Lab Exercise 4

Ace Bagilidad

2024-03-13

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

library(dplyr)
```

```
##
## 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
```

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

url <- 'https://arxiv.org/search/?query=machine+learning&searchtype=all&abstracts=show&order=-announced_

parse_url(url)
```

```
## $scheme
## [1] "https"
##
## $hostname
## [1] "arxiv.org"
##
## $port
## NULL
##
## $path
## [1] "search/"
##
## $query
## $query$query
## [1] "machine+learning"
##
## $query$searchtype
## [1] "all"
```

```
##
## $query$abstracts
## [1] "show"
##
## $query$order
## [1] "-announced_date_first"
##
## $query$size
## [1] "50"
##
##
## $params
## NULL
##
## $fragment
## NULL
##
## $username
## NULL
##
## $password
## NULL
##
## attr(,"class")
## [1] "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 %>% st
    meta <- c(meta, tmp_meta)
    cat(j, "paper\n")
    Sys.sleep(1)

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

}
```

```
## 1 paper
## 2 paper
## 3 paper
## 4 paper
## 5 paper
## 6 paper
## 7 paper
## 8 paper
## 9 paper
## 10 paper
## 11 paper
## 12 paper
## 13 paper
## 14 paper
## 15 paper
## 16 paper
## 17 paper
## 18 paper
## 19 paper
## 20 paper
## 21 paper
## 22 paper
## 23 paper
## 24 paper
## 25 paper
## 26 paper
```

```
## 27 paper
## 28 paper
## 29 paper
## 30 paper
## 31 paper
## 32 paper
## 33 paper
## 34 paper
## 35 paper
## 36 paper
## 37 paper
## 38 paper
## 39 paper
## 40 paper
## 41 paper
## 42 paper
## 43 paper
## 44 paper
## 45 paper
## 46 paper
## 47 paper
## 48 paper
## 49 paper
## 50 paper
## 1 / 9 page
## 1 paper
## 2 paper
## 3 paper
## 4 paper
## 5 paper
## 6 paper
## 7 paper
## 8 paper
## 9 paper
## 10 paper
## 11 paper
## 12 paper
## 13 paper
## 14 paper
## 15 paper
## 16 paper
## 17 paper
## 18 paper
## 19 paper
## 20 paper
## 21 paper
## 22 paper
## 23 paper
## 24 paper
## 25 paper
## 26 paper
## 27 paper
## 28 paper
## 29 paper
```

```
## 30 paper
## 31 paper
## 32 paper
## 33 paper
## 34 paper
## 35 paper
## 36 paper
## 37 paper
## 38 paper
## 39 paper
## 40 paper
## 41 paper
## 42 paper
## 43 paper
## 44 paper
## 45 paper
## 46 paper
## 47 paper
## 48 paper
## 49 paper
## 50 paper
## 2 / 9 page
## 1 paper
## 2 paper
## 3 paper
## 4 paper
## 5 paper
## 6 paper
## 7 paper
## 8 paper
## 9 paper
## 10 paper
## 11 paper
## 12 paper
## 13 paper
## 14 paper
## 15 paper
## 16 paper
## 17 paper
## 18 paper
## 19 paper
## 20 paper
## 21 paper
## 22 paper
## 23 paper
## 24 paper
## 25 paper
## 26 paper
## 27 paper
## 28 paper
## 29 paper
## 30 paper
## 31 paper
## 32 paper
```

```
## 33 paper
## 34 paper
## 35 paper
## 36 paper
## 37 paper
## 38 paper
## 39 paper
## 40 paper
## 41 paper
## 42 paper
## 43 paper
## 44 paper
## 45 paper
## 46 paper
## 47 paper
## 48 paper
## 49 paper
## 50 paper
## 3 / 9 page
```

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

```
##    user  system elapsed
##   3.506   0.173 157.950
```

```
# Export the result
save(Arxiv_papers, file = "Arxiv_Machine_Learning.RData")
write.csv(Arxiv_papers, file = "Machine Learning.csv")
```

INSERTING DATA TO THE DATABASE

```
install.packages("dbplyr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
library(DBI)
library(odbc)
library(RMySQL)
library(dplyr,dbplyr)
# connection <- dbConnect(RMySQL::MySQL(),
#                         dsn="MariaDB-connection",
#                         Server = "localhost",
#                         dbname = "datascience_bagilidad2c",
#                         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 datascience_bagilidad2c.arXivArticles")
# glimpse(review_data)
```