Session 3:Create Corpus

Lucien Baumgartner

5/27/2021

```
library(jsonlite)
library(stringi)
library(tidyverse)
library(quanteda)
library(readtext)

rm(list = ls())
```

```
setwd('~/coliphi21/create_corpora/src')
```

```
## the SEP corpus
df <- readtext('../output/stanfordEnc/txt/*.json', text_field = "body.text", verbosity = 0)
df <- corpus(df)
docvars(df) <- mutate(docvars(df), doc = docnames(df))
df <- corpus_subset(df, !duplicated(docvars(df)$url))
docvars(df) <- mutate(docvars(df), doc = gsub('\\.[0-9]', '', doc))

## the university metadata
path <- "../output/philpeople"
files <- dir(path, pattern = "*.json")

meta <- files %>%
  map_df(~mutate(fromJSON(file.path(path, .), flatten = TRUE), doc = .))
sort(table(meta$geo.type), decreasing = T)
```

```
##
##              university                  museum             information                 college
##                    1455                     227                     153                      96
##                optician                    cafe educational_institution              tram_stop
##                       7                       6                       6                       6
##                  retail   sustainability issues                  ticket                    town
##                       2                       2                       2                       2
```

```
length(unique(meta$doc))
```

```
## [1] 1712
```

```
meta <-
  meta %>%
  group_by(doc) %>%
  filter(geo.type%in%c('university', 'college', 'community_centre', 'dormitory') | (geo.type%in%c('stati
  slice(1, .preserve = T)

## join
```

```
table(meta$doc%in%df$doc)
```

```
##
## TRUE
##  884
```

```
docvars(df) <- left_join(docvars(df), meta, by='doc')
docvars(df) <- mutate(docvars(df), lat = as.numeric(lat), lon = as.numeric(lon))
sfe <- df

sfe
```

```
## Corpus consisting of 1,712 documents and 21 docvars.
## 18thGerman-preKant.json :
## " In Germany, the eighteenth century was the age of enlighten..."
##
## abduction.json :
## " In the philosophical literature, the term "abduction" is us..."
##
## abelard.json :
## " Peter Abelard (1079-21 April 1142) [‘Abailard’ or ‘Abaelard..."
##
## abhidharma.json :
## " The first centuries after Śākyamuni Buddha’s death saw the ..."
##
## abilities.json :
## " In the accounts we give of one another, claims about our ab..."
##
## abner-burgos.json :
## " Abner of Burgos (Alfonso de Valladolid; c. 1260-1347) was p..."
##
## [ reached max_ndoc ... 1,706 more documents ]
```