EDS 231: Assignment 5: Topic Analysis

Scout Leonard

05/10/2022

Load Libraries

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

Set Up

Read in data:

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

Corpus:

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

epa_corp.stats <- summary(epa_corp)

head(epa_corp.stats, n = 5) %>%
  kable()
```

Text	Types	Tokens	Sentences	Document
text1	1196	3973	178	1_Air Alliance.pdf
text2	830	2509	111	10_Bus NEJ.pdf
text3	279	571	31	11_Carlton Ginny.pdf
text4	1745	6904	251	15_City Project.pdf
text5	581	1534	49	16_Corporate EEC.pdf

Tokenize Corpus:

```
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")
```

Convert tokens to a document frame matrix:

```
#construct dfm from tokens
dfm_comm <- dfm(toks1,
                tolower = TRUE)

#apply a stemmer to words in dfm
dfm <- dfm_wordstem(dfm_comm)

#remove terms only appearing in one doc (min_termfreq = 10)
dfm <- dfm_trim(dfm,
                min_docfreq = 2)

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

#comments_df <- dfm[sel_idx, ]
dfm <- dfm[sel_idx, ]
```

LDA Modelling:

Write the model:

```
topicModel_k9 <- LDA(dfm,
                     k = 9,
                     method = "Gibbs",
                     control = list(iter = 500, verbose = 25))
```

```
## K = 9; 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!
```

Return results:

```
tmResult <- posterior(topicModel_k9)

beta <- tmResult$terms #get beta from results

terms(topicModel_k9, 10)
```

```
##       Topic 1     Topic 2     Topic 3    Topic 4     Topic 5
##  [1,] "communiti" "communiti" "prison"   "pollut"    "water"
##  [2,] "plan"      "enforc"    "facil"    "state"     "communiti"
##  [3,] "local"     "monitor"   "new"      "rule"      "econom"
##  [4,] "use"       "air"       "peopl"    "impact"    "energi"
##  [5,] "strategi"  "comment"   "center"   "communiti" "comment"
##  [6,] "comment"   "provid"    "report"   "air"       "clean"
##  [7,] "govern"    "requir"    "project"  "also"      "infrastructur"
##  [8,] "help"      "permit"    "sourc"    "health"    "counti"
##  [9,] "work"      "pollut"    "contamin" "popul"     "site"
## [10,] "particip"  "data"      "problem"  "plan"      "area"
##       Topic 6      Topic 7     Topic 8   Topic 9
##  [1,] "framework"  "health"    "agenc"   "state"
##  [2,] "communiti"  "park"      "issu"    "permit"
##  [3,] "draft"      "access"    "program" "feder"
##  [4,] "effort"     "peopl"     "feder"   "consid"
##  [5,] "action"     "right"     "titl"    "comment"
##  [6,] "develop"    "green"     "right"   "organ"
##  [7,] "agenda"     "citi"      "epa"     "use"
##  [8,] "agenc"      "see"       "includ"  "meet"
##  [9,] "overburden" "color"     "vi"      "air"
## [10,] "state"      "communiti" "civil"   "requir"
```

Visualize results:

```
#load libraries
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)
```

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 = 5`

Write model:

```
topicModel_k5 <- LDA(dfm,
                     k = 5,
                     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!
```

Return results:

```
tmResult_k5 <- posterior(topicModel_k5)

beta <- tmResult_k5$terms #get beta from results

terms(topicModel_k5, 10)
```

```
##       Topic 1     Topic 2     Topic 3     Topic 4   Topic 5
##  [1,] "communiti" "state"     "state"     "right"   "communiti"
##  [2,] "plan"      "pollut"    "framework" "civil"   "enforc"
##  [3,] "local"     "impact"    "draft"     "health"  "comment"
##  [4,] "water"     "communiti" "agenc"     "prison"  "includ"
##  [5,] "can"       "rule"      "program"   "vi"      "action"
##  [6,] "strategi"  "also"      "epa"       "includ"  "monitor"
##  [7,] "agenda"    "health"    "permit"    "peopl"   "region"
##  [8,] "econom"    "air"       "consid"    "titl"    "need"
##  [9,] "comment"   "popul"     "effort"    "project" "use"
## [10,] "work"      "guidanc"   "goal"      "nation"  "permit"
```

Visualize Results:

```
json_k5 <- createJSON(
  phi = tmResult_k5$terms,
  theta = tmResult_k5$topics,
  doc.length = rowSums(dfm),
  vocab = colnames(dfm),
  term.frequency = colSums(dfm),
  mds.method = svd_tsne,
  plot.opts = list(xlab="",
                   ylab="")
)

serVis(json_k5)
```

Model 2: `k = 20`

Write model:

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

```
## 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!
```

Return results:

```
tmResult_k20 <- posterior(topicModel_k20)

beta <- tmResult_k20$terms #get beta from results

terms(topicModel_k20, 10)
```

```
##       Topic 1    Topic 2     Topic 3      Topic 4      Topic 5   Topic 6
##  [1,] "requir"   "communiti" "like"       "framework"  "right"   "air"
##  [2,] "work"     "comment"   "sent"       "draft"      "agenc"   "health"
##  [3,] "mercuri"  "use"       "need"       "effort"     "titl"    "area"
##  [4,] "individu" "provid"    "help"       "epa"        "vi"      "issu"
##  [5,] "econom"   "exampl"    "thank"      "agenda"     "civil"   "clean"
##  [6,] "distress" "concern"   "tai"        "develop"    "act"     "standard"
##  [7,] "year"     "will"      ">"          "impact"     "feder"   "can"
##  [8,] "educ"     "includ"    "ejstrategi" "overburden" "plan"    "sourc"
##  [9,] "peopl"    "program"   "strategi"   "opportun"   "issu"    "risk"
## [10,] "make"     "also"      "<"          "lee"        "address" "address"
##       Topic 7    Topic 8      Topic 9    Topic 10  Topic 11    Topic 12
##  [1,] "data"     "communiti"  "plan"     "prison"  "communiti" "state"
##  [2,] "particip" "pollut"     "process"  "facil"   "local"     "comment"
##  [3,] "texa"     "polici"     "govern"   "popul"   "injustic"  "action"
##  [4,] "feder"    "overburden" "resourc"  "project" "sustain"   "enforc"
##  [5,] "citizen"  "will"       "use"      "report"  "work"      "epa"
##  [6,] "resourc"  "reduct"     "strategi" "center"  "impact"    "work"
##  [7,] "air"      "reduc"      "develop"  "sourc"   "fund"      "import"
##  [8,] "process"  "power"      "particip" "one"     "administr" "support"
##  [9,] "success"  "protect"    "land"     "impact"  "provid"    "within"
## [10,] "access"   "new"        "engag"    "peopl"   "goal"      "complianc"
##       Topic 13   Topic 14    Topic 15    Topic 16   Topic 17    Topic 18
##  [1,] "impact"   "farmwork"  "permit"    "program"  "communiti" "park"
##  [2,] "rule"     "health"    "state"     "requir"   "monitor"   "health"
##  [3,] "state"    "pesticid"  "consid"    "tribe"    "enforc"    "peopl"
##  [4,] "popul"    "work"      "grant"     "polici"   "permit"    "see"
##  [5,] "asthma"   "exposur"   "carolina"  "regul"    "report"    "citi"
##  [6,] "ejscreen" "polici"    "meet"      "guidanc"  "requir"    "green"
##  [7,] "guidanc"  "agenc"     "north"     "will"     "action"    "color"
##  [8,] "pollut"   "implement" "framework" "follow"   "includ"    "project"
##  [9,] "also"     "risk"      "use"       "principl" "complianc" "area"
## [10,] "ozon"     "includ"    "air"       "feder"    "assess"    "space"
##       Topic 19   Topic 20
##  [1,] "agenc"    "water"
##  [2,] "consid"   "site"
##  [3,] "director" "juli"
##  [4,] "health"   "clean"
##  [5,] "action"   "job"
##  [6,] "program"  "counti"
##  [7,] "develop"  "brownfield"
##  [8,] "scienc"   "health"
##  [9,] "tool"     "comment"
## [10,] "committe" "can"
```

Visualize results:

```
json_k20 <- createJSON(
  phi = tmResult_k20$terms,
  theta = tmResult_k20$topics,
  doc.length = rowSums(dfm),
  vocab = colnames(dfm),
  term.frequency = colSums(dfm),
  mds.method = svd_tsne,
  plot.opts = list(xlab="",
                   ylab="")
)

serVis(json_k20)
```

Model 3: `k = 3`

Write model:

```
topicModel_k3 <- LDA(dfm,
                     k = 3,
                     method = "Gibbs",
                     control = list(iter = 500, verbose = 25))
```

```
## K = 3; 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!
```

Return results:

```
tmResult_k3 <- posterior(topicModel_k3)

beta <- tmResult_k3$terms #get beta from results

terms(topicModel_k3, 10)
```

```
##       Topic 1     Topic 2     Topic 3
##  [1,] "communiti" "communiti" "communiti"
##  [2,] "state"     "right"     "pollut"
##  [3,] "framework" "peopl"     "health"
##  [4,] "agenc"     "plan"      "air"
##  [5,] "comment"   "civil"     "provid"
##  [6,] "draft"     "health"    "impact"
##  [7,] "permit"    "prison"    "comment"
##  [8,] "action"    "water"     "state"
##  [9,] "program"   "project"   "enforc"
## [10,] "polici"    "can"       "also"
```

Visualize results:

```
json_k3 <- createJSON(
  phi = tmResult_k3$terms,
  theta = tmResult_k3$topics,
  doc.length = rowSums(dfm),
  vocab = colnames(dfm),
  term.frequency = colSums(dfm),
  mds.method = svd_tsne,
  plot.opts = list(xlab="",
                   ylab="")
)

serVis(json_k3)
```

I think that overall, my model with k = 3 was the best number of topics for any of the models I tried. I think this based on the ratio of overall term freqnecy to estimated term frequency with the selected topic from the model, as output by the `serVis()` function. For the model with `k = 3`, these seemed to be the closest compared to the models with `k = 5` and `k = 20`.

Also, there are fewer topics with overlapping dimensions according to the intertopic distance map for the `k = 3` model compared to the other two models. `k = 5` similarly had more intertopic distance, but the term salience was not as strong as described in the previous paragraph.

Finally, the predicted topics terms outputs make the most sense to me for `k = 3`. I can see that the topics have some sort of meaning - one for administrative topics like different levels of agencies and kinds of policies and procedures they make, one for public health and the environment terms, and one for civil rights and the environment terms.