Lab_Exercise#4_Octaviano

Jirraine Octaviano

2024-03-17

```
#install.packages("dplyr")
#install.packages("stringr")
#install.packages("httr")
#install.packages("rvest")

library(dplyr)
library(stringr)
library(httr)
library(rvest)

start <- proc.time()

## initializing empty vectors
title <- author <- subject <- abstract <- meta <- vector("character")

base_url <- 'https://arxiv.org/search/?query=programming+language&searchtype=all&abstracts=show&order=-a
## There are 50 articles per page / 3 Pages
pages <- seq(from = 0, to = 100, by = 50)

for(page in pages) {

  url <- paste0(base_url, page)

  article_urls <- read_html(url) %>%
    html_nodes('p.list-title.is-inline-block') %>%
    html_nodes('a[href^="https://arxiv.org/abs"]') %>%
    html_attr('href')


  # loop through all article urls in each page
  for(article_url in article_urls) {

    article_page <- read_html(article_url)

    ## TITLE
    scrapedTitle <- article_page %>% html_nodes('h1.title.mathjax') %>% html_text(TRUE)
    scrapedTitle <- gsub('Title:', '', scrapedTitle)
    title <- c(title, scrapedTitle)

    ## AUTHOR
    scrapedAuthor <- article_page %>% html_nodes('div.authors') %>% html_text(TRUE)
    scrapedAuthor <- gsub('Authors:','',scrapedAuthor)
```

```
    author <- c(author, scrapedAuthor)

    ## SUBJECT
    scrapedSubject <- article_page %>% html_nodes('span.primary-subject') %>% html_text(TRUE)
    subject <- c(subject, scrapedSubject)

    ## ABSTRACT
    scrapedAbstract <- article_page %>% html_nodes('blockquote.abstract.mathjax') %>% html_text(TRUE)
    scrapedAbstract <- sub('Abstract:','',scrapedAbstract)
    abstract <- c(abstract, scrapedAbstract)

    ## META
    scrapedMeta <- article_page %>% html_nodes('div.submission-history') %>% html_text(TRUE)
    scrapedMeta <- gsub('\\s+', ' ',scrapedMeta)
    scrapedMeta <- strsplit(scrapedMeta, '[v1]', fixed = T)
    scrapedMeta <- scrapedMeta[[1]][2] %>% unlist %>% str_trim
    meta <- c(meta, scrapedMeta)


    cat("Scraped article:", length(title), "\n")
    Sys.sleep(1)
  }
}

# merge all vectors to a data frame
papers <- data.frame(title, author, subject, abstract, meta)
#View(papers)

end <- proc.time()
end - start # Total Elapsed Time
```

//saved to csv and rdata

```
save(papers, file = "data/arxiv_ai.RData")
write.csv(papers, file = "data/arxiv_ai.csv")
```