lab_exercise#4_Esmalla

Janessa Marie Esmalla

2024-03-07

Scraping article data

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

start <- proc.time()


url <- 'https://arxiv.org/search/?query=%22mathematics%22&searchtype=all&source=header&start=0'

parse_url(url)

start <- proc.time()
title <- NULL
author <- NULL
subject <- NULL
abstract <- NULL
meta <- NULL

pages <- seq(from = 0, to = 100, by = 50)

for( i in pages){

  tmp_url <- modify_url(url, query = list(start = i))
  tmp_list <- read_html(tmp_url) %>%
    html_nodes('p.list-title.is-inline-block') %>%
    html_nodes('a[href^="https://arxiv.org/abs"]') %>%
    html_attr('href')

  for(j in 1:length(tmp_list)){

    tmp_paragraph <- read_html(tmp_list[j])

    # title
    tmp_title <- tmp_paragraph %>% html_nodes('h1.title.mathjax') %>% html_text(T)
    tmp_title <-  gsub('Title:', '', tmp_title)
    title <- c(title, tmp_title)

    # author
```

```
    tmp_author <- tmp_paragraph %>% html_nodes('div.authors') %>% html_text
    tmp_author <- gsub('\\s+',' ',tmp_author)
    tmp_author <- gsub('Authors:','',tmp_author) %>% str_trim
    author <- c(author, tmp_author)

    # subject
    tmp_subject <- tmp_paragraph %>% html_nodes('span.primary-subject') %>% html_text(T)
    subject <- c(subject, tmp_subject)

    # abstract
    tmp_abstract <- tmp_paragraph %>% html_nodes('blockquote.abstract.mathjax') %>% html_text(T)
    tmp_abstract <- gsub('\\s+',' ',tmp_abstract)
    tmp_abstract <- sub('Abstract:','',tmp_abstract) %>% str_trim
    abstract <- c(abstract, tmp_abstract)

    # meta
    tmp_meta <- tmp_paragraph %>% html_nodes('div.submission-history') %>% html_text
    tmp_meta <- lapply(strsplit(gsub('\\s+', ' ',tmp_meta), '[v1]', fixed = T),'[',2) %>% unlist %>% st
    meta <- c(meta, tmp_meta)
    cat(j, "paper\n")
    Sys.sleep(1)

  }
  cat((i/50) + 1,'/ 9 page\n')

}
papers <- data.frame(title, author, subject, abstract, meta)
end <- proc.time()
end - start # Total Elapsed Time

# Export the result
save(papers, file = "Arxiv_Mathematics.RData")
write.csv(papers, file = "Arxiv papers on Mathematics.csv")
```

==========
INSERT ALL REVIEWS DATA FRAME TO DATABASE ==========

USED RMYSQL

```
library(dplyr,dbplyr)
library(DBI)
library(RMariaDB)
library(odbc)

#creating connections
connection <- dbConnect(RMariaDB::MariaDB(),
                        dsn="MariaDB-connection",
                        Server = "localhost",
                        dbname = "esmalla_lab4",
                        user = "root",
```

```
                        password = "")


#install.packages("readr")
library(readr)

articles <- read.csv("Arxiv papers on Mathematics.csv")
tail(articles)
#dbWriteTable(connection,'lab4_articles', articles, append = TRUE)


dbListTables(connection)
dbListFields(connection,'lab4_articles')


review_data <- dbGetQuery(connection, "SELECT * FROM esmalla_lab4.lab4_articles")
glimpse(review_data)




dbDisconnect(connection)
```