title: “CS102_JUNTANILLA_LAB4” author: “JOSE ROLAND JUNTANILLA” date: “2024-03-14” output: html_document: default pdf_document: default

#Scraping informations in Arxiv

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

begin <- proc.time()

Title <- Author <- Subject <- Abstract <- Meta <- vector("character")

urllink <- 'https://arxiv.org/search/?query=%22computer+program%22&searchtype=all&source=header&start='
pages <- seq(from = 0, to = 100, by = 50)

for(page in pages) {

  pasteurl <- paste0(urllink, page)

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


  for(articlescrape in articlescrapes) {

    articlepage <- read_html(articlescrape)


    articletitle <- articlepage %>% html_nodes('h1.title.mathjax') %>% html_text(TRUE)
    articletitle <- gsub('Title:', '', articletitle)
    Title <- c(Title, articletitle)


    articleauthor <- articlepage %>% html_nodes('div.authors') %>% html_text(TRUE)
    articleauthor <- gsub('Authors:','',articleauthor)
    Author <- c(Author, articleauthor)


    articlesubject <- articlepage %>% html_nodes('span.primary-subject') %>% html_text(TRUE)
    Subject <- c(Subject, articlesubject)


    articleabstract <- articlepage %>% html_nodes('blockquote.abstract.mathjax') %>% html_text(TRUE)
    articleabstract <- sub('Abstract:','',articleabstract)
    Abstract <- c(Abstract, articleabstract)


    articlemeta <- articlepage %>% html_nodes('div.submission-history') %>% html_text(TRUE)
    articlemeta <- gsub('\\s+', ' ',articlemeta)
    articlemeta <- strsplit(articlemeta, '[v1]', fixed = T)
    articlemeta <- articlemeta[[1]][2] %>% unlist %>% str_trim
    Meta <- c(Meta, articlemeta)


   cat("Scraped article:", length(Title), "\n")
    Sys.sleep(2)
  }
}

finalpaper <- data.frame(Title, Author, Subject, Abstract, Meta)
View(finalpaper)

end <- proc.time()
end - begin
```

#Saving the data into a CSV and Rdata

```
save(finalpaper, file = "extractarxiv.RData")
```

```
## Error in save(finalpaper, file = "extractarxiv.RData"): object 'finalpaper' not found
```

```
write.csv(finalpaper, file = "extractarxiv.csv")
```

```
## Error in eval(expr, p): object 'finalpaper' not found
```

#Connecting in DATABASE

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

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
connection <- dbConnect(RMySQL::MySQL(),
                        dsn="MariaDB-connection",
                        Server = "localhost",
                        dbname = "bsit2c_juntanilla",
                        user = "root",
                        password = "")
```

```
## Error in .local(drv, ...): Failed to connect to database: Error: Can't connect to local MySQL server
```

```
library(readr)

arxivart <- read.csv("extractarxiv.csv")
```

```
tail(arxivart)
```

```
# Writing in database

dbWriteTable(connection, 'paperscrape', arxivart, append = TRUE)
```

```
## Error in h(simpleError(msg, call)): error in evaluating the argument 'conn' in selecting a method fo
```

```
# Creating Tables and Fields
dbListTables(connection)
```

```
## Error in h(simpleError(msg, call)): error in evaluating the argument 'conn' in selecting a method fo
```

```
dbListFields(connection,'paperscrape')
```

```
## Error in h(simpleError(msg, call)): error in evaluating the argument 'conn' in selecting a method fo
```

```
# Reading Data
arxiv_readata <- dbGetQuery(connection, "SELECT * FROM bsit2c_juntanilla.paperscrape")
```

```
## Error in h(simpleError(msg, call)): error in evaluating the argument 'conn' in selecting a method fo
```

```
glimpse(arxiv_readata)
```

```
## Error in eval(expr, envir, enclos): object 'arxiv_readata' not found
```