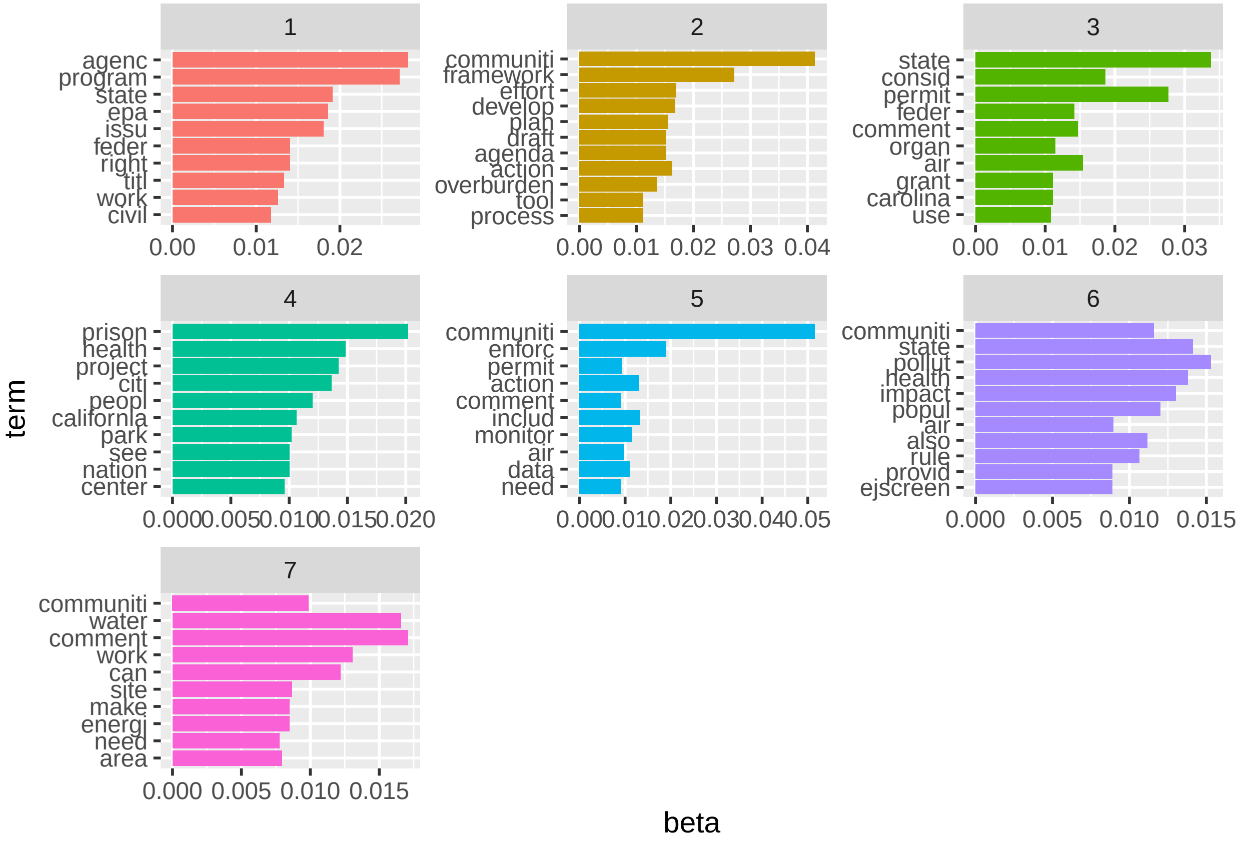
```r
comment_topics <- tidy(topicModel_k7, matrix = "beta")

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

top_terms
```

```
## # A tibble: 72 x 3
##    topic term      beta
##    <int> <chr>    <dbl>
##  1     1 agenc   0.0281
##  2     1 program 0.0271
##  3     1 state   0.0191
##  4     1 epa     0.0186
##  5     1 issu    0.0180
##  6     1 feder   0.0140
##  7     1 right   0.0140
##  8     1 titl    0.0133
##  9     1 work    0.0126
## 10     1 civil   0.0118
## # ... with 62 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()
```



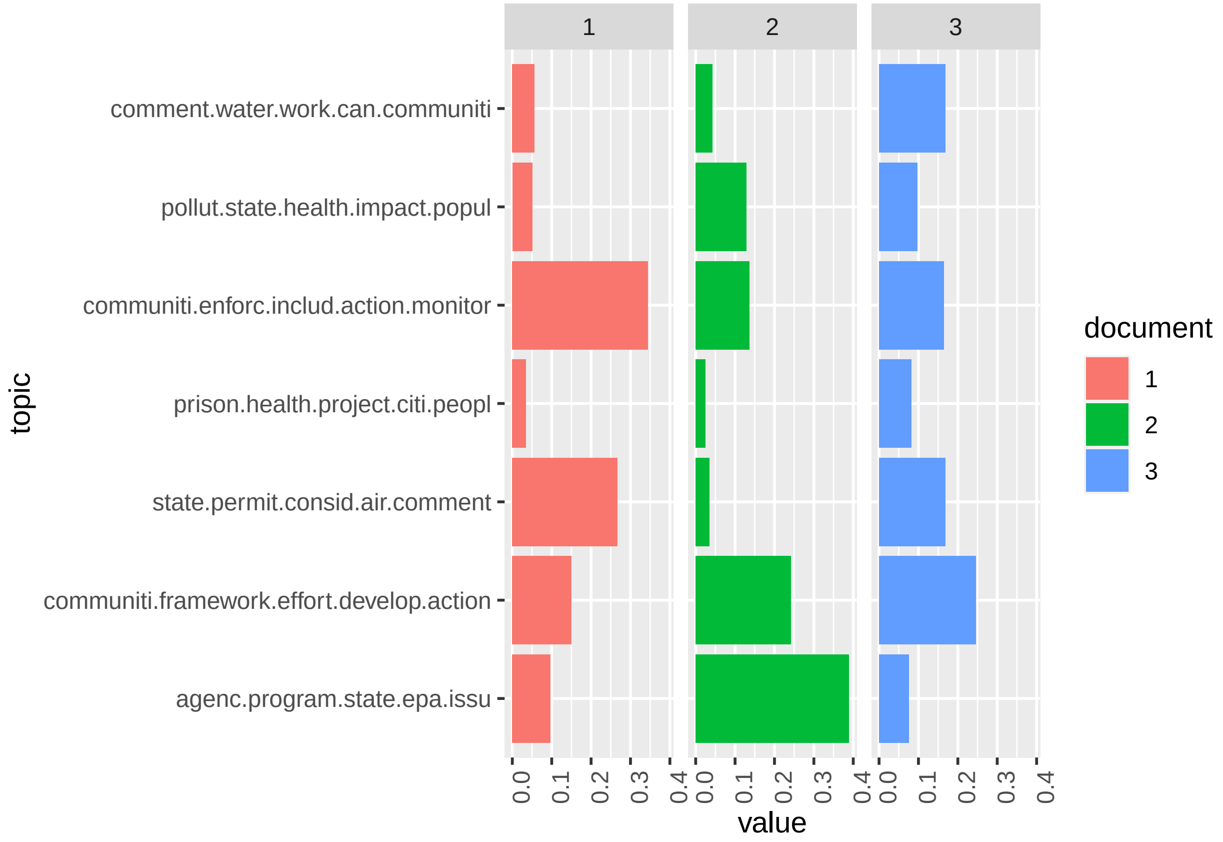
Let's assign names to the topics so we know what we are working with. We can name them by their top terms

```
top5termsPerTopic <- terms(topicModel_k7, 5)
topicNames <- apply(top5termsPerTopic, 2, paste, collapse=" ")
```

We can explore the theta matrix, which contains the distribution of each topic over each document

```
exampleIds <- c(1, 2, 3)
N <- length(exampleIds)

#lapply(epa_corp[exampleIds], as.character) #uncomment to view example text
# get topic proportions form example documents
topicProportionExamples <- theta[exampleIds,]
colnames(topicProportionExamples) <- topicNames
vizDataFrame <- melt(cbind(data.frame(topicProportionExamples), document=factor(1:N)), variable.name =
ggplot(data = vizDataFrame, aes(topic, value, fill = document), ylab = "proportion") +
  geom_bar(stat="identity") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  coord_flip() +
  facet_wrap(~ document, ncol = N)
```



Here's a neat JSON-based model visualizer

```
library(LDAvis)
library("tsne")
svd_tsne <- function(x) tsne(svd(x)$u)
json <- createJSON(
  phi = tmResult$terms,
  theta = tmResult$topics,
  doc.length = rowSums(dfm),
  vocab = colnames(dfm),
  term.frequency = colSums(dfm),
  mds.method = svd_tsne,
  plot.opts = list(xlab="", ylab="")
)
serVis(json)
```

Analysis continued

14 topics

```
k <- 14

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

```r
tmResult <- posterior(topicModel_k14)
attributes(tmResult)
```

```
## $names
## [1] "terms"  "topics"
```

```r
beta <- tmResult$terms
dim(beta)
```

```
## [1]   14 2781
```

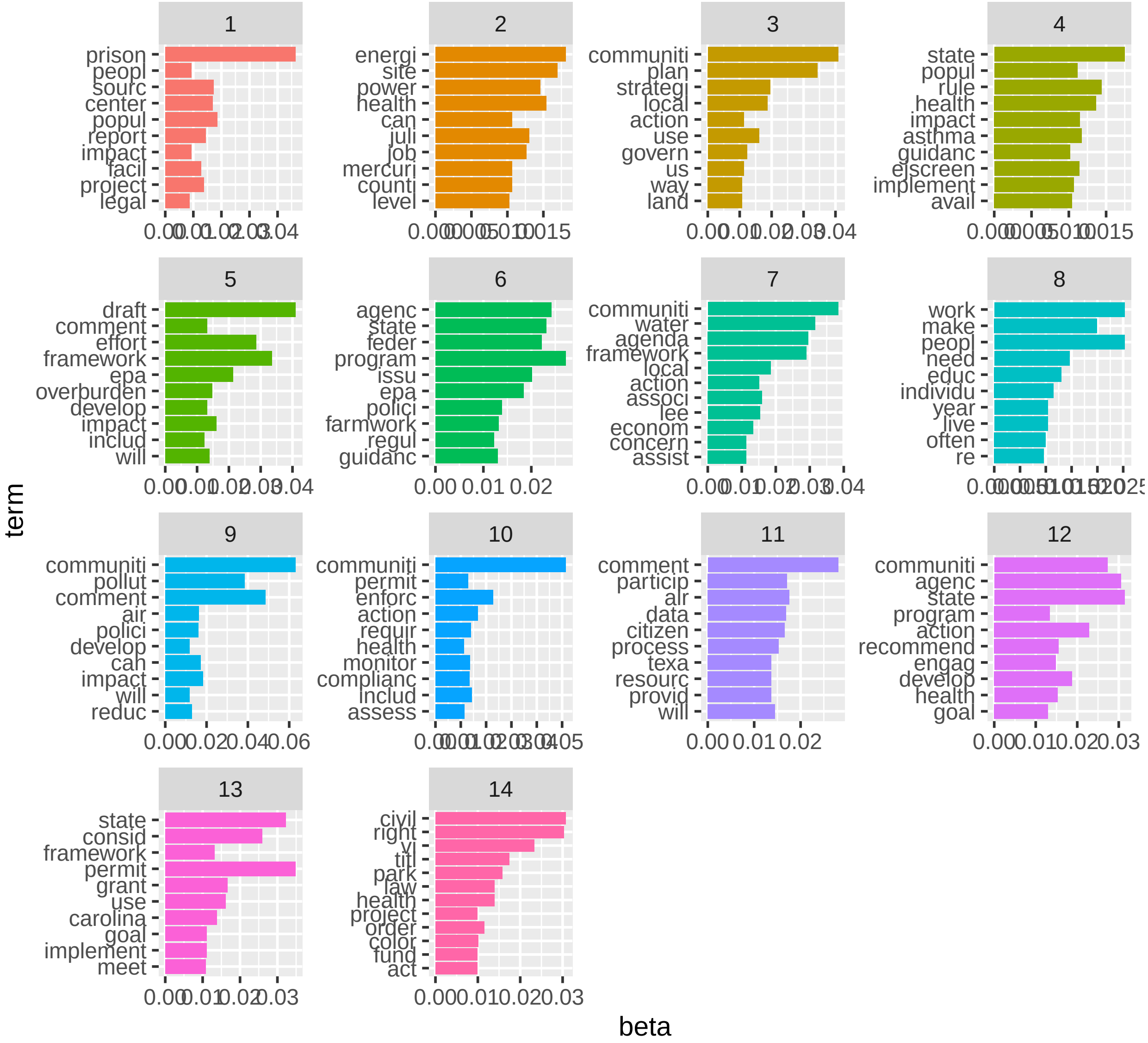
```r
terms(topicModel_k14, 10)
```

```
##       Topic 1   Topic 2   Topic 3     Topic 4     Topic 5      Topic 6
##  [1,] "prison"  "energi"  "communiti" "state"     "draft"      "program"
##  [2,] "popul"   "site"    "plan"      "rule"      "framework"  "agenc"
##  [3,] "sourc"   "health"  "strategi"  "health"    "effort"     "state"
##  [4,] "center"  "power"   "local"     "asthma"    "epa"        "feder"
##  [5,] "report"  "juli"    "use"       "impact"    "impact"     "issu"
##  [6,] "project" "job"     "govern"    "ejscreen"  "overburden" "epa"
##  [7,] "facil"   "mercuri" "action"    "popul"     "will"       "polici"
##  [8,] "impact"  "can"     "us"        "implement" "comment"    "farmwork"
##  [9,] "peopl"   "counti"  "way"       "avail"     "develop"    "guidanc"
## [10,] "legal"   "level"   "land"      "guidanc"   "includ"     "regul"
##       Topic 7     Topic 8    Topic 9     Topic 10    Topic 11   Topic 12
##  [1,] "communiti" "work"     "communiti" "communiti" "comment"  "state"
##  [2,] "water"     "peopl"    "comment"   "enforc"    "air"      "agenc"
##  [3,] "agenda"    "make"     "pollut"    "action"    "particip" "communiti"
##  [4,] "framework" "need"     "impact"    "includ"    "data"     "action"
##  [5,] "local"     "educ"     "can"       "requir"    "citizen"  "develop"
##  [6,] "associ"    "individu" "air"       "monitor"   "process"  "recommend"
##  [7,] "lee"       "year"     "polici"    "complianc" "will"     "health"
##  [8,] "action"    "live"     "reduc"     "permit"    "provid"   "engag"
##  [9,] "econom"    "often"    "will"      "assess"    "texa"     "program"
## [10,] "assist"    "re"       "develop"   "health"    "resourc"  "goal"
##       Topic 13    Topic 14
##  [1,] "permit"    "civil"
##  [2,] "state"     "right"
##  [3,] "consid"    "vi"
##  [4,] "grant"     "titl"
##  [5,] "use"       "park"
##  [6,] "carolina"  "health"
##  [7,] "framework" "law"
##  [8,] "goal"      "order"
##  [9,] "implement" "color"
## [10,] "meet"      "act"
```

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

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


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()
```



10 topics

```
k <- 10

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

tmResult <- posterior(topicModel_k10)
attributes(tmResult)
```

```
## $names
## [1] "terms"  "topics"
```

```
beta <- tmResult$terms
dim(beta)
```

```
## [1]   10 2781
```

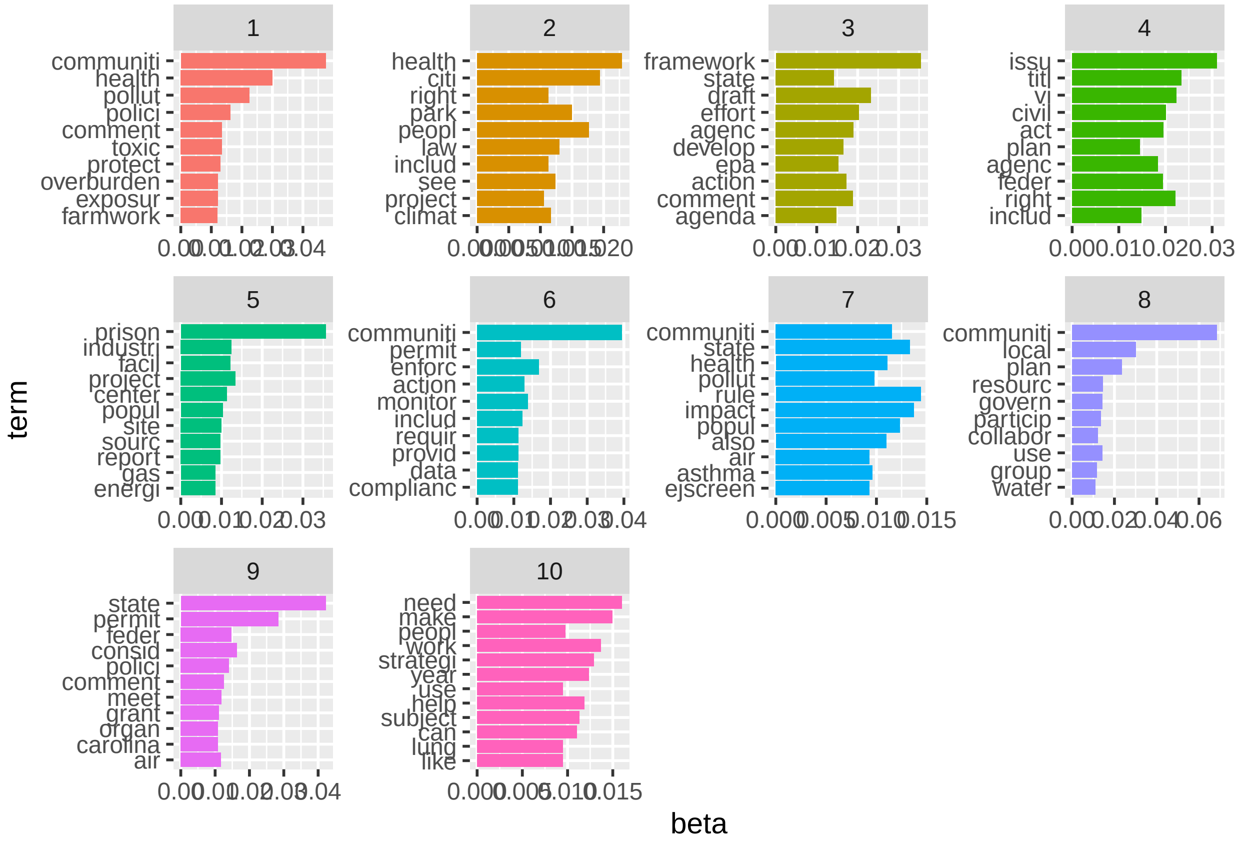
```
terms(topicModel_k10, 10)
```

```
##       Topic 1      Topic 2   Topic 3     Topic 4  Topic 5    Topic 6    
##  [1,] "communiti"  "health"  "framework" "issu"   "prison"   "communiti"
##  [2,] "health"     "citi"    "draft"     "titl"   "project"  "enforc"   
##  [3,] "pollut"     "peopl"   "effort"    "vi"     "industri" "monitor"  
##  [4,] "polici"     "park"    "agenc"     "right"  "facil"    "action"   
##  [5,] "comment"    "law"     "comment"   "civil"  "center"   "includ"   
##  [6,] "toxic"      "see"     "action"    "act"    "popul"    "permit"   
##  [7,] "protect"    "climat"  "develop"   "feder"  "site"     "provid"   
##  [8,] "exposur"    "includ"  "epa"       "agenc"  "report"   "requir"   
##  [9,] "overburden" "right"   "agenda"    "includ" "sourc"    "complianc"
## [10,] "farmwork"   "project" "state"     "plan"   "gas"      "data"     
##       Topic 7     Topic 8     Topic 9   Topic 10  
##  [1,] "rule"      "communiti" "state"   "need"    
##  [2,] "impact"    "local"     "permit"  "make"    
##  [3,] "state"     "plan"      "consid"  "work"    
##  [4,] "popul"     "resourc"   "feder"   "strategi"
##  [5,] "communiti" "use"       "polici"  "year"    
##  [6,] "health"    "govern"    "comment" "help"    
##  [7,] "also"      "particip"  "meet"    "subject" 
##  [8,] "pollut"    "collabor"  "air"     "can"     
##  [9,] "asthma"    "group"     "grant"   "peopl"   
## [10,] "air"       "water"     "organ"   "use"     
```

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

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

```
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()
```



8 Topics

```
k <- 8

topicModel_k8 <- LDA(dfm, k, method="Gibbs", control=list(iter = 500, verbose = FALSE))

tmResult <- posterior(topicModel_k8)
attributes(tmResult)
```

```
## $names
## [1] "terms"  "topics"
```

```
beta <- tmResult$terms
dim(beta)
```

```
## [1]    8 2781
```

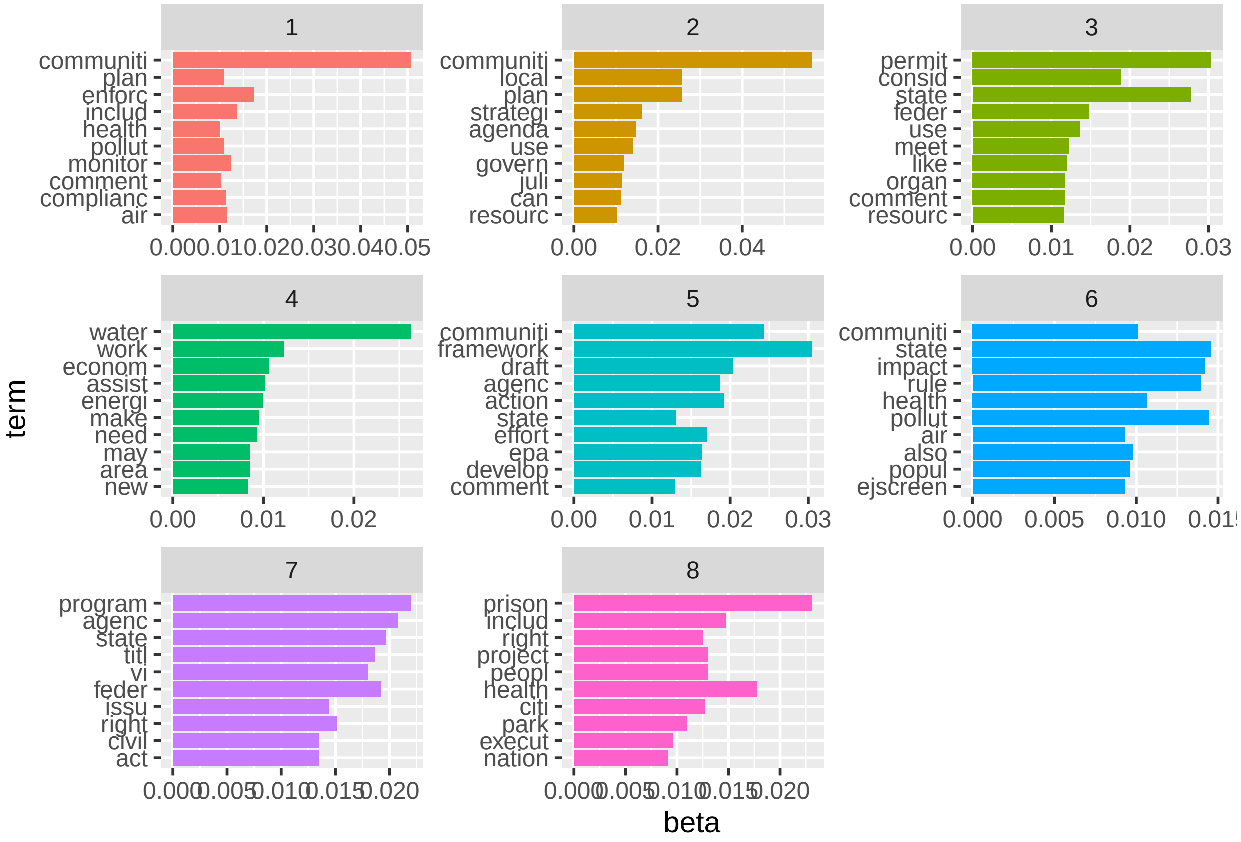
```
terms(topicModel_k8, 10)
```

```
##       Topic 1     Topic 2     Topic 3   Topic 4  Topic 5     Topic 6    
##  [1,] "communiti" "communiti" "permit"  "water"  "framework" "state"    
##  [2,] "enforc"    "local"     "state"   "work"   "communiti" "pollut"   
##  [3,] "includ"    "plan"      "consid"  "econom" "draft"     "impact"   
##  [4,] "monitor"   "strategi"  "feder"   "assist" "action"    "rule"     
##  [5,] "air"       "agenda"    "use"     "energi" "agenc"     "health"   
##  [6,] "complianc" "use"       "meet"    "make"   "effort"    "communiti"
##  [7,] "pollut"    "govern"    "like"    "need"   "epa"       "also"     
##  [8,] "plan"      "juli"      "comment" "may"    "develop"   "popul"    
##  [9,] "comment"   "can"       "organ"   "area"   "state"     "air"      
## [10,] "health"    "resourc"   "resourc" "new"    "comment"   "ejscreen" 
##       Topic 7   Topic 8  
##  [1,] "program" "prison" 
##  [2,] "agenc"   "health" 
##  [3,] "state"   "includ" 
##  [4,] "feder"   "project"
##  [5,] "titl"    "peopl"  
##  [6,] "vi"      "citi"   
##  [7,] "right"   "right"  
##  [8,] "issu"    "park"   
##  [9,] "act"     "execut" 
## [10,] "civil"   "nation" 
```

```
comment_topics <- tidy(topicModel_k8, matrix = "beta")

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


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 value for k

Based on the Deveaud2014 metric in the `FindTopicsNumber()` analysis, I chose to try 10, 14, and 8 as possible numbers of topics. I assessed these models by looking at the frequency of top terms in each topic and with the `LDAvis` app. After running these additional models, I think that 8 topics has been the most successful so far. In my opinion it seems like when more than 8 topics are formed, they start to become more redundant and the lines between them start to blur. In part, I based this on looking at the top words in each supposed topic and feeling out how cohesive and unique each was. Using `LDAvis`, it appears using 8 topics creates a fairly equidistant spacing between the topics—as the topics increase beyond this amount, certain topics start to become closer to one another. This choice would align fairly well with Deveaud2014 metric, as 8 was one of the topic numbers with a higher value, albeit not among the very highest values.