Week 5: Assignment 4: Word relationship analysis

Scout Leonard

2022-04-27

Load Libraries

```
packages=c("tidyr",
           "pdftools",
           "lubridate",
           "tidyverse",
           "tidytext",
           "readr",
           "quanteda",
           "readtext",
           "quanteda.textstats",
           "quanteda.textplots",
           "ggplot2",
           "forcats",
           "stringr",
           "quanteda.textplots",
           "widyr",
           "igraph",
           "ggraph",
           "here")

for (i in packages) {
  if (require(i,character.only=TRUE)==FALSE) {
    install.packages(i,repos='http://cran.us.r-project.org')
  }
  else {
    require(i,character.only=TRUE)
  }
}
```

Read in data

```
files <- list.files(path = here("data/week5_data/"),
                    pattern = "*pdf$",
                    full.names = TRUE)

ej_reports <- lapply(files, pdf_text)

ej_pdf <- readtext(file = here("data/week5_data/*.pdf"),
                   docvarsfrom = "filenames",
                   docvarnames = c("type", "subj", "year"),
                   sep = "_")

#creating an initial corpus containing our data
epa_corp <- corpus(x = ej_pdf, text_field = "text" )
summary(epa_corp)
```

```
## Corpus consisting of 6 documents, showing 6 documents:
##
##              Text Types Tokens Sentences type subj year
##  EPA_EJ_2015.pdf  2136   8944       263  EPA   EJ 2015
##  EPA_EJ_2016.pdf  1599   7965       176  EPA   EJ 2016
##  EPA_EJ_2017.pdf  2774  16658       447  EPA   EJ 2017
##  EPA_EJ_2018.pdf  3973  30564       653  EPA   EJ 2018
##  EPA_EJ_2019.pdf  3773  22648       672  EPA   EJ 2019
##  EPA_EJ_2020.pdf  4493  30523       987  EPA   EJ 2020
```

```
#I'm adding some additional, context-specific stop words to stop word lexicon
more_stops <-c("2015","2016", "2017", "2018",
               "2019", "2020", "www.epa.gov", "https")
add_stops<- tibble(word = c(stop_words$word, more_stops))
stop_vec <- as_vector(add_stops)
```

```
#convert to tidy format and apply my stop words
raw_text <- tidy(epa_corp)

#Distribution of most frequent words across documents
raw_words <- raw_text %>%
  mutate(year = as.factor(year)) %>%
  unnest_tokens(word, text) %>%
  anti_join(add_stops, by = 'word') %>%
  count(year, word, sort = TRUE)

#number of total words by document
total_words <- raw_words %>%
  group_by(year) %>%
  summarize(total = sum(n))

report_words <- left_join(raw_words, total_words)
```

```
## Joining, by = "year"
```

```
par_tokens <- unnest_tokens(raw_text,
                            output = paragraphs,
                            input = text,
                            token = "paragraphs")

par_tokens <- par_tokens %>%
 mutate(par_id = 1:n())

par_words <- unnest_tokens(par_tokens,
                           output = word,
                           input = paragraphs,
                           token = "words")
```

Part 1

What are the most frequent trigrams in the dataset? How does this compare to the most frequent bigrams? Which n-gram seems more informative here, and why?

Part 2

Choose a new focal term to replace “justice” and recreate the correlation table and network (see corr_paragraphs and corr_network chunks). Explore some of the plotting parameters in the cor_network chunk to see if you can improve the clarity or amount of information your plot conveys. Make sure to use a different color for the ties!

Part 3

Write a function that allows you to conduct a keyness analysis to compare two individual EPA reports (hint: that means target and reference need to both be individual reports). Run the function on 3 pairs of reports, generating 3 keyness plots.

Part 4

Select a word or multi-word term of interest and identify words related to it using windowing and keyness comparison. To do this you will create to objects: one containing all words occurring within a 10-word window of your term of interest, and the second object containing all other words. Then run a keyness comparison on these objects. Which one is the target, and which the reference? Hint