EDS 231: Assignment 1

Scout Leonard

05/06/2022

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 = 10) %>%
  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
text6	469	1187	53	17_Detriot Sierra Club.pdf
text7	424	903	38	18_District DOE.pdf
text8	3622	22270	655	19_Earth Justice.pdf
text9	373	717	25	2_Alex Kidd.pdf

Text	Types	Tokens	Sentences	Document
text10	404	971	42	20_Elizabeth Mooney.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)

print(head(dfm)) %>%
  kable()
```

```
## 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 ]
```

|| || || ||

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

Model 2

Model 3