RWorksheet#5_Soteo-Group.Rmd

2023-12-22

```
install.packages("rvest")
```

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

```
library(rvest)
install.packages("polite")
```

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

```
library(polite)
install.packages("dplyr")
```

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

```
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
```

```
install.packages("httr")
```

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

```
library(httr)
install.packages("kableExtra")
```

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

```
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
polite::use_manners(save_as = 'polite_scrape.R')
```

```
## v Setting active project to '/cloud/project'
```

```
urlLinks <-'https://www.imdb.com/chart/top/?ref_=nv_mv_250&sort=rank%2Casc'
session <- bow(urlLinks,
               user_agent = "Educational")
session
```

```
## <polite session> https://www.imdb.com/chart/top/?ref_=nv_mv_250&sort=rank%2Casc
##     User-agent: Educational
##     robots.txt: 34 rules are defined for 2 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

```
rank_title <- character(0)
links <- character(0)

title_list <- scrape(session) %>%
  html_nodes('h3.ipc-title__text') %>%
  html_text

class(title_list)
```

```
## [1] "character"
```

```
title_list_sub <- as.data.frame(title_list[2:251])
head(title_list_sub)
```

```
##             title_list[2:251]
## 1 1. The Shawshank Redemption
## 2            2. The Godfather
## 3          3. The Dark Knight
## 4    4. The Godfather Part II
## 5             5. 12 Angry Men
## 6         6. Schindler's List
```

```
tail(title_list_sub)
```

```
##           title_list[2:251]
## 245    245. Pather Panchali
## 246      246. The 400 Blows
## 247             247. Aladdin
## 248 248. Dances with Wolves
## 249      249. Life of Brian
## 250           250. Drishyam
```

```
colnames(title_list_sub) <- "ranks"
colnames(title_list_sub) <- "ranks"
split_df <- strsplit(as.character(title_list_sub$ranks),".",fixed = TRUE)
split_df <- data.frame(do.call(rbind,split_df))
```

```
## Warning in (function (..., deparse.level = 1) : number of columns of result is
## not a multiple of vector length (arg 72)
```

```
split_df <- split_df[-c(3:4)]
colnames(split_df) <- c("ranks","title")
str(split_df)
```

```
## 'data.frame':    250 obs. of  2 variables:
##  $ ranks: chr  "1" "2" "3" "4" ...
##  $ title: chr  " The Shawshank Redemption" " The Godfather" " The Dark Knight" " The Godfather Part I
```

```
class(split_df)
```

```
## [1] "data.frame"
```

```
head(split_df)
```

```
##   ranks                     title
## 1     1  The Shawshank Redemption
## 2     2             The Godfather
## 3     3           The Dark Knight
## 4     4     The Godfather Part II
## 5     5              12 Angry Men
## 6     6          Schindler's List
```

```
rank_title <- data.frame(
  rank_title = split_df)

write.csv(rank_title,file = "title.csv")
```

```
link_list <- scrape(session) %>%
  html_nodes('a.ipc-title-link-wrapper') %>%
  html_attr('href')

head(link_list)
```

```
## [1] "/title/tt0111161/?ref_=chttp_t_1" "/title/tt0068646/?ref_=chttp_t_2"
## [3] "/title/tt0468569/?ref_=chttp_t_3" "/title/tt0071562/?ref_=chttp_t_4"
## [5] "/title/tt0050083/?ref_=chttp_t_5" "/title/tt0108052/?ref_=chttp_t_6"
```

```
link_list[245:257]
```

```
##  [1] "/title/tt0048473/?ref_=chttp_t_245"
##  [2] "/title/tt0053198/?ref_=chttp_t_246"
##  [3] "/title/tt0103639/?ref_=chttp_t_247"
##  [4] "/title/tt0099348/?ref_=chttp_t_248"
##  [5] "/title/tt0079470/?ref_=chttp_t_249"
##  [6] "/title/tt4430212/?ref_=chttp_t_250"
##  [7] "/chart/boxoffice/?ref_=chttp_ql_1"
##  [8] "/chart/moviemeter/?ref_=chttp_ql_2"
##  [9] "/chart/top-english-movies/?ref_=chttp_ql_4"
## [10] "/chart/tvmeter/?ref_=chttp_ql_5"
## [11] "/chart/toptv/?ref_=chttp_ql_6"
## [12] "/chart/bottom/?ref_=chttp_ql_7"
## [13] "/chart/starmeter/?ref_=chttp_ql_8"
```

```
link <- as.vector(link_list[1:250])
names(link) <- "links"

head(link)
```

```
##                               links                                    <NA>
## "/title/tt0111161/?ref_=chttp_t_1" "/title/tt0068646/?ref_=chttp_t_2"
##                                <NA>                                    <NA>
## "/title/tt0468569/?ref_=chttp_t_3" "/title/tt0071562/?ref_=chttp_t_4"
```

```
##                                <NA>                                <NA>
## "/title/tt0050083/?ref_=chttp_t_5" "/title/tt0108052/?ref_=chttp_t_6"
```

```
tail(link)
```

```
##                                  <NA>                                  <NA>
## "/title/tt0048473/?ref_=chttp_t_245" "/title/tt0053198/?ref_=chttp_t_246"
##                                  <NA>                                  <NA>
## "/title/tt0103639/?ref_=chttp_t_247" "/title/tt0099348/?ref_=chttp_t_248"
##                                  <NA>                                  <NA>
## "/title/tt0079470/?ref_=chttp_t_249" "/title/tt4430212/?ref_=chttp_t_250"
```

```
for (i in 1:250) {
  link[i] <- paste0("https://imdb.com", link[i], sep = "")
}

links <- as.data.frame(link)

rank_title <- data.frame(
  rank_title = split_df, link)
scrape_df <- data.frame(rank_title,links)
names(scrape_df) <- c("Rank","Title","Link")

head(scrape_df)
```

```
##   Rank                     Title
## 1    1  The Shawshank Redemption
## 2    2             The Godfather
## 3    3           The Dark Knight
## 4    4     The Godfather Part II
## 5    5              12 Angry Men
## 6    6          Schindler's List
##                                                 Link
## 1 https://imdb.com/title/tt0111161/?ref_=chttp_t_1
## 2 https://imdb.com/title/tt0068646/?ref_=chttp_t_2
## 3 https://imdb.com/title/tt0468569/?ref_=chttp_t_3
## 4 https://imdb.com/title/tt0071562/?ref_=chttp_t_4
## 5 https://imdb.com/title/tt0050083/?ref_=chttp_t_5
## 6 https://imdb.com/title/tt0108052/?ref_=chttp_t_6
##                                                   NA
## 1 https://imdb.com/title/tt0111161/?ref_=chttp_t_1
## 2 https://imdb.com/title/tt0068646/?ref_=chttp_t_2
## 3 https://imdb.com/title/tt0468569/?ref_=chttp_t_3
## 4 https://imdb.com/title/tt0071562/?ref_=chttp_t_4
## 5 https://imdb.com/title/tt0050083/?ref_=chttp_t_5
## 6 https://imdb.com/title/tt0108052/?ref_=chttp_t_6
```

```
imdb_top_50 <- data.frame()

current_row <- 1

for (row in 1:2) {
  url <- links$link[current_row]

  session2 <- bow(url,
                  user_agent = "Educational")
```

```
  webpage <- scrape(session2)

  rating <- html_text(html_nodes(webpage, ".sc-bde20123-1.cMEQkK"))
  rating <- rating[-2]

  votecount <- html_text(html_nodes(webpage,
                                    'div.sc-bde20123-3.gPVQxL'))
  votecount <- votecount[-2]

  movie_desc <- html_text(html_nodes(webpage,
                                     '.sc-466bb6c-1.dWufeH'))
  movie_desc <- movie_desc[-2]

  meta_score <- html_text(html_nodes(
    webpage,
    '.sc-b0901df4-0.bcQdDJ.metacritic-score-box'))
  meta_score <- meta_score[-2]

  cat("Rating for", url, "is:", rating, "vote count is", votecount, 'and metascore is',meta_score, "\n")

  imdb_top_50[current_row,1] <- rating
  imdb_top_50[current_row,2] <- votecount
  imdb_top_50[current_row,3] <- movie_desc
  imdb_top_50[current_row,4] <- meta_score

  current_row <- current_row + 1

  Sys.sleep(3)
}
```

```
## Rating for https://imdb.com/title/tt0111161/?ref_=chttp_t_1 is: 9.3 vote count is 2.8M and metascore
## Rating for https://imdb.com/title/tt0068646/?ref_=chttp_t_2 is: 9.2 vote count is 2M and metascore is
```

```
names(imdb_top_50) <- c("Rating","VoteCount","Description","MetaScore")

write.csv(imdb_top_50,file = "title.csv")
imdb_top_250 <- data.frame(
  scrape_df,imdb_top_50)
```

```
## Warning in data.frame(scrape_df, imdb_top_50): row names were found from a
## short variable and have been discarded
```

```
write.csv(imdb_top_250,file = "title.csv")

library(kableExtra)

df_d <- imdb_top_250[c(1:2),]

knitr::kable(df_d, caption = "IMDB Top 250 Movies") %>%
  kable_classic(full_width = T, html_font = "Arial Narrow") %>%
  kable_styling(font_size = 9)
```

Table 1: IMDB Top 250 Movies

Rank	Title	Link	NA.	Rating	VoteCount	Description	MetaScore
1	The Shawshank Redemption	https://imdb.com/title/tt0111161/?ref_=chttp_t_1	https://imdb.com/title/tt0111161/?ref_=chttp_t_1	9.3	2.8M	Over the course of several years, two convicts form a friendship, seeking consolation and, eventually, redemption through basic compassion.	82
2	The Godfather	https://imdb.com/title/tt0068646/?ref_=chttp_t_2	https://imdb.com/title/tt0068646/?ref_=chttp_t_2	9.2	2.0M	Don Vito Corleone, head of a mafia family, decides to hand over his empire to his youngest son Michael. However, his decision unintentionally puts the lives of his loved ones in grave danger.	100