Topic 6: Topic Analysis

```
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
```

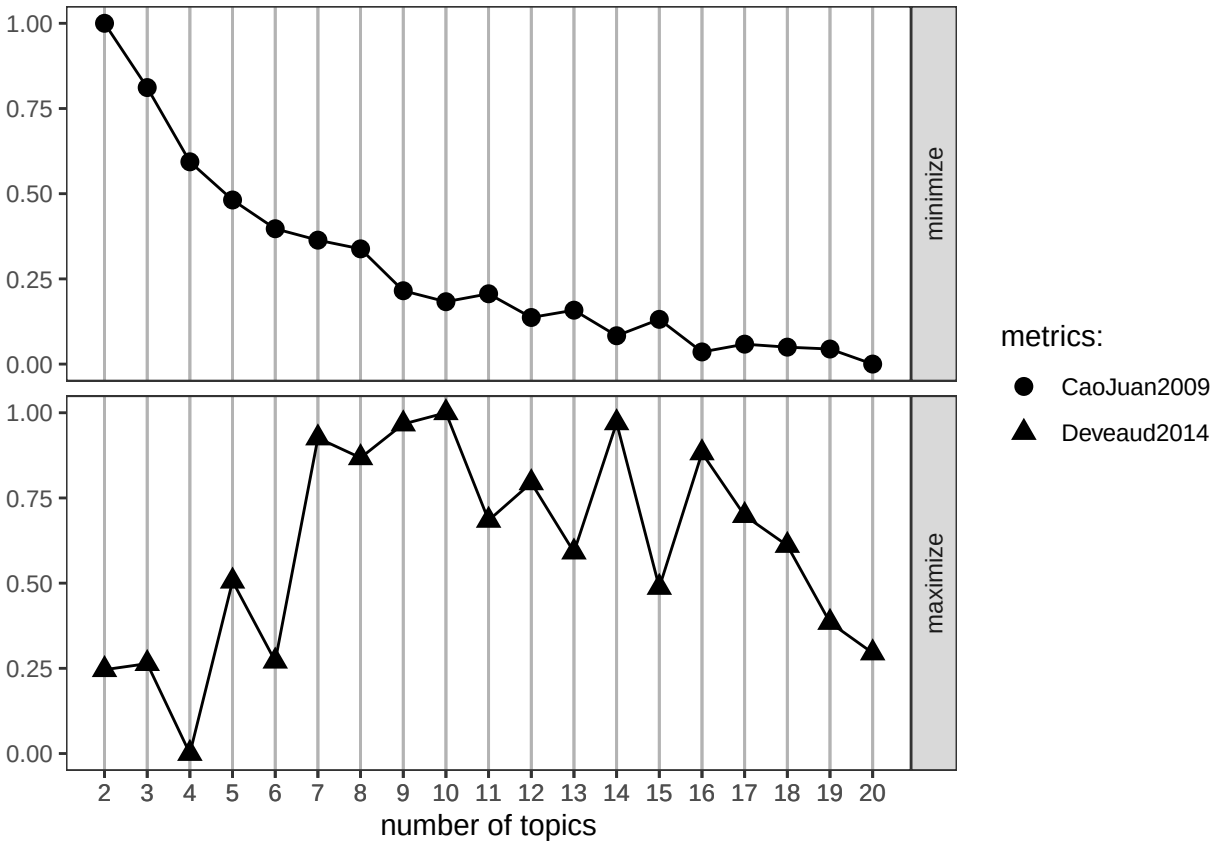
```
##   text5     4   5      1      1      1         1     0     1      1
##   text6     1   1      1      3      1         3     0     4      0
##        features
## docs    pennsylvania
##   text1            1
##   text2            0
##   text3            0
##   text4            0
##   text5            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:

Either:

A) continue on with the analysis we started:

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!
```

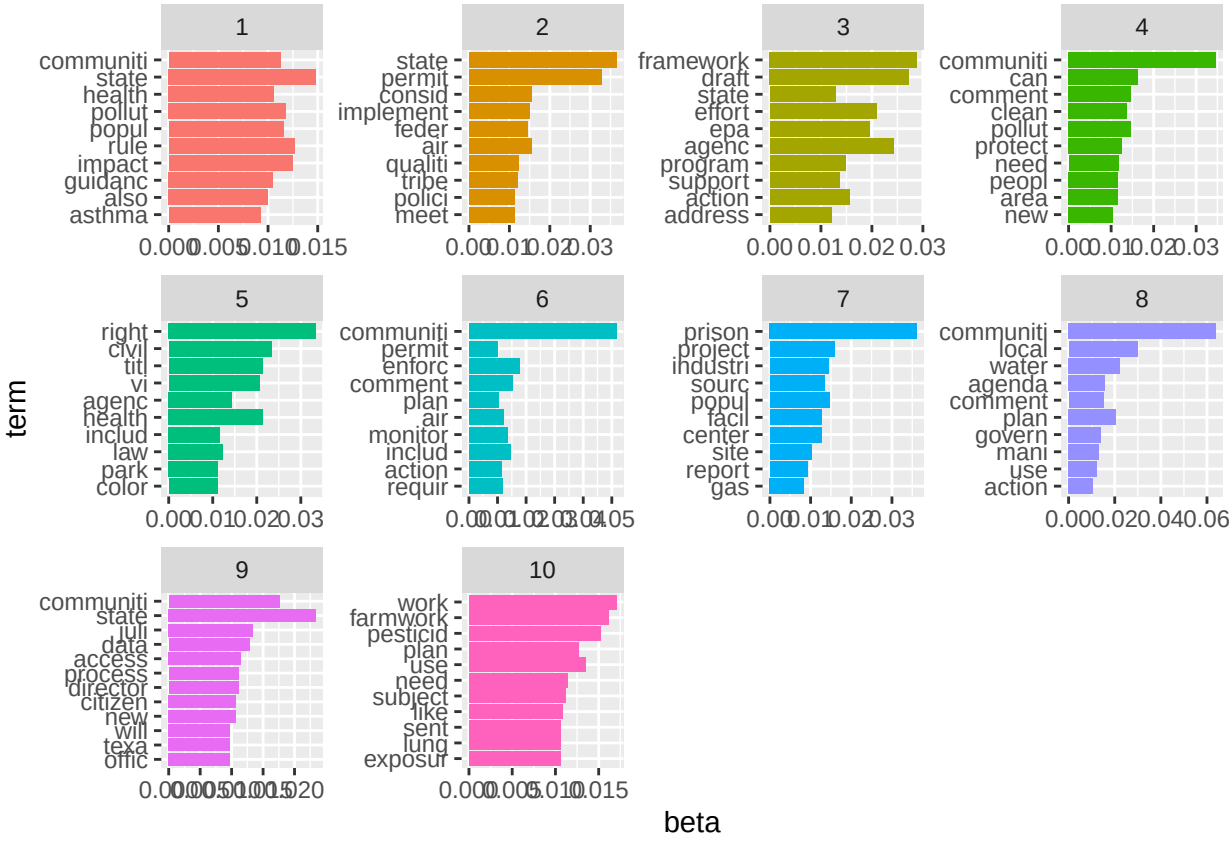
```
tmResult <- posterior(topicModel_k10)
terms(topicModel_k10, 10)
```

```
##       Topic 1     Topic 2     Topic 3     Topic 4     Topic 5  Topic 6
##  [1,] "state"     "state"     "framework" "communiti" "right"  "communiti"
##  [2,] "rule"      "permit"    "draft"     "can"       "civil"  "enforc"
##  [3,] "impact"    "air"       "agenc"     "comment"   "health" "comment"
##  [4,] "pollut"    "consid"    "effort"    "pollut"    "titl"   "includ"
##  [5,] "popul"     "implement" "epa"       "clean"     "vi"     "monitor"
##  [6,] "communiti" "feder"     "action"    "protect"   "agenc"  "air"
##  [7,] "health"    "qualiti"   "program"   "need"      "law"    "requir"
##  [8,] "guidanc"   "tribe"     "support"   "area"      "includ" "action"
##  [9,] "also"      "polici"    "state"     "peopl"     "park"   "plan"
## [10,] "asthma"    "meet"      "address"   "new"       "color"  "permit"
##       Topic 7    Topic 8     Topic 9     Topic 10
##  [1,] "prison"   "communiti" "state"     "work"
##  [2,] "project"  "local"     "communiti" "farmwork"
##  [3,] "popul"    "water"     "juli"      "pesticid"
##  [4,] "industri" "plan"      "data"      "use"
##  [5,] "sourc"    "agenda"    "access"    "plan"
##  [6,] "facil"    "comment"   "process"   "need"
##  [7,] "center"   "govern"    "director"  "subject"
##  [8,] "site"     "mani"      "citizen"   "like"
##  [9,] "report"   "use"       "new"       "exposur"
## [10,] "gas"      "action"    "offic"     "lung"
```

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

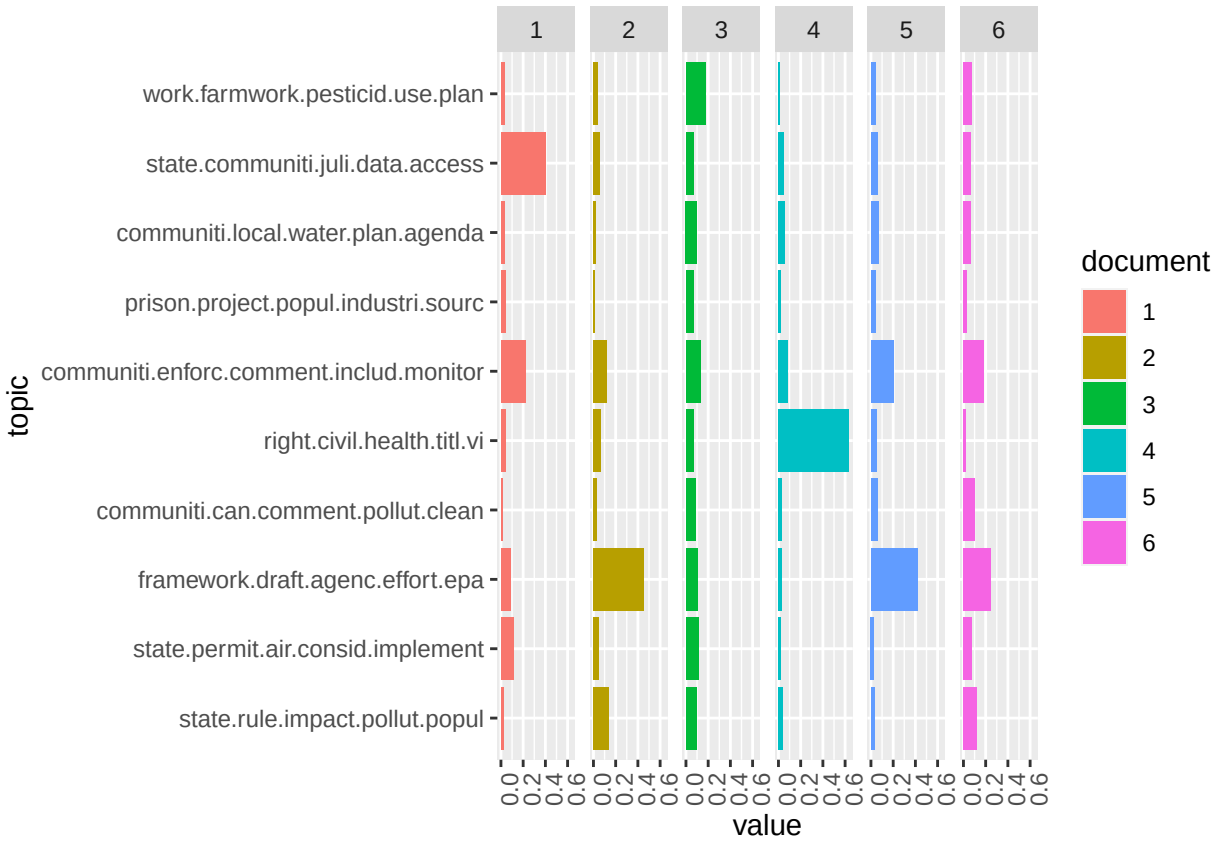
```
comment_topics <- tidy(topicModel_k10, matrix = "beta")

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

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

top_terms %>%
  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=" ")
```

```
exampleIds <- c(1, 2, 3, 4, 5, 6)
N <- length(exampleIds)

#lapply(epa_corp[exampleIds], as.character) #uncomment to view example text
# get topic proportions form example documents
topicProportions <- theta[exampleIds,]
colnames(topicProportions) <- topicNames_k10
vizDataFrame <- melt(cbind(data.frame(topicProportions), document=factor(1:N)), variable.name = "topic"
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)
```



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 ...
```

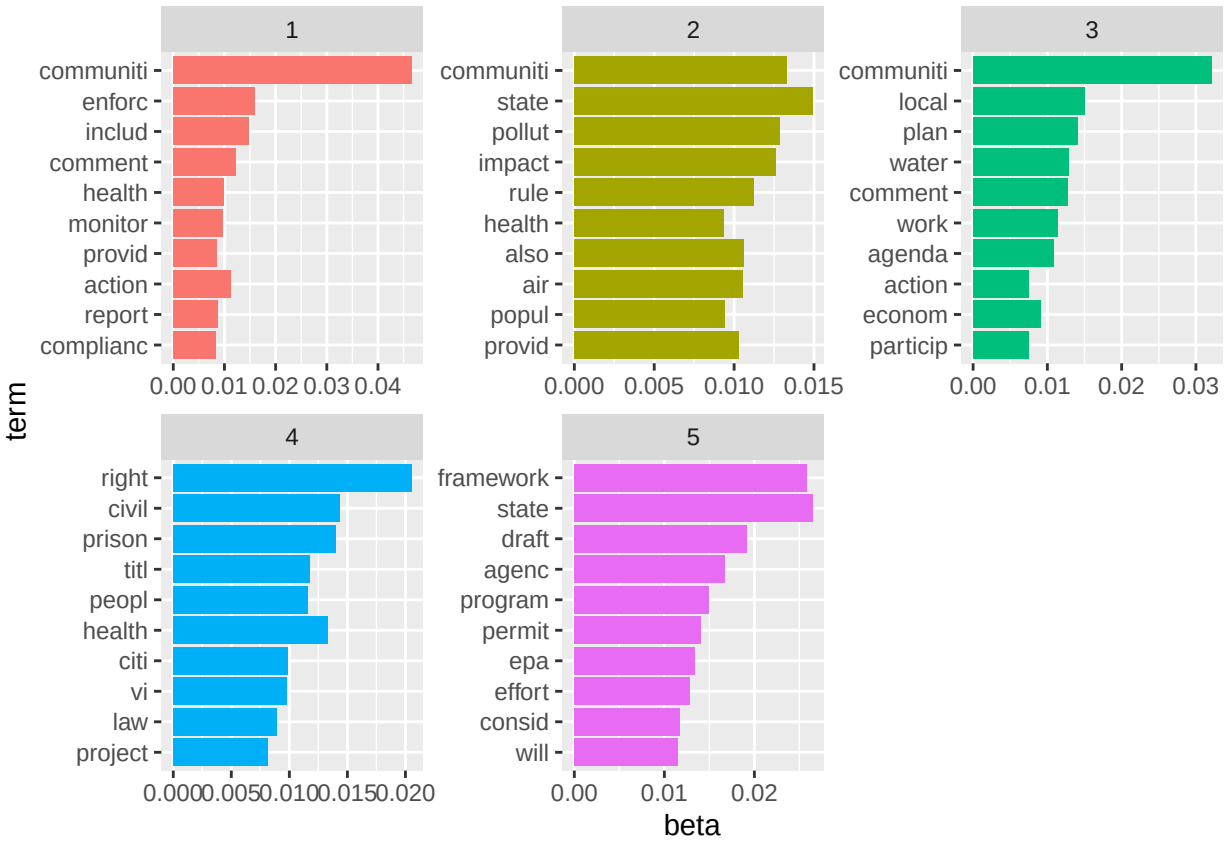
```
## Iteration 400 ...
## Iteration 425 ...
## Iteration 450 ...
## Iteration 475 ...
## Iteration 500 ...
## Gibbs sampling completed!
```

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

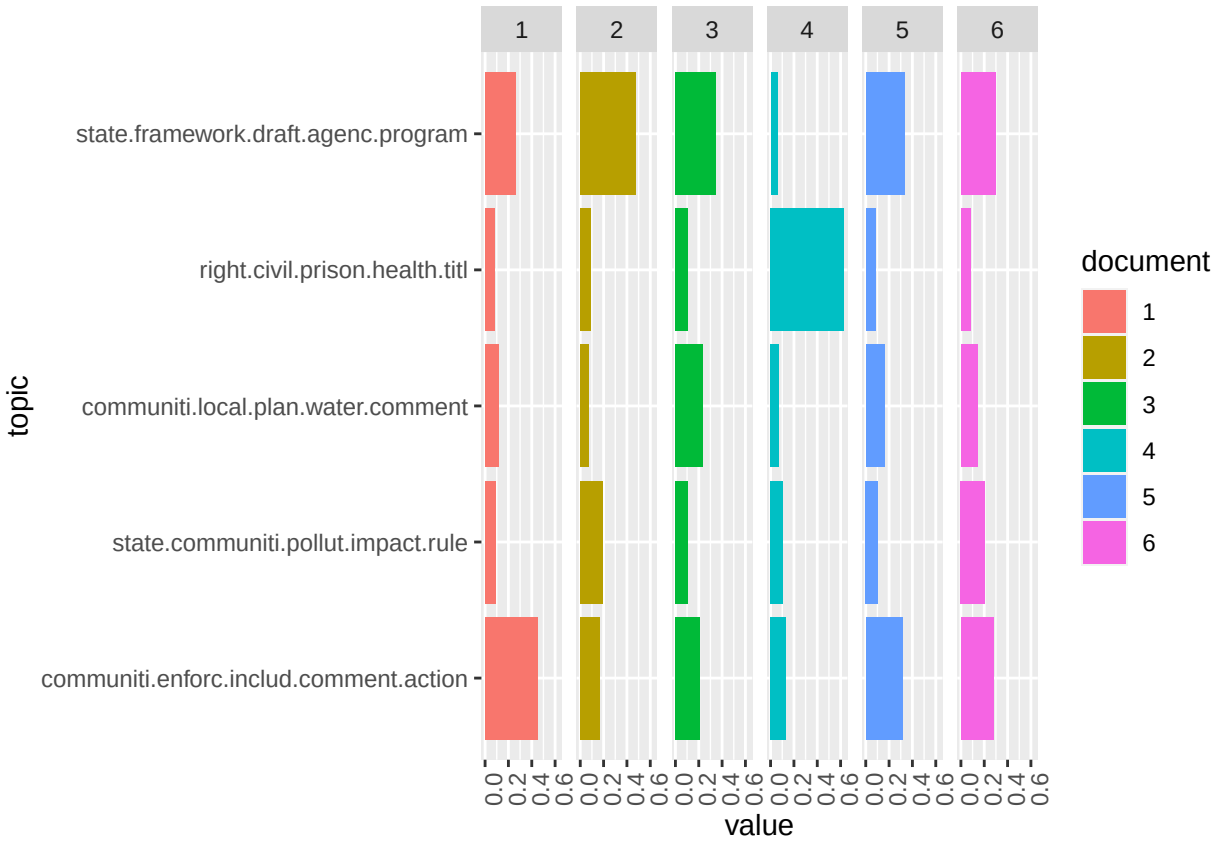
```
##       Topic 1     Topic 2     Topic 3     Topic 4   Topic 5
##  [1,] "communiti" "state"     "communiti" "right"   "state"
##  [2,] "enforc"    "communiti" "local"     "civil"   "framework"
##  [3,] "includ"    "pollut"    "plan"      "prison"  "draft"
##  [4,] "comment"   "impact"    "water"     "health"  "agenc"
##  [5,] "action"    "rule"      "comment"   "titl"    "program"
##  [6,] "health"    "also"      "work"      "peopl"   "permit"
##  [7,] "monitor"   "air"       "agenda"    "citi"    "epa"
##  [8,] "report"    "provid"    "econom"    "vi"      "effort"
##  [9,] "provid"    "popul"     "particip"  "law"     "consid"
## [10,] "complianc" "health"    "action"    "project" "will"
```

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

```
comment_topics <- tidy(topicModel_k5, 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()
```




```
top5termsPerTopic <- terms(topicModel_k5, 5)
topicNames_k5 <- apply(top5termsPerTopic, 2, paste, collapse=" ")

exampleIds <- c(1, 2, 3, 4, 5, 6)
N <- length(exampleIds)

#lapply(epa_corp[exampleIds], as.character) #uncomment to view example text
# get topic proportions form example documents
topicProportions <- theta[exampleIds,]
colnames(topicProportions) <- topicNames_k5
vizDataFrame <- melt(cbind(data.frame(topicProportions), document=factor(1:N)), variable.name = "topic"
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)
```



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!
```

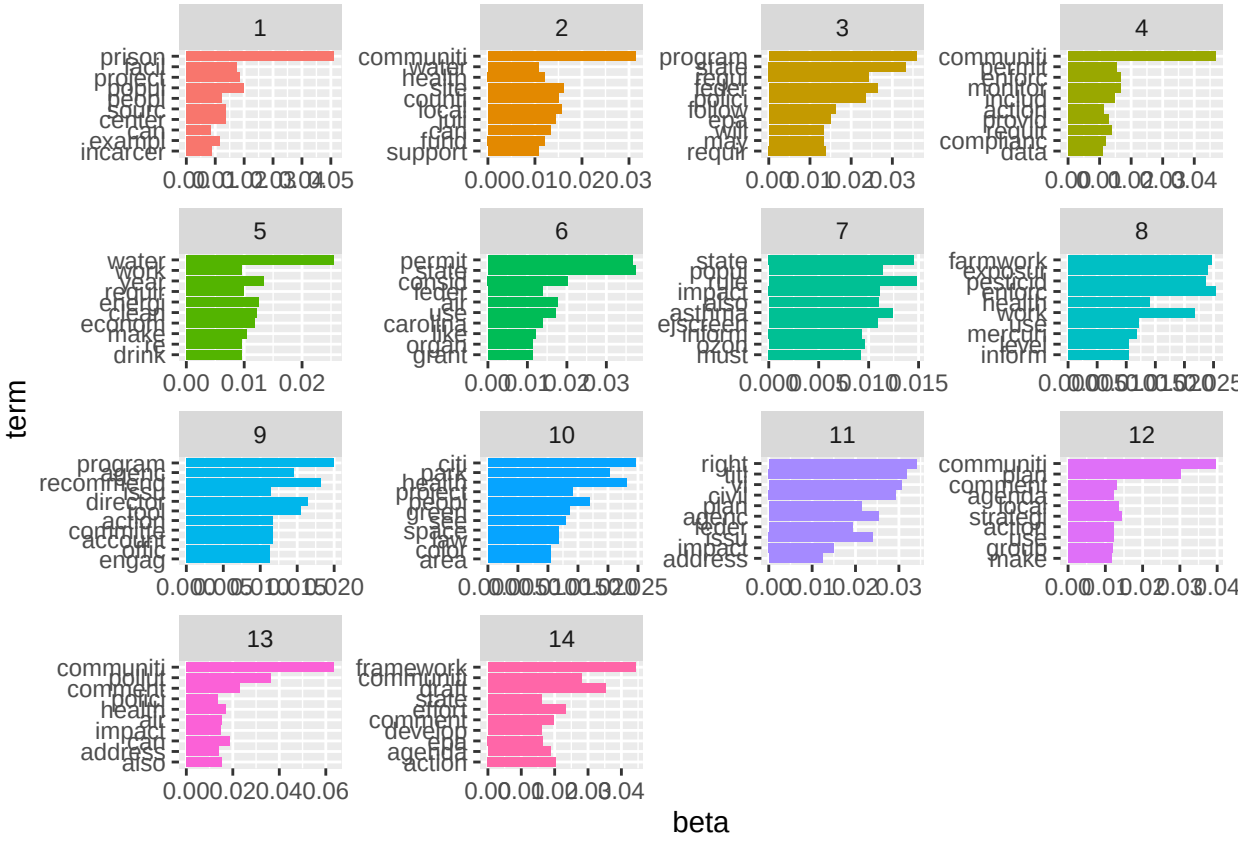
```
tmResult <- posterior(topicModel_k14)
terms(topicModel_k14, 10)
```

```
##       Topic 1    Topic 2     Topic 3   Topic 4     Topic 5  Topic 6
##  [1,] "prison"   "communiti" "program" "communiti" "water"  "state"
##  [2,] "popul"    "site"      "state"   "monitor"   "year"   "permit"
##  [3,] "project"  "local"     "feder"   "enforc"    "energi" "consid"
##  [4,] "facil"    "counti"    "regul"   "permit"    "clean"  "air"
##  [5,] "center"   "juli"      "polici"  "includ"    "econom" "use"
##  [6,] "sourc"    "can"       "follow"  "requir"    "make"   "feder"
##  [7,] "peopl"    "health"    "epa"     "provid"    "requir" "carolina"
##  [8,] "exampl"   "fund"      "requir"  "complianc" "re"     "like"
##  [9,] "incarcer" "support"   "will"    "action"    "work"   "organ"
## [10,] "can"      "water"     "may"     "data"      "drink"  "grant"
##       Topic 7    Topic 8    Topic 9     Topic 10  Topic 11  Topic 12
##  [1,] "rule"     "enforc"   "program"   "citi"    "right"   "communiti"
##  [2,] "state"    "farmwork" "recommend" "health"  "titl"    "plan"
##  [3,] "asthma"   "exposur"  "director"  "park"    "vi"      "strategi"
##  [4,] "popul"    "pesticid" "tool"      "peopl"   "civil"   "local"
##  [5,] "impact"   "work"     "agenc"     "project" "agenc"   "comment"
##  [6,] "also"     "health"   "action"    "green"   "issu"    "action"
##  [7,] "ejscreen" "use"      "committe"  "see"     "plan"    "agenda"
##  [8,] "ozon"     "mercuri"  "account"   "law"     "feder"   "use"
##  [9,] "inform"   "inform"   "issu"      "space"   "impact"  "group"
## [10,] "must"     "level"    "offic"     "area"    "address" "make"
##       Topic 13    Topic 14
##  [1,] "communiti" "framework"
##  [2,] "pollut"    "draft"
##  [3,] "comment"   "communiti"
##  [4,] "can"       "effort"
##  [5,] "health"    "action"
##  [6,] "also"      "comment"
##  [7,] "air"       "agenda"
##  [8,] "impact"    "epa"
##  [9,] "address"   "state"
## [10,] "polici"    "develop"
```

```
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 %>%
  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_k14, 5)
topicNames_k14 <- apply(top5termsPerTopic, 2, paste, collapse=" ")

exampleIds <- c(1, 2, 3, 4, 5, 6)
N <- length(exampleIds)

#lapply(epa_corp[exampleIds], as.character) #uncomment to view example text
# get topic proportions form example documents
topicProportions <- theta[exampleIds,]
colnames(topicProportions) <- topicNames_k14
vizDataFrame <- melt(cbind(data.frame(topicProportions), document=factor(1:N)), variable.name = "topic"
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)
```

Best k-value

Based on the plot we made in class from Deveaud2014 method, it looks like 10 number of topics maximizes the metric. We know there are 9 priority areas from the EPA so it makes sense that there may be an extra topic that catches another topic from the pdfs.

OR

B) use the data you plan to use for your final project:

Prepare the data so that it can be analyzed in the topicmodels package

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