

CS102__4th__Act

Cedric

2024-03-14

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

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

url <- 'https://arxiv.org/search/?query=drugs&searchtype=all&abstracts=show&order=-announced_date_first'

parse_url(url)

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

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

#Getting several pages
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])

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

```

# For AUTHORS
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)

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

# For ABSTRACTS
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)

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

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

}

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

# Export the result
save(Arxiv_papers, file = "drugs.RData")
write.csv(Arxiv_papers, file = "drugs.csv")

INSERTING DATA TO THE DATABASE

#install.packages("odbc")
#install.packages("RMySQL")
library(DBI)
library(odbc)
library(RMySQL)
library(dplyr, dbplyr)
connection <- dbConnect(RMySQL::MySQL(),
                        dsn="MariaDB-connection",
                        Server = "localhost",
                        dbname = "talon",
                        user = "root",
                        password = "")

```

Writing Table to Database

```
dbWriteTable(connection, 'arXivArticles', articles, append = TRUE)
```

LISTING TABLES AND FIELDS

```
dbListTables(connection)  
dbListFields(connection, 'arXivArticles')
```

READING DATA FROM TABLE

```
review_data <- dbGetQuery(connection, "SELECT * FROM .arXivArticles")  
glimpse(review_data)
```