Assignment 01: Text Data in R

Julia Parish

2022/04/11

Text Data in R

This text analysis was completed as an assignment for the course, Environmental Data Science 231: Text and Sentiment Analysis for Environmental Problems. The data was sourced from the New York Times using the New York Times API.

Load Packages

```
library(jsonlite) #convert results from API queries into R-friendly formats
library(tidyverse)
library(tidytext) #text data management and analysis
library(ggplot2) #plot word frequencies and publication dates
```

Key Word Selection & Query NY Times API

Pick an interesting environmental key word(s) and use the jsonlite package to query the API. Pick something high profile enough and over a large enough time frame that your query yields enough articles for an interesting examination.

```
# the from JSON flatten the JSON object
coalash <- fromJSON("http://api.nytimes.com/svc/search/v2/articlesearch.json?q=coal+ash&api-key=qUngwwal
```

```
# convert to a data frame
coalash <- coalash %>%
  data.frame()

#Inspect data
class(coalash)
```

```
## [1] "data.frame"
```

```
dim(coalash) # how big is it? 10 articles, 34 variables
```

```
## [1] 10 34
```

```
names(coalash) # list of variables
```

```
##  [1] "status"
##  [2] "copyright"
##  [3] "response.docs.abstract"
##  [4] "response.docs.web_url"
##  [5] "response.docs.snippet"
##  [6] "response.docs.lead_paragraph"
##  [7] "response.docs.print_section"
```

```
##  [8] "response.docs.print_page"
##  [9] "response.docs.source"
## [10] "response.docs.multimedia"
## [11] "response.docs.keywords"
## [12] "response.docs.pub_date"
## [13] "response.docs.document_type"
## [14] "response.docs.news_desk"
## [15] "response.docs.section_name"
## [16] "response.docs.type_of_material"
## [17] "response.docs._id"
## [18] "response.docs.word_count"
## [19] "response.docs.uri"
## [20] "response.docs.slideshow_credits"
## [21] "response.docs.subsection_name"
## [22] "response.docs.headline.main"
## [23] "response.docs.headline.kicker"
## [24] "response.docs.headline.content_kicker"
## [25] "response.docs.headline.print_headline"
## [26] "response.docs.headline.name"
## [27] "response.docs.headline.seo"
## [28] "response.docs.headline.sub"
## [29] "response.docs.byline.original"
## [30] "response.docs.byline.person"
## [31] "response.docs.byline.organization"
## [32] "response.meta.hits"
## [33] "response.meta.offset"
## [34] "response.meta.time"
```

Recreate the publications per day and word frequency plots using the first paragraph

Make some (at least 3) transformations to the corpus (add stopword(s), stem a key term and its variants, remove numbers)

Recreate the publications per day and word frequency plots using the headlines variable (response.docs.headline.main). Compare the distributions of word frequencies between the first paragraph and headlines. Do you see any difference?