Homework 5

Paloma Cartwright

2022-05-10

Read in the Data

```
comments_df <- read_csv("https://raw.githubusercontent.com/MaRo406/EDS_231-text-sentiment/main/dat/comme
```

Build the Corpus

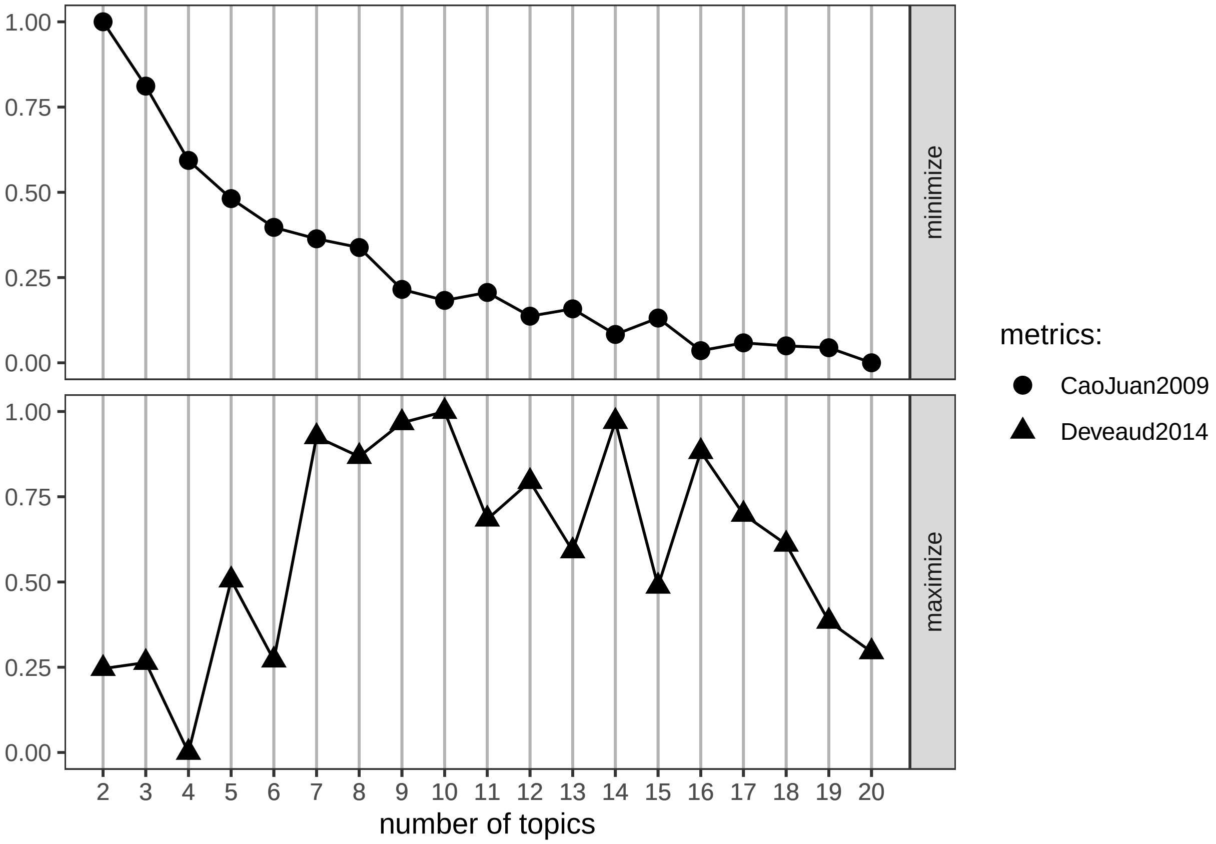
```
epa_corp <- corpus(x = comments_df, text_field = "text")
epa_corp_stats <- summary(epa_corp)

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")

dfm_comm<- dfm(toks1, tolower = TRUE)
dfm <- dfm_wordstem(dfm_comm)
dfm <- dfm_trim(dfm, min_docfreq = 2)

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

Find number of topics (Class Example)

```
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)
```



```
k <- 7

topicModel_k7 <- LDA(dfm, k, method="Gibbs", control=list(iter = 500, verbose = 25))
```

Try with k = 7 (Class Example)

```
## K = 7; 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!
```

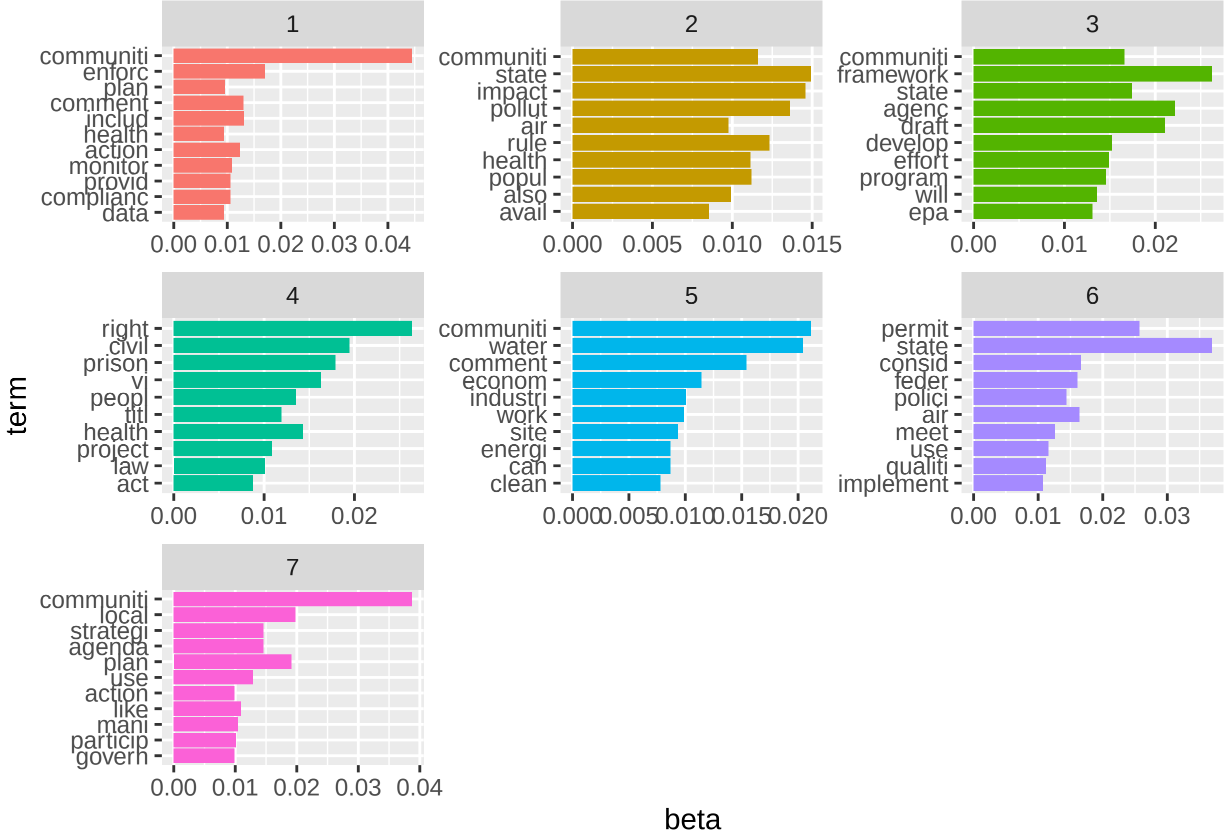
```
tmResult <- posterior(topicModel_k7)
terms(topicModel_k7, 10)
```

```
##       Topic 1     Topic 2     Topic 3     Topic 4   Topic 5     Topic 6
##  [1,] "communiti" "state"     "framework" "right"   "communiti" "state"
##  [2,] "enforc"    "impact"    "agenc"     "civil"   "water"     "permit"
##  [3,] "includ"    "pollut"    "draft"     "prison"  "comment"   "consid"
##  [4,] "comment"   "rule"      "state"     "vi"      "econom"    "air"
##  [5,] "action"    "communiti" "communiti" "health"  "industri"  "feder"
##  [6,] "monitor"   "popul"     "develop"   "peopl"   "work"      "polici"
##  [7,] "provid"    "health"    "effort"    "titl"    "site"      "meet"
##  [8,] "complianc" "also"      "program"   "project" "can"       "use"
##  [9,] "plan"      "air"       "will"      "law"     "energi"    "qualiti"
## [10,] "health"    "avail"     "epa"       "act"     "clean"     "implement"
##       Topic 7
##  [1,] "communiti"
##  [2,] "local"
##  [3,] "plan"
##  [4,] "agenda"
##  [5,] "strategi"
##  [6,] "use"
##  [7,] "like"
##  [8,] "mani"
##  [9,] "particip"
## [10,] "action"
```

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

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

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

top_terms
```

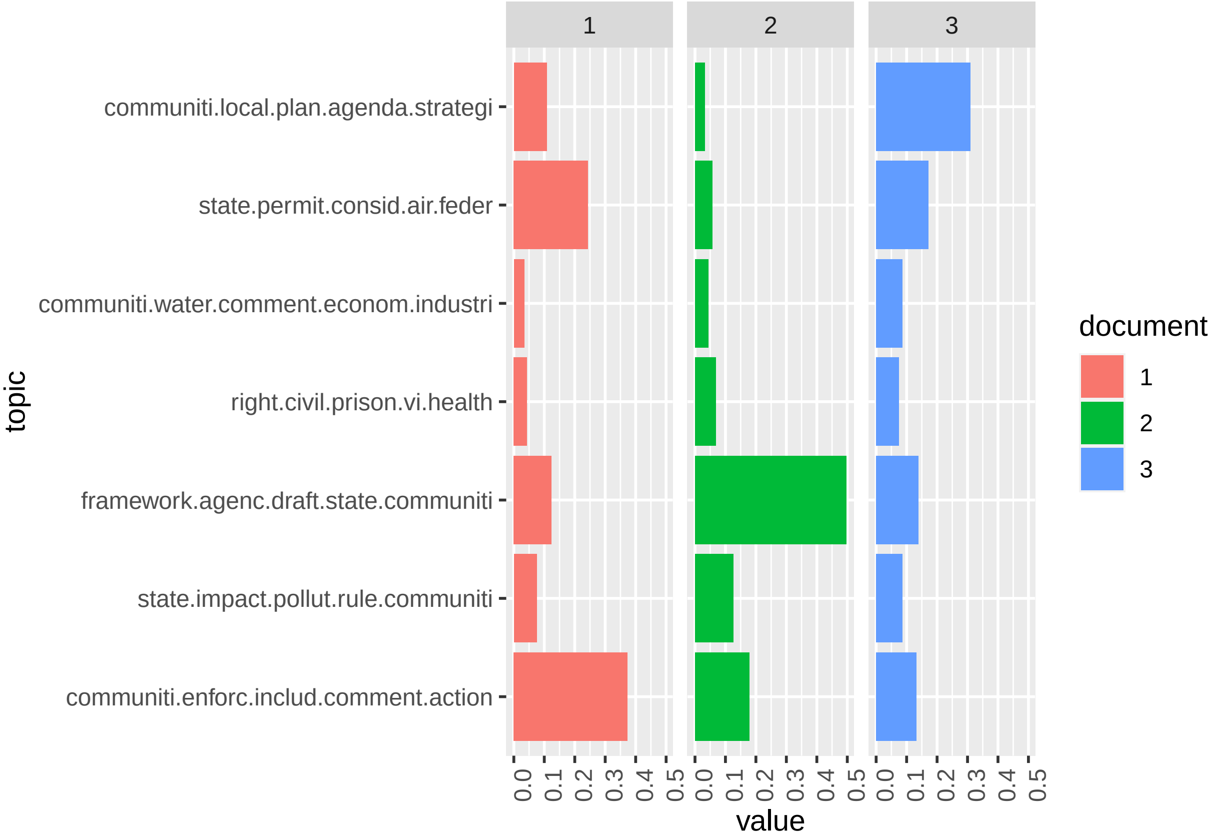
```
## # A tibble: 72 x 3
##    topic term         beta
##    <int> <chr>       <dbl>
##  1     1 communiti 0.0445
##  2     1 enforc    0.0170
##  3     1 includ    0.0131
##  4     1 comment   0.0130
##  5     1 action    0.0124
##  6     1 monitor   0.0109
##  7     1 provid    0.0106
##  8     1 complianc 0.0106
## 9      1 plan      0.00954
## 10     1 health    0.00940
## # ... with 62 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_k7, 5)
topicNames <- apply(top5termsPerTopic, 2, paste, collapse=" ")

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() +
```

```r
  facet_wrap(~ document, ncol = N)
```



Class Example

Assignment:

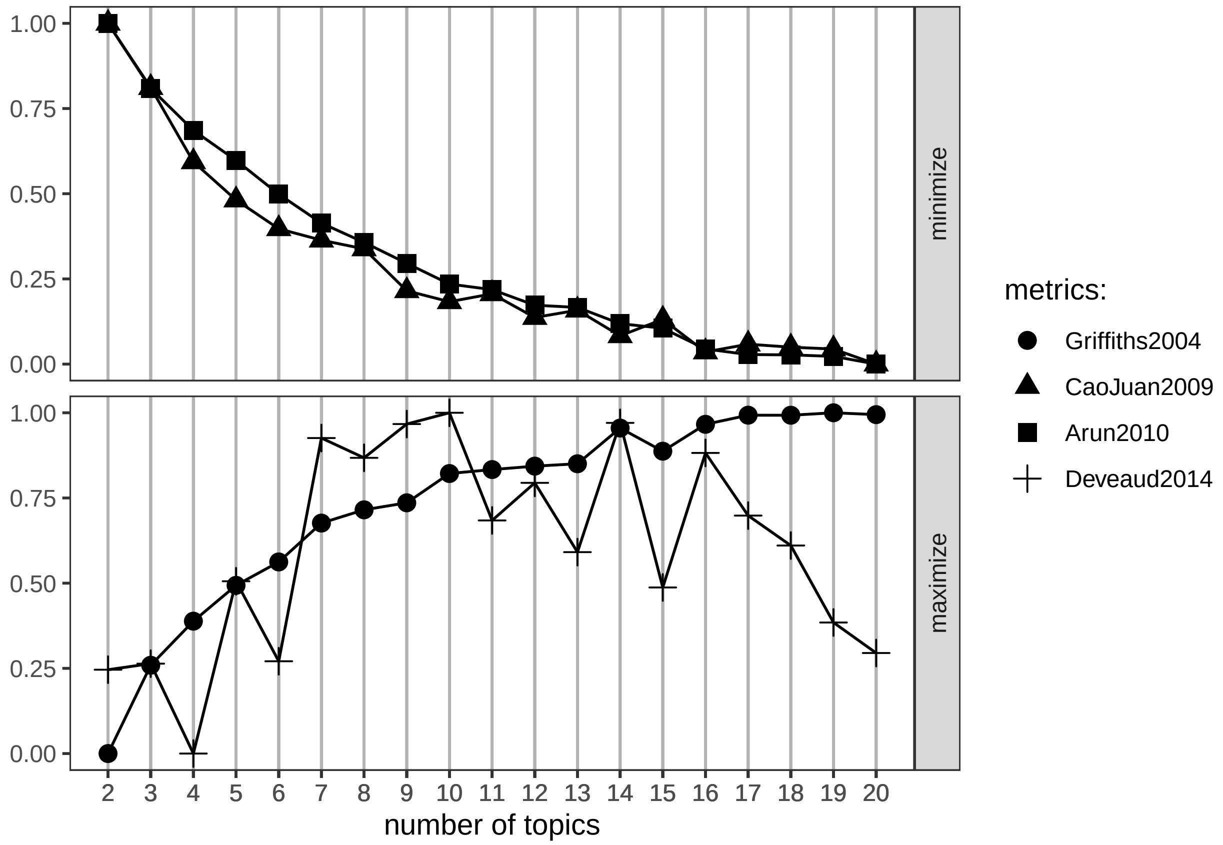
Run three more models and select the overall best value for k (the number of topics) - include some justification for your selection: theory, FindTopicsNumber() optimization metrics, interpretability, LDAvis

First find the topics number using `FindTopicsNumber()`

```r
hw_result <- FindTopicsNumber(
  dfm,
  topics = seq(from = 2, to = 20, by = 1),
  metrics = c("Griffiths2004", "CaoJuan2009", "Arun2010", "Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 77),
  verbose = TRUE
)
```

```
## fit models... done.
## calculate metrics:
##   Griffiths2004... done.
##   CaoJuan2009... done.
##   Arun2010... done.
##   Deveaud2014... done.
```

```
FindTopicsNumber_plot(hw_result)
```



Based on the minimization of "CaoJuan2009" and "Arun2010", the ideal number of topics is likely 20 and then based on "Griffiths2004" and "Deveaud2014" the ideal number of topics is as little as 7. So I will try 10, 14, and 20 for my k value in the models. I chose 10 instead of 7 because we already did this earlier in the analysis and 14 is the value of a large peak.

```
k <- 10
topicModel_k10 <- LDA(dfm, k, method="Gibbs", control=list(iter = 500, verbose = 25))
```

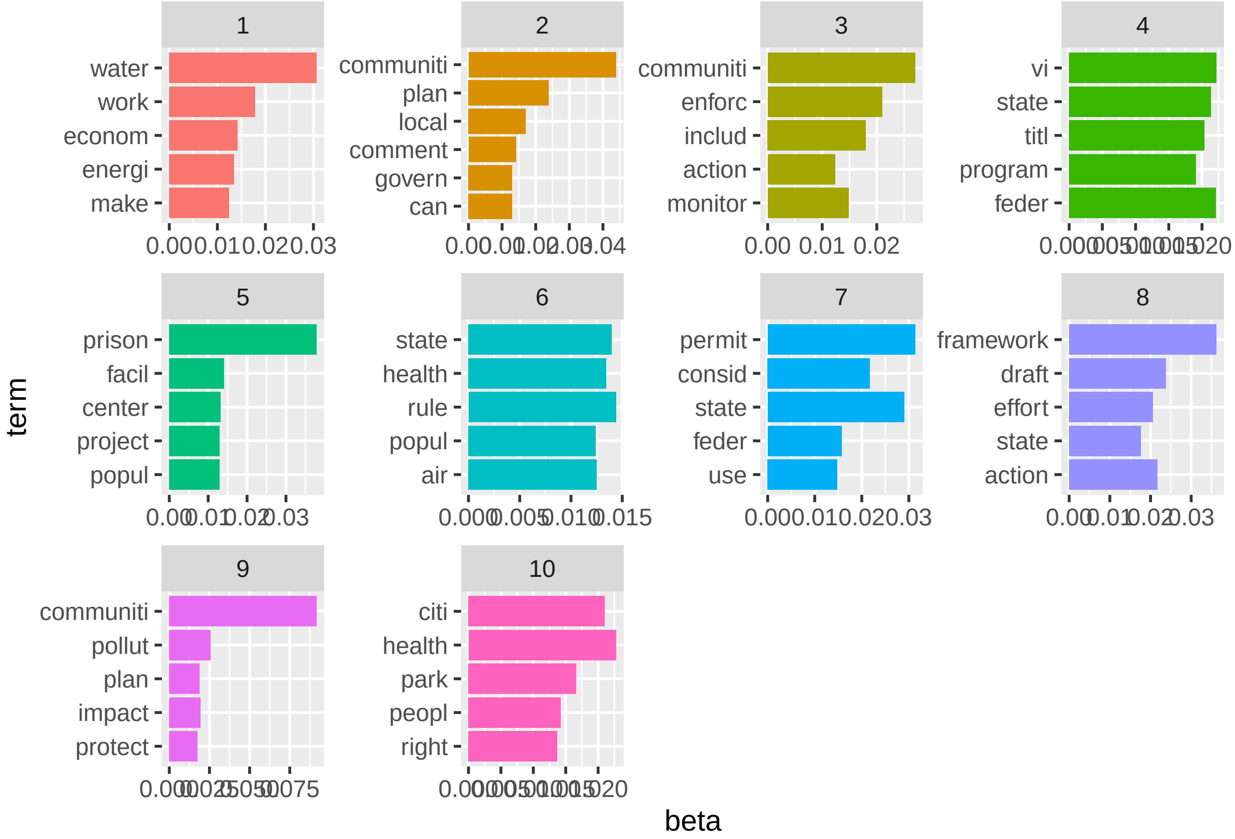
Try for k = 10

```
## 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,] "water"         "communiti" "communiti" "vi"      "prison"   "rule"
##  [2,] "work"          "plan"      "enforc"    "feder"   "facil"    "state"
##  [3,] "econom"        "local"     "includ"    "state"   "center"   "health"
##  [4,] "energi"        "comment"   "monitor"   "titl"    "project"  "air"
##  [5,] "make"          "can"       "action"    "program" "popul"    "popul"
##  [6,] "individu"      "govern"    "data"      "agenc"   "sourc"    "impact"
##  [7,] "clean"         "strategi"  "permit"    "issu"    "contamin" "ejscreen"
##  [8,] "area"          "help"      "comment"   "civil"   "peopl"    "asthma"
##  [9,] "infrastructur" "use"       "complianc" "right"   "address"  "provid"
## [10,] "industri"      "land"      "health"    "polici"  "site"     "also"
##       Topic 7    Topic 8     Topic 9     Topic 10
##  [1,] "permit"   "framework" "communiti" "health"
##  [2,] "state"    "draft"     "pollut"    "citi"
##  [3,] "consid"   "action"    "impact"    "park"
##  [4,] "feder"    "effort"    "plan"      "peopl"
##  [5,] "use"      "state"     "protect"   "right"
##  [6,] "grant"    "agenc"     "comment"   "law"
##  [7,] "organ"    "develop"   "also"      "project"
##  [8,] "carolina" "agenda"    "will"      "green"
##  [9,] "opportun" "epa"       "need"      "includ"
## [10,] "comment"  "goal"      "result"    "nation"
```

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

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

top_terms <- comment_topics %>%
  group_by(topic) %>%
  top_n(5, 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()
```



```
k <- 14
topicModel_k14 <- LDA(dfm, k, method="Gibbs", control=list(iter = 500, verbose = 25))
```

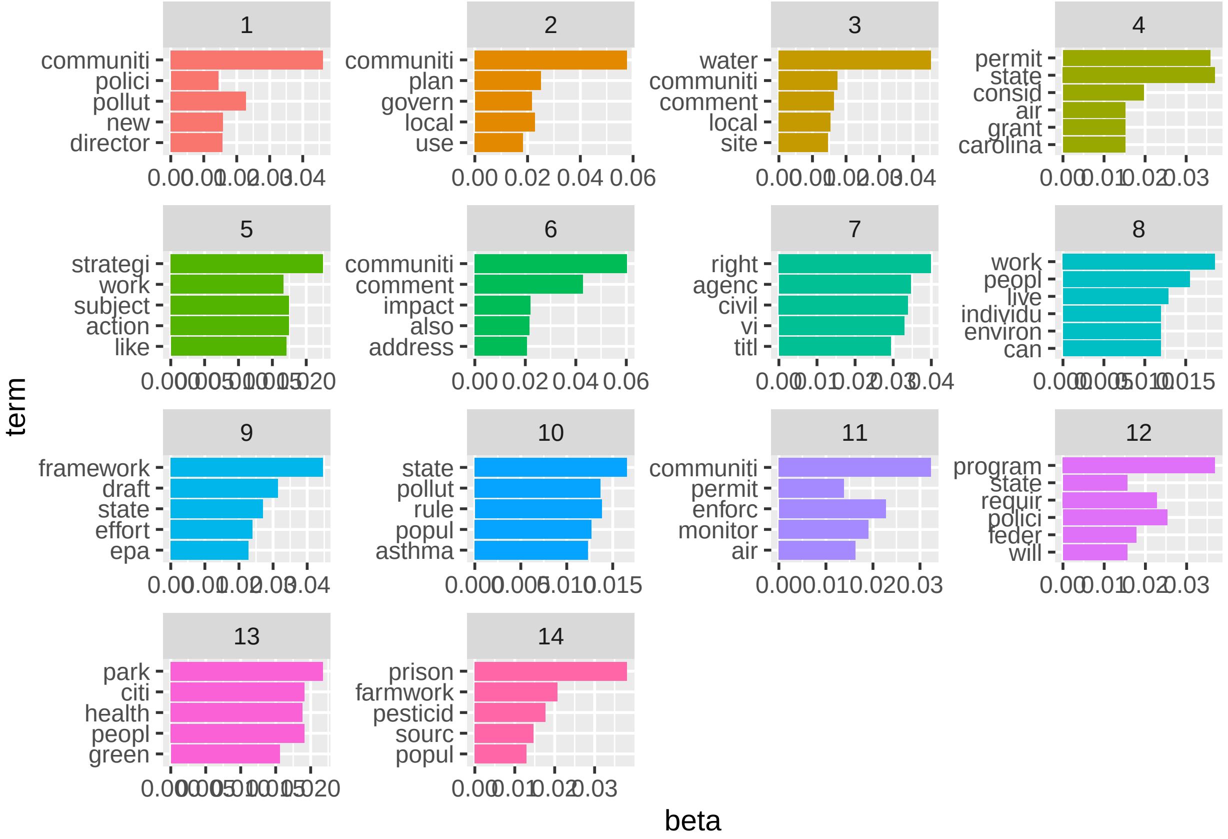
Try again for k = 14

```
## 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" "communiti" "water"         "state"    "strategi" "communiti"
##   [2,] "pollut"    "plan"      "communiti"     "permit"   "action"   "comment"
##   [3,] "new"       "local"     "comment"       "consid"   "subject"  "impact"
##   [4,] "director"  "govern"    "local"         "air"      "like"     "also"
##   [5,] "polici"    "use"       "site"          "grant"    "work"     "address"
##   [6,] "protect"   "particip"  "clean"         "carolina" "agenda"   "use"
##   [7,] "action"    "process"   "june"          "comment"  "plan"     "concern"
##   [8,] "comment"   "group"     "job"           "use"      "make"     "exampl"
##   [9,] "develop"   "texa"      "infrastructur" "opportun" "sent"     "issu"
##  [10,] "air"       "engag"     "counti"        "north"    "help"     "result"
##       Topic 7     Topic 8     Topic 9     Topic 10   Topic 11    Topic 12
##   [1,] "right"     "work"      "framework" "state"    "communiti" "program"
##   [2,] "agenc"     "peopl"     "draft"     "rule"     "enforc"    "polici"
##   [3,] "civil"     "live"      "state"     "pollut"   "monitor"   "requir"
##   [4,] "vi"        "environ"   "effort"    "popul"    "air"       "feder"
##   [5,] "titl"      "can"       "epa"       "asthma"   "permit"    "will"
##   [6,] "feder"     "individu"  "agenc"     "health"   "complianc" "state"
##   [7,] "issu"      "re"        "action"    "air"      "requir"    "regul"
##   [8,] "plan"      "econom"    "agenda"    "ejscreen" "assess"    "tribe"
##   [9,] "act"       "requir"    "develop"   "avail"    "action"    "import"
##  [10,] "implement" "underserv" "support"   "guidanc"  "region"    "order"
##       Topic 13  Topic 14
##   [1,] "park"    "prison"
##   [2,] "peopl"   "farmwork"
##   [3,] "citi"    "pesticid"
##   [4,] "health"  "sourc"
##   [5,] "green"   "popul"
##   [6,] "color"   "facil"
##   [7,] "law"     "project"
##   [8,] "project" "report"
##   [9,] "poor"    "center"
##  [10,] "includ"  "offic"
```

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

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

top_terms <- comment_topics %>%
  group_by(topic) %>%
  top_n(5, 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()
```



```
k <- 20
topicModel_k20 <- LDA(dfm, k, method="Gibbs", control=list(iter = 500, verbose = 25))
```

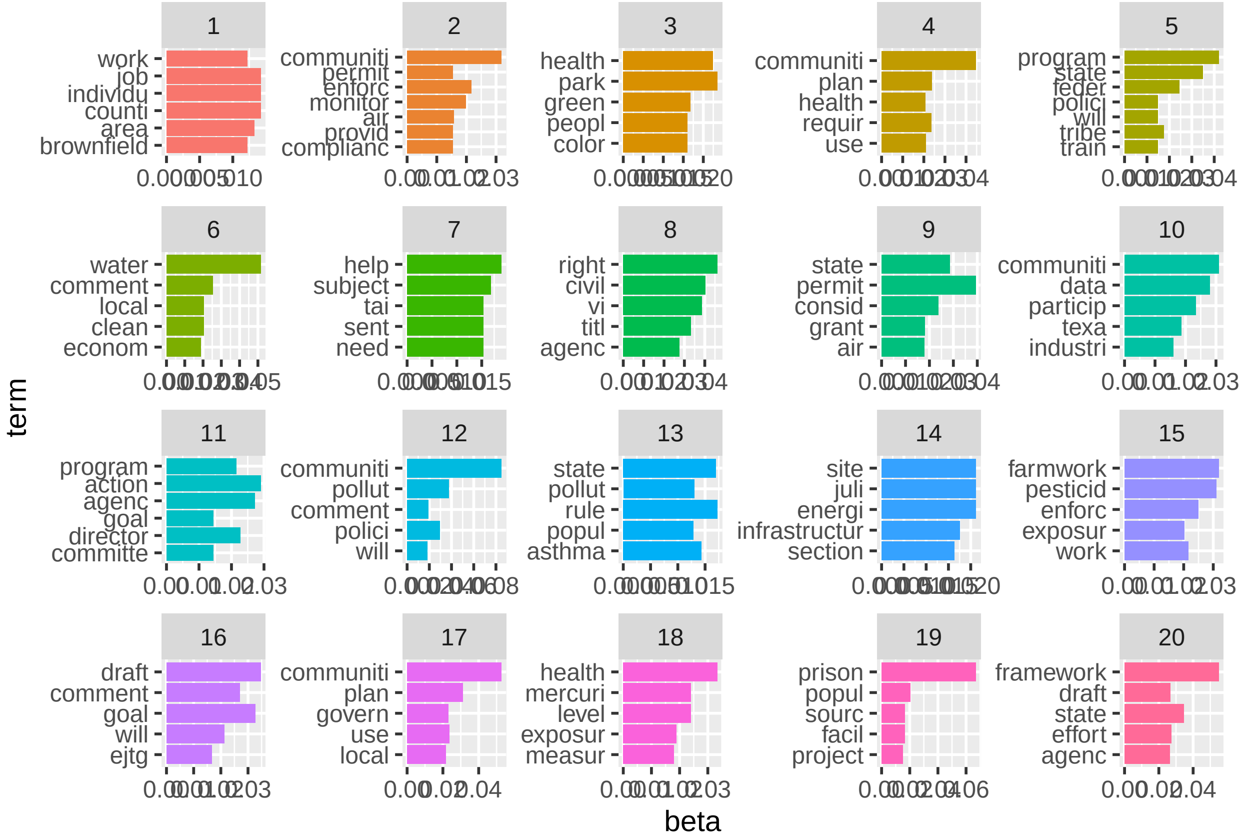
Try again for k = 20

```
## K = 20; 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_k20)
terms(topicModel_k20, 10)
```

```
##       Topic 1      Topic 2     Topic 3     Topic 4     Topic 5   Topic 6
##  [1,] "individu"   "communiti" "park"      "communiti" "program" "water"
##  [2,] "job"        "enforc"    "health"    "plan"      "state"   "comment"
##  [3,] "counti"     "monitor"   "green"     "requir"    "feder"   "local"
##  [4,] "area"       "air"       "peopl"     "use"       "tribe"   "clean"
##  [5,] "work"       "provid"    "color"     "health"    "will"    "econom"
##  [6,] "brownfield" "permit"    "law"       "comment"   "polici"  "fund"
##  [7,] "^"          "complianc" "see"       "includ"    "train"   "new"
##  [8,] "econom"     "report"    "complianc" "impact"    "requir"  "citi"
##  [9,] "can"        "action"    "citi"      "also"      "specif"  "popul"
## [10,] "increas"    "avail"     "access"    "concern"   "follow"  "drink"
##       Topic 7      Topic 8  Topic 9    Topic 10    Topic 11    Topic 12
##  [1,] "help"       "right"  "permit"   "communiti" "action"    "communiti"
##  [2,] "subject"    "civil"  "state"    "data"      "agenc"     "pollut"
##  [3,] "need"       "vi"     "consid"   "particip"  "director"  "polici"
##  [4,] "tai"        "titl"   "grant"    "texa"      "program"   "comment"
##  [5,] "sent"       "agenc"  "air"      "industri"  "goal"      "will"
##  [6,] ">"          "act"    "feder"    "process"   "committe"  "overburden"
##  [7,] "ejstrategi" "issu"   "carolina" "citizen"   "advisori"  "impact"
##  [8,] "<"          "nation" "use"      "resourc"   "includ"    "reduc"
##  [9,] "lung"       "feder"  "comment"  "comment"   "feder"     "air"
## [10,] "pm"         "impact" "north"    "permit"    "recommend" "new"
##       Topic 13   Topic 14        Topic 15    Topic 16   Topic 17    Topic 18
##  [1,] "rule"     "site"          "farmwork"  "draft"    "communiti" "health"
##  [2,] "state"    "energi"        "pesticid"  "goal"     "plan"      "mercuri"
##  [3,] "asthma"   "juli"          "enforc"    "comment"  "use"       "level"
##  [4,] "pollut"   "infrastructur" "work"      "will"     "govern"    "exposur"
##  [5,] "popul"    "section"       "exposur"   "ejtg"     "local"     "measur"
##  [6,] "ejscreen" "natur"         "agenc"     "may"      "land"      "depart"
##  [7,] "air"      "gas"           "offic"     "year"     "mani"      "hous"
##  [8,] "guidanc"  "pipelin"       "use"       "regul"    "group"     "pleas"
##  [9,] "provid"   "access"        "implement" "industri" "develop"   "attach"
## [10,] "avail"    "district"      "includ"    "busi"     "comment"   "contamin"
##       Topic 19     Topic 20
##  [1,] "prison"     "framework"
##  [2,] "popul"      "state"
##  [3,] "facil"      "effort"
##  [4,] "sourc"      "draft"
##  [5,] "project"    "agenc"
##  [6,] "peopl"      "epa"
##  [7,] "center"     "agenda"
##  [8,] "report"     "communiti"
##  [9,] "incarcer"   "develop"
## [10,] "california" "action"
```

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

comment_topics <- tidy(topicModel_k20, matrix = "beta")

top_terms <- comment_topics %>%
  group_by(topic) %>%
  top_n(5, 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()
```


term
beta
work
job
individu
counti
area
brownfield
communiti
permit
enforc
monitor
air
provid
complianc
health
park
green
peopl
color
communiti
plan
health
requir
use
program
state
feder
polici
will
tribe
train
water
comment
local
clean
econom
help
subject
tai
sent
need
right
civil
vi
titl
agenc
state
permit
consid
grant
air
communiti
data
particip
texa
industri
program
action
agenc
goal
director
committe
communiti
pollut
comment
polici
will
state
pollut
rule
popul
asthma
site
juli
energi
infrastructur
section
farmwork
pesticid
enforc
exposur
work
draft
comment
goal
will
ejtg
communiti
plan
govern
use
local
health
mercuri
level
exposur
measur
prison
popul
sourc
facil
project
framework
draft
state
effort
agenc