Topic 6: Topic Analysis Homework

Steven Cognac

2022-05-10

Load the data

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

Now we'll build and clean the corpus

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

Now let's tokenize our dataset and remove stopwords

```r
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

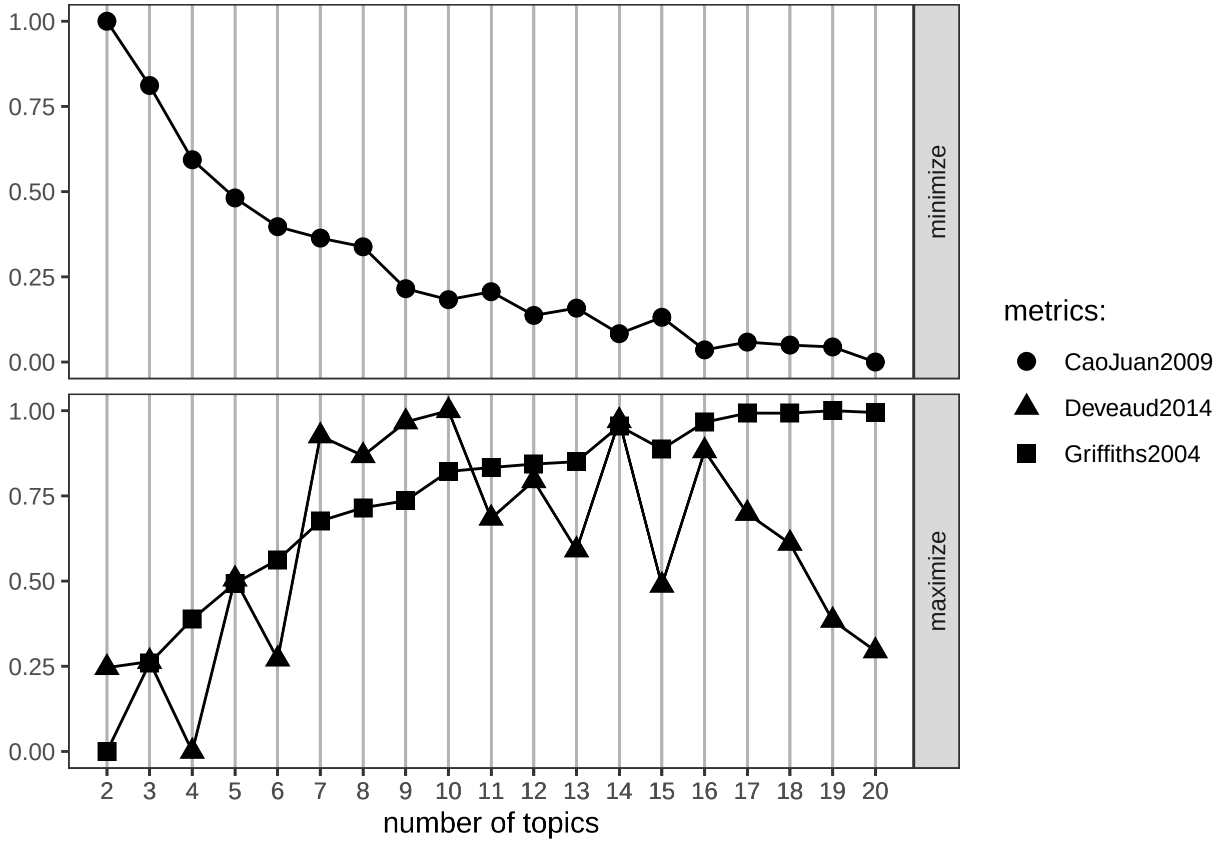
```r
dfm_comm<- dfm(toks1, tolower = TRUE)
dfm <- dfm_wordstem(dfm_comm)
dfm <- dfm_trim(dfm, min_docfreq = 2) #remove terms only appearing in one doc (min_termfreq = 10)

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

Find optimal number of topics

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

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

```
FindTopicsNumber_plot(result)
```



Selecting number of latent topics present from the comment letters

- Model 1 - **k=9** Latent Topics
- Model 2 - **k=4** Latent Topics
- Model 3 - **k=10** Latent topics
- Model 4 - **k=16** Latent topics

Picking k values to check

Model 1 (k=9) is the original number of topics selected in the lab based on the 9 EPA priority areas: Rulemaking, Permitting, Compliance and Enforcement, Science, States and Local Governments, Federal Agencies, Community-based Work, Tribes and Indigenous People, National Measures. Model 2 (k=4) was chosen from reviewing the 2017-2020 EPA Environmental Justice Report's where I noticed there are 4 major goals/themes: (1) delivering environmental results; (2) cooperative federalism; (3) rule of law and fair process; and (4) building community capacity and engagement. Model's 3 (k=10) and 4 (k=16) were based on the `FindTopicNumber()` optimization metrics. Based on the plots, 10 has the highest mximum value around where the minimization plot begins to descrease more slowely. In the minimization plot, 16 is where the plot begins to level out, indicating a stable number of topics is near. We also know 16 is a good option because there are 7 additional topics identified in the EPA's response to the public comments.

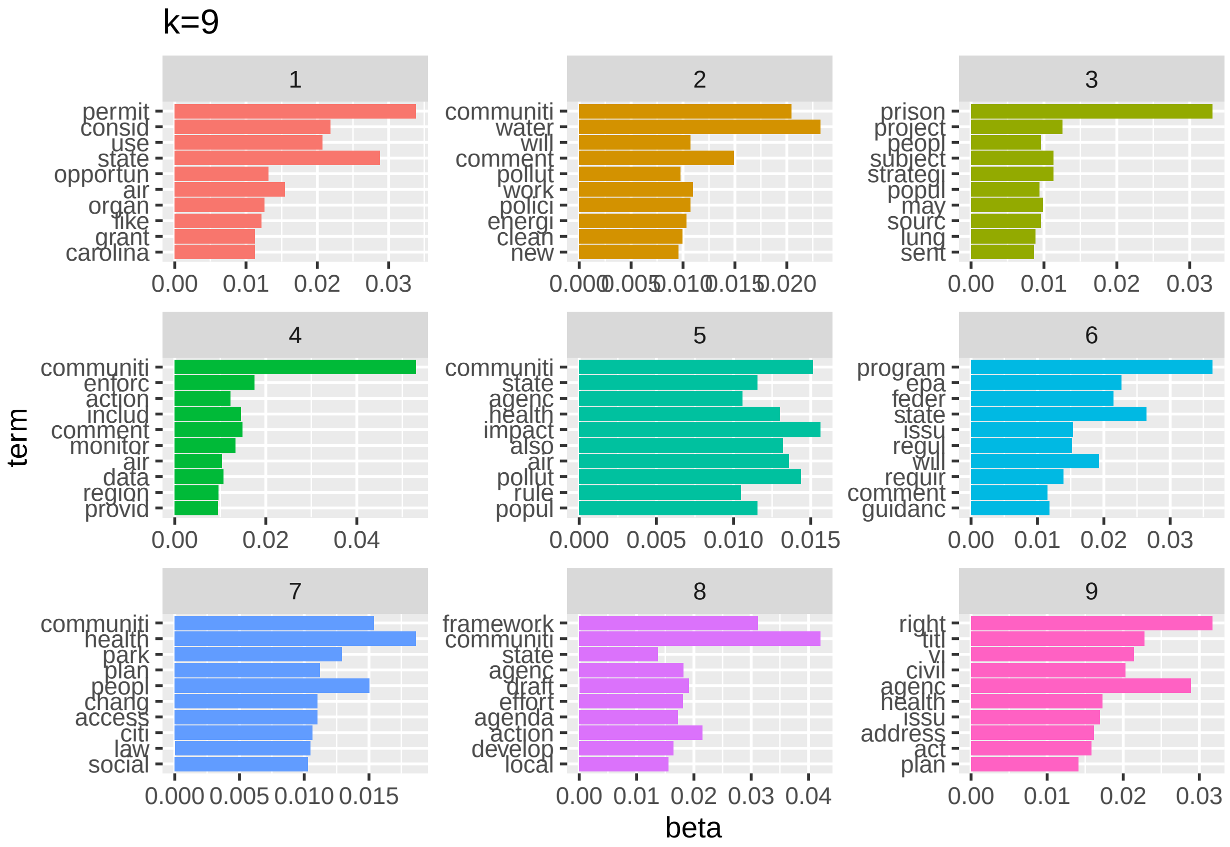
```
# select number of topics
k9 <- 9
k4 <- 4
k10 <- 10
k16 <- 16

# run LDA model
topicModel_k9 <- LDA(dfm, k9, method="Gibbs", control=list(iter = 500, verbose = 25))
topicModel_k4 <- LDA(dfm, k4, method="Gibbs", control=list(iter = 500, verbose = 25))
topicModel_k10 <- LDA(dfm, k10, method="Gibbs", control=list(iter = 500, verbose = 25))
topicModel_k16 <- LDA(dfm, k16, method="Gibbs", control=list(iter = 500, verbose = 25))

#nTerms(dfm_comm)
tmResult9 <- posterior(topicModel_k9)
tmResult4 <- posterior(topicModel_k4)
tmResult10 <- posterior(topicModel_k10)
tmResult16 <- posterior(topicModel_k16)
# attributes(tmResult9)


# nTerms(dfm_comm)
# get beta from results
theta9 <- tmResult9$topics
theta4 <- tmResult4$topics
theta10 <- tmResult10$topics
theta16 <- tmResult16$topics

beta <- tmResult9$terms
```

Let's pull out the top 10 likelihood / probability of frequency in each topic for each latent topic number. Remember, just because we choose those number of topics, it doesn't mean LDS actually picked up on the correct topic.

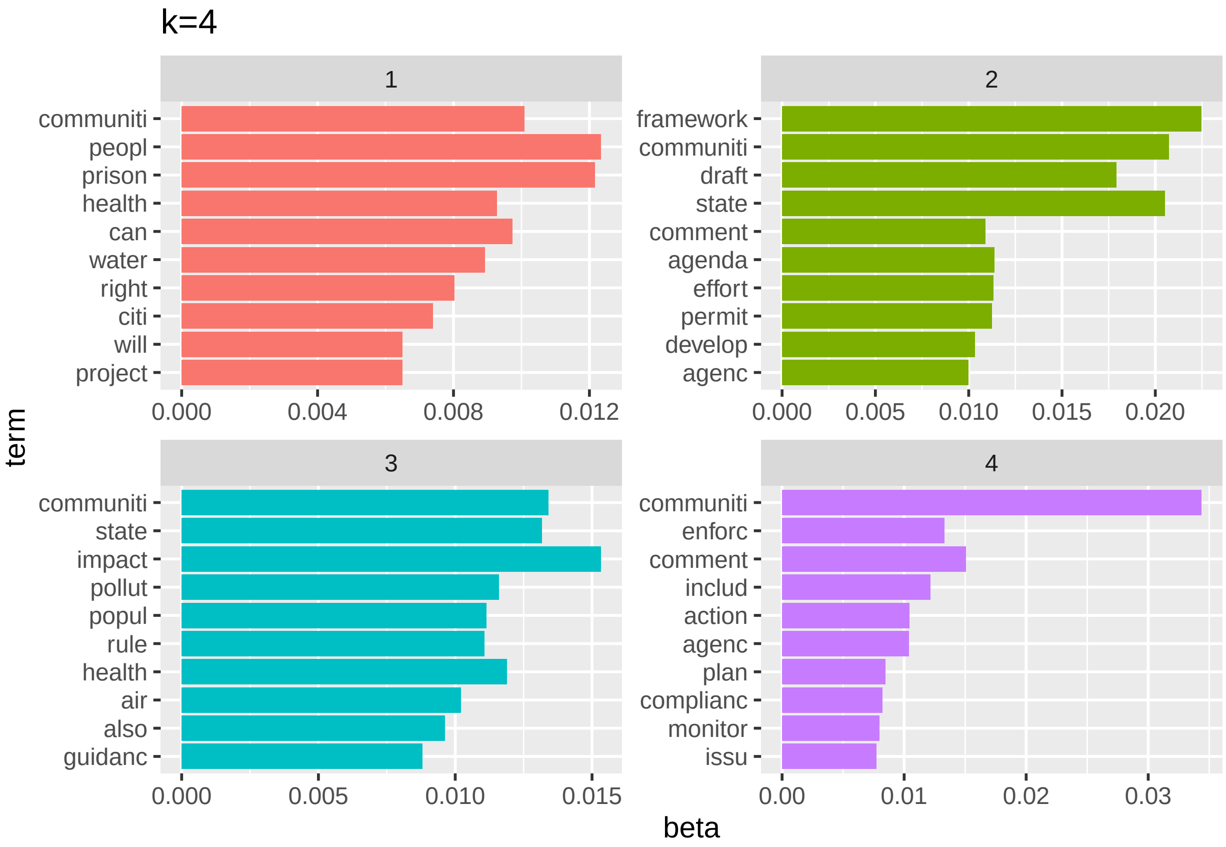
```
#terms(topicModel_k9, 10)
#terms(topicModel_k4, 10)
#terms(topicModel_k10, 10)
#terms(topicModel_k16, 10)
```

```
# tidy terms
comment_topics9 <- tidy(topicModel_k9, matrix = "beta")
comment_topics4 <- tidy(topicModel_k4, matrix = "beta")
comment_topics10 <- tidy(topicModel_k10, matrix = "beta")
comment_topics16 <- tidy(topicModel_k16, matrix = "beta")

top_terms9 <- comment_topics9 %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
```

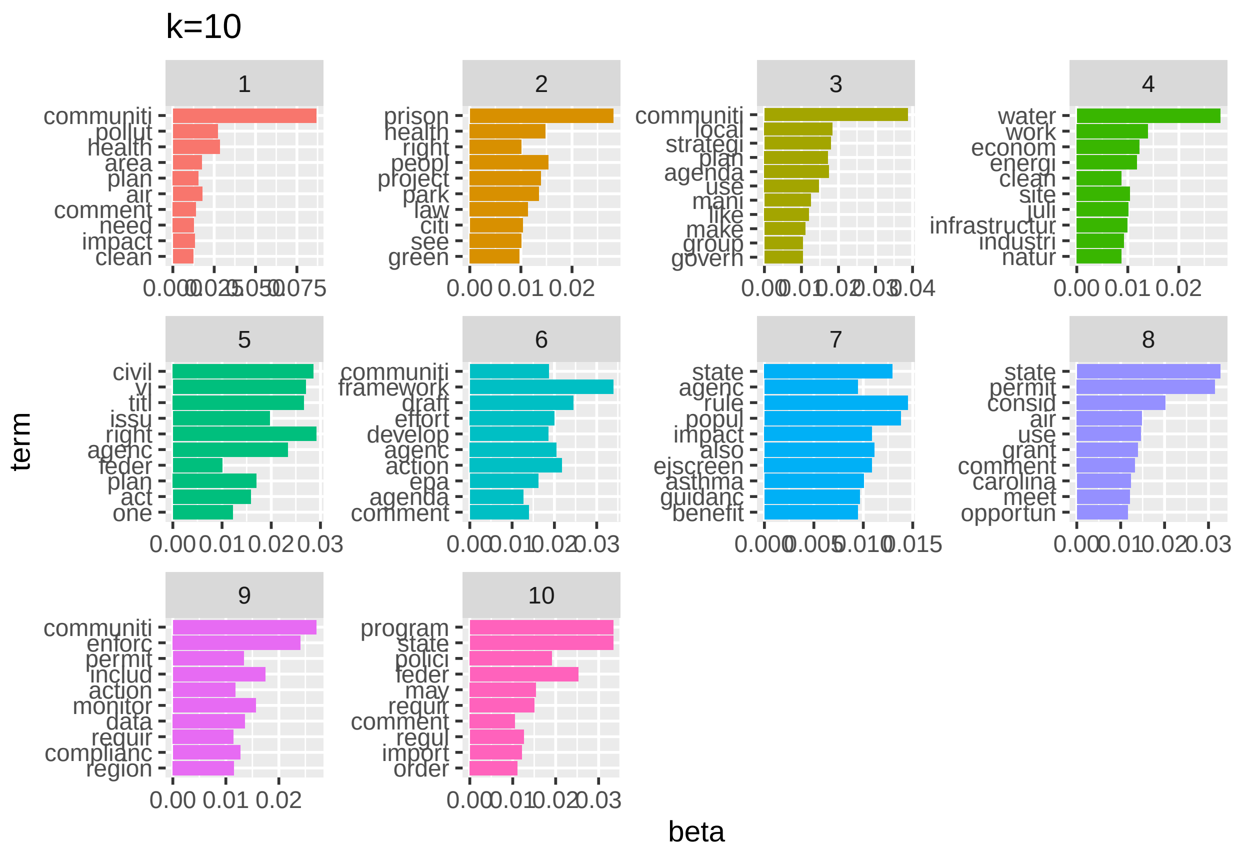
```r
  arrange(topic, -beta) %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  labs(title = "k=9")

top_terms4 <- comment_topics4 %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta) %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  labs(title = "k=4")

top_terms10 <- comment_topics10 %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta) %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  labs(title = "k=10")

top_terms16 <- comment_topics16 %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta) %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  labs(title = "k=16")

top_terms9
```


k=9
term
beta
1
permit
consid
use
state
opportun
air
organ
like
grant
carolina
0.00
0.01
0.02
0.03
2
communiti
water
will
comment
pollut
work
polici
energi
clean
new
0.000
0.005
0.010
0.015
0.020
3
prison
project
peopl
subject
strategi
popul
may
sourc
lung
sent
0.00
0.01
0.02
0.03
4
communiti
enforc
action
includ
comment
monitor
air
data
region
provid
0.00
0.02
0.04
5
communiti
state
agenc
health
impact
also
air
pollut
rule
popul
0.000
0.005
0.010
0.015
6
program
epa
feder
state
issu
regul
will
requir
comment
guidanc
0.00
0.01
0.02
0.03
7
communiti
health
park
plan
peopl
chang
access
citi
law
social
0.000
0.005
0.010
0.015
8
framework
communiti
state
agenc
draft
effort
agenda
action
develop
local
0.00
0.01
0.02
0.03
0.04
9
right
titl
vi
civil
agenc
health
issu
address
act
plan
0.00
0.01
0.02
0.03

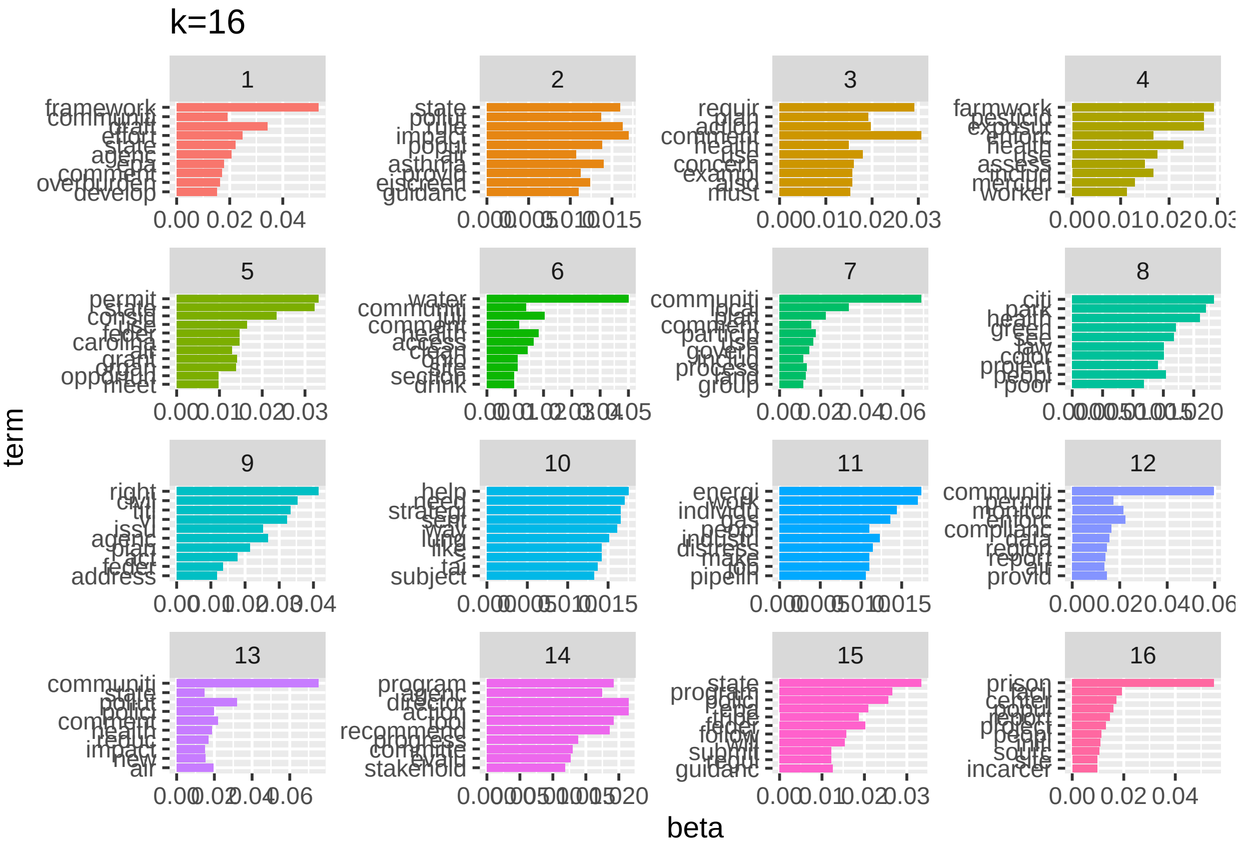
```
top_terms4
```


k=4
1
communiti
peopl
prison
health
can
water
right
citi
will
project
0.000
0.004
0.008
0.012
2
framework
communiti
draft
state
comment
agenda
effort
permit
develop
agenc
0.000
0.005
0.010
0.015
0.020
3
communiti
state
impact
pollut
popul
rule
health
air
also
guidanc
0.000
0.005
0.010
0.015
4
communiti
enforc
comment
includ
action
agenc
plan
complianc
monitor
issu
0.00
0.01
0.02
0.03
term
beta

```
top_terms10
```


k=10
1
communiti
pollut
health
area
plan
air
comment
need
impact
clean
0.000
0.025
0.050
0.075
2
prison
health
right
peopl
project
park
law
citi
see
green
0.00
0.01
0.02
3
communiti
local
strategi
plan
agenda
use
mani
like
make
group
govern
0.00
0.01
0.02
0.03
0.04
4
water
work
econom
energi
clean
site
juli
infrastructur
industri
natur
0.00
0.01
0.02
5
civil
vi
titl
issu
right
agenc
feder
plan
act
one
0.00
0.01
0.02
0.03
6
communiti
framework
draft
effort
develop
agenc
action
epa
agenda
comment
0.00
0.01
0.02
0.03
7
state
agenc
rule
popul
impact
also
ejscreen
asthma
guidanc
benefit
0.000
0.005
0.010
0.015
8
state
permit
consid
air
use
grant
comment
carolina
meet
opportun
0.00
0.01
0.02
0.03
9
communiti
enforc
permit
includ
action
monitor
data
requir
complianc
region
0.00
0.01
0.02
10
program
state
polici
feder
may
requir
comment
regul
import
order
0.00
0.01
0.02
0.03
term
beta

```
top_terms16
```



Let's assign names to the topics so we know what we are working with. We can name them by their top terms. Let's reassign the topic names to the top 5 words per topic.

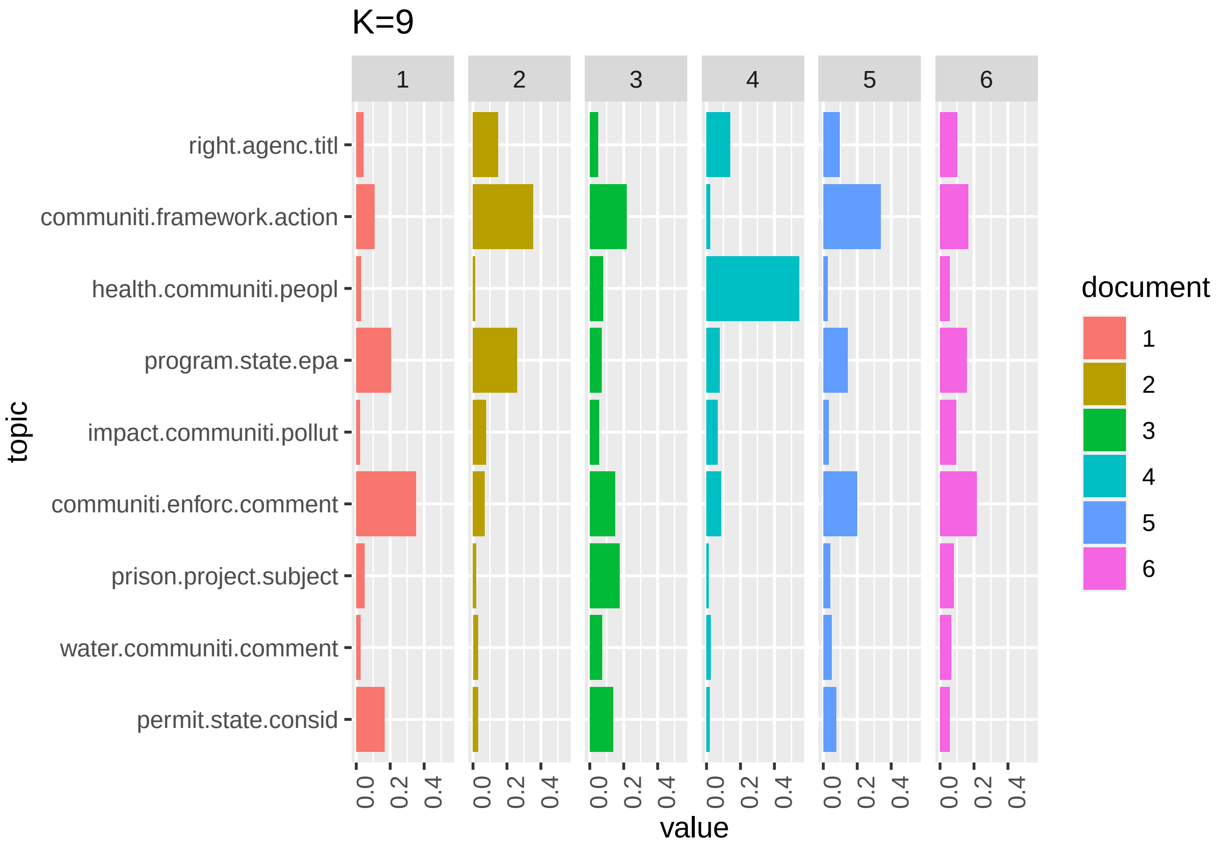
```
# re-assign names
top5termsPerTopic9 <- terms(topicModel_k9, 3)
top5termsPerTopic4 <- terms(topicModel_k4, 3)
top5termsPerTopic10 <- terms(topicModel_k10, 3)
top5termsPerTopic16 <- terms(topicModel_k16, 3)

# remove spaces
topicNames9 <- apply(top5termsPerTopic9, 2, paste, collapse=" ")
topicNames4 <- apply(top5termsPerTopic4, 2, paste, collapse=" ")
topicNames10 <- apply(top5termsPerTopic10, 2, paste, collapse=" ")
topicNames16 <- apply(top5termsPerTopic16, 2, paste, collapse=" ")
```

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

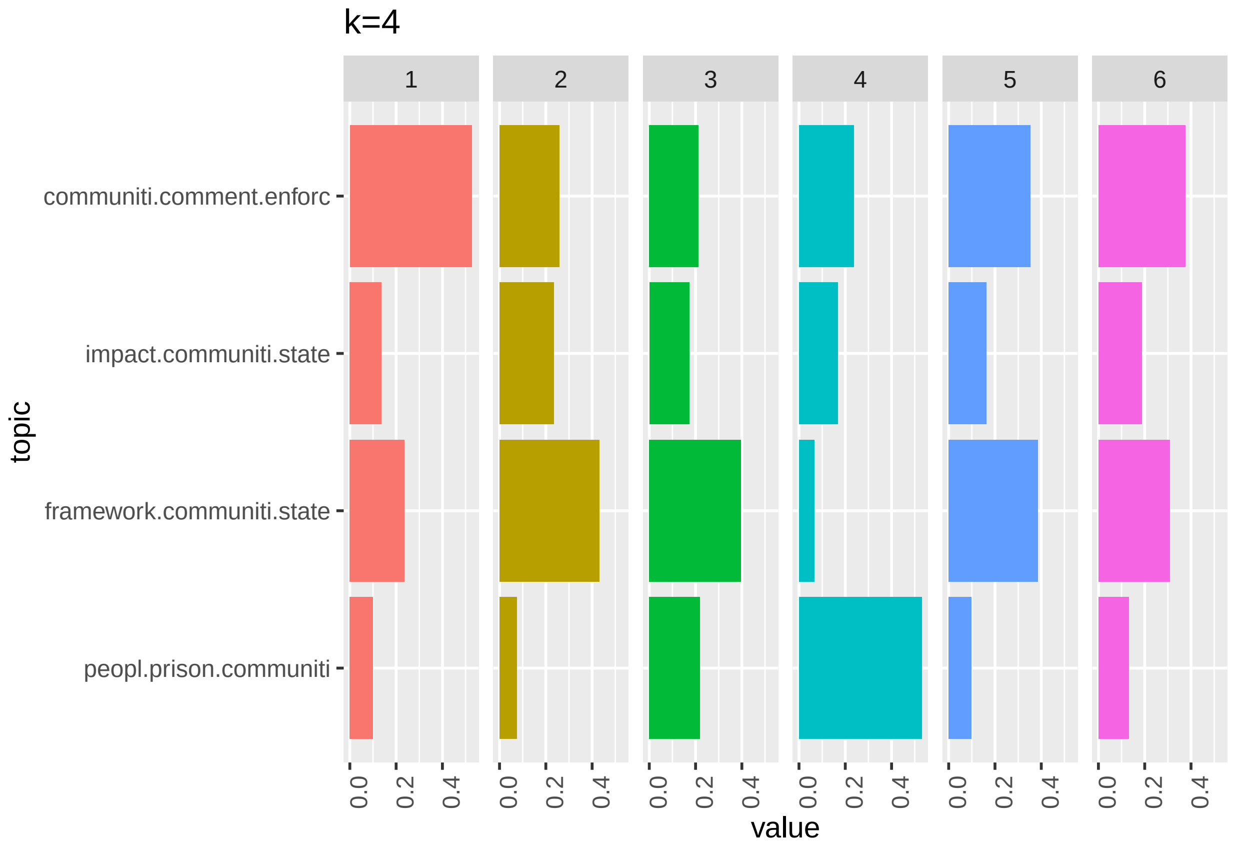
```
exampleIds <- c(1, 2, 3, 4, 5, 6)
N <- length(exampleIds)

# get topic proportions from example documents
topicProportionExamples9 <- theta9[exampleIds,]
colnames(topicProportionExamples9) <- topicNames9
vizDataFrame9 <- melt(cbind(data.frame(topicProportionExamples9), document=factor(1:N)), variable.name =

topicProportionExamples4 <- theta4[exampleIds,]
```

```
colnames(topicProportionExamples4) <- topicNames4
vizDataFrame4 <- melt(cbind(data.frame(topicProportionExamples4), document=factor(1:N)), variable.name =

topicProportionExamples10 <- theta10[exampleIds,]
colnames(topicProportionExamples10) <- topicNames10
vizDataFrame10 <- melt(cbind(data.frame(topicProportionExamples10), document=factor(1:N)), variable.name

topicProportionExamples16 <- theta16[exampleIds,]
colnames(topicProportionExamples16) <- topicNames16
vizDataFrame16 <- melt(cbind(data.frame(topicProportionExamples16), document=factor(1:N)), variable.name
```

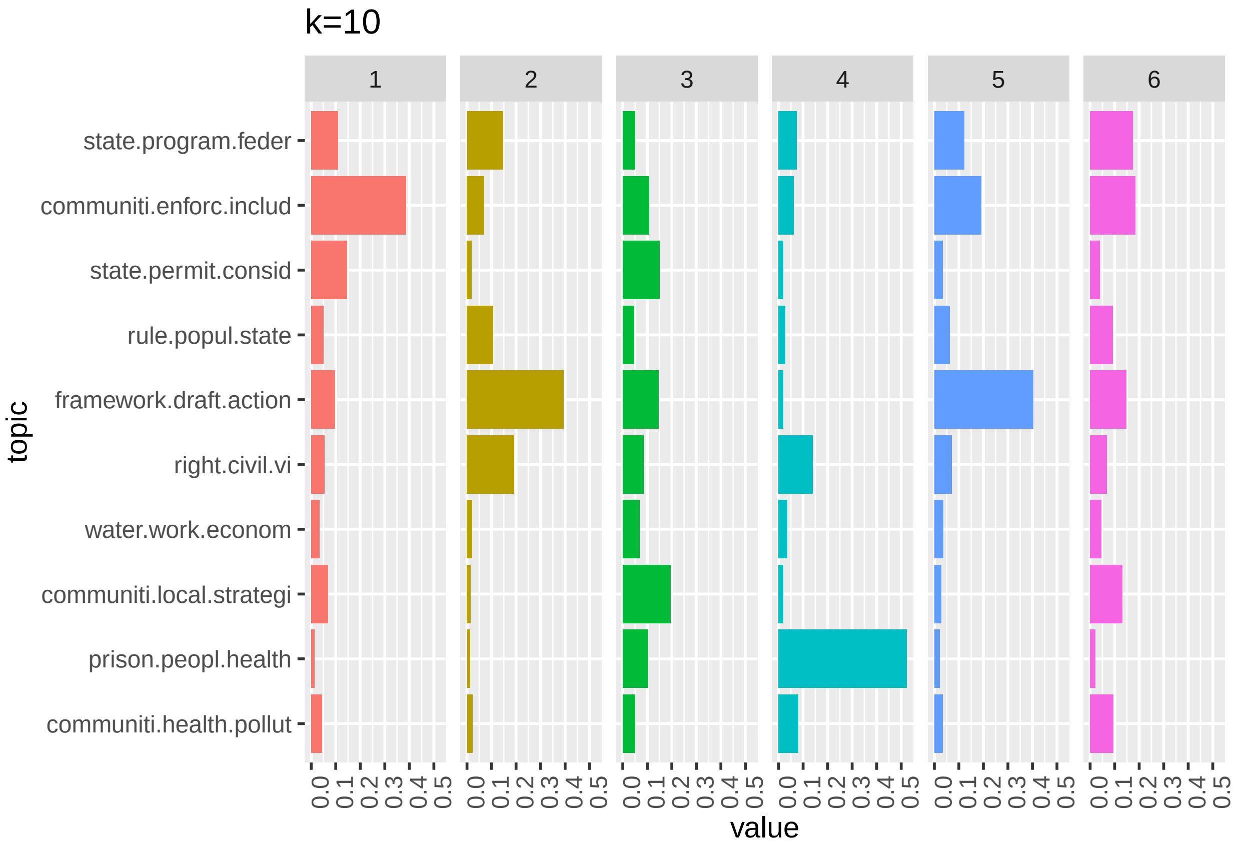
```
# plot
terms9 <- ggplot(data = vizDataFrame9, aes(topic, value, fill = document), ylab = "proportion") +
  geom_bar(stat="identity") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  coord_flip() +
  facet_wrap(~ document, ncol = N) +
  labs(title = "K=9")

terms4 <- ggplot(data = vizDataFrame4, 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) +
  theme(legend.position="none") +
  labs(title = "k=4")

terms10 <- ggplot(data = vizDataFrame10, 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) +
  theme(legend.position="none") +
  labs(title = "k=10")

terms16 <- ggplot(data = vizDataFrame16, 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) +
  theme(legend.position="none") +
  labs(title = "k=16")

terms9
```

K=9
1
2
3
4
5
6
right.agenc.titl
communiti.framework.action
health.communiti.peopl
program.state.epa
impact.communiti.pollut
communiti.enforc.comment
prison.project.subject
water.communiti.comment
permit.state.consid
topic
0.0
0.2
0.4
value
document
1
2
3
4
5
6

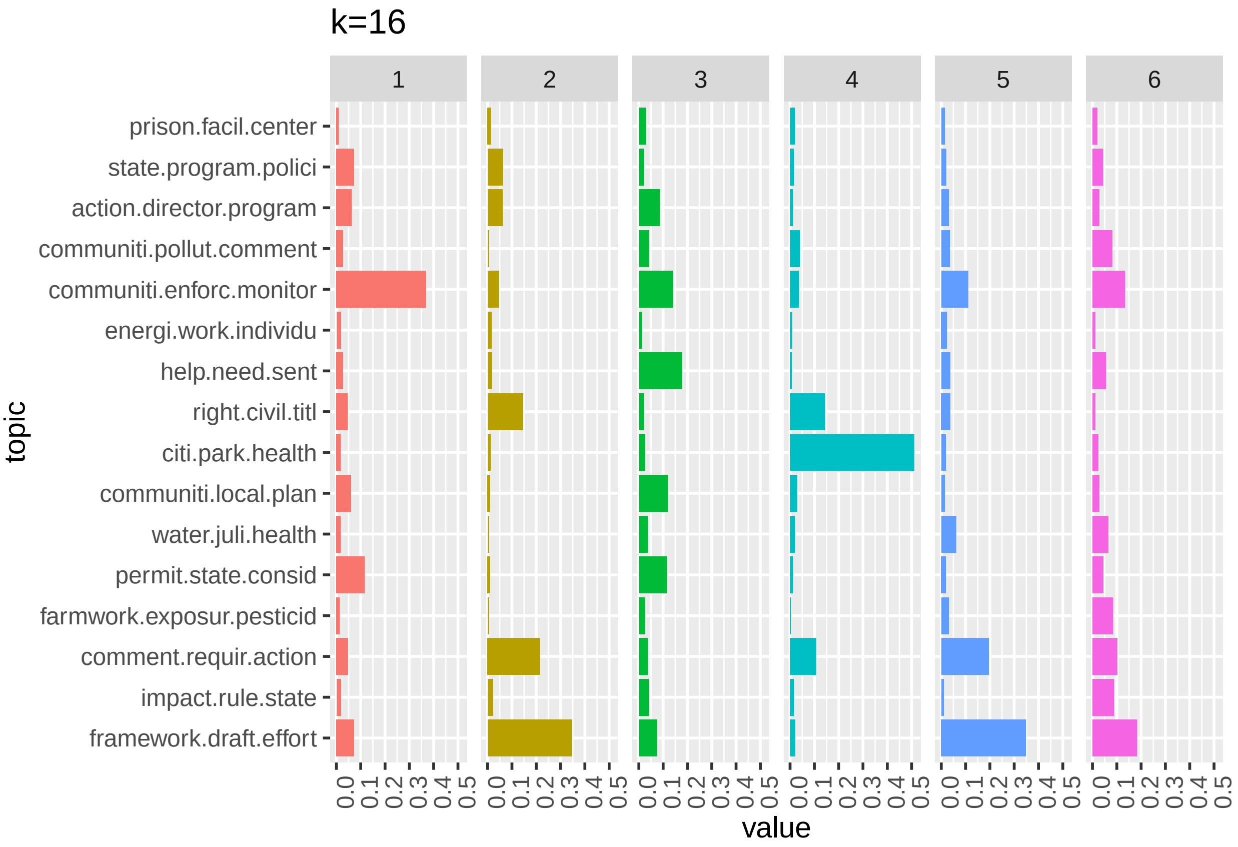
```
terms4
```


k=4
1
2
3
4
5
6
topic
communiti.comment.enforc
impact.communiti.state
framework.communiti.state
peopl.prison.communiti
0.0
0.2
0.4
value

```
terms10
```


k=10
1
2
3
4
5
6
topic
state.program.feder
communiti.enforc.includ
state.permit.consid
rule.popul.state
framework.draft.action
right.civil.vi
water.work.econom
communiti.local.strategi
prison.peopl.health
communiti.health.pollut
0.0
0.1
0.2
0.3
0.4
0.5
value

```
terms16
```



```r
svd_tsne <- function(x){
  tsne(svd(x)$u)
}

json9 <- createJSON(
  phi = tmResult9$terms,
  theta = tmResult9$topics,
  doc.length = rowSums(dfm),
  vocab = colnames(dfm),
  term.frequency = colSums(dfm),
  mds.method = svd_tsne,
  plot.opts = list(xlab="",
                   ylab="")
)

json4 <- createJSON(
  phi = tmResult4$terms,
  theta = tmResult4$topics,
  doc.length = rowSums(dfm),
  vocab = colnames(dfm),
  term.frequency = colSums(dfm),
  mds.method = svd_tsne,
  plot.opts = list(xlab="",
                   ylab="")
)
```

```
json10 <- createJSON(
  phi = tmResult10$terms,
  theta = tmResult10$topics,
  doc.length = rowSums(dfm),
  vocab = colnames(dfm),
  term.frequency = colSums(dfm),
  mds.method = svd_tsne,
  plot.opts = list(xlab="",
                   ylab="")
)

json16 <- createJSON(
  phi = tmResult16$terms,
  theta = tmResult16$topics,
  doc.length = rowSums(dfm),
  vocab = colnames(dfm),
  term.frequency = colSums(dfm),
  mds.method = svd_tsne,
  plot.opts = list(xlab="",
                   ylab="")
)

#serVis(json9)
#serVis(json4)
#serVis(json10)
#serVis(json16)
```

Picking the best model k-value

Based on the LDAvis and looking at the graphs of the top terms in each topic I would move forward with using 16 as the overall best value for k. When looking at the term plots for k=4, we get a lot of overlap between topics and minimual variation in unique topic subjects. When looking at k=10 as compared to k=16, there were again too much topic overlap with k=10. With the interactive LDAvis plots, the size of k=16 circles vary in size much more than k=10, indicating greater topic distriution.