RWorksheet#5_group(Lomibao,rabago and andigan)

2024-11-18

```
library(kableExtra)
library("rvest")
library("polite")
library("dplyr")
```

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

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

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

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

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

```
## v Setting active project to "/cloud/project".
```

```
url <- "https://www.imdb.com/chart/toptv/"
webpage <- read_html(url)
 session <- bow(url,
                 user_agent = "Student education purpose")
 session
```

```
## <polite session> https://www.imdb.com/chart/toptv/
##     User-agent: Student education purpose
##     robots.txt: 35 rules are defined for 3 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

```
 page <- scrape(session)
```

scraping the title

```
 title <- page%>%html_nodes('h3.ipc-title__text')%>%html_text()
 title <- title[2:26]
 title
```

```
##  [1] "1. Breaking Bad"
##  [2] "2. Planet Earth II"
##  [3] "3. Planet Earth"
##  [4] "4. Band of Brothers"
##  [5] "5. Chernobyl"
##  [6] "6. The Wire"
##  [7] "7. Avatar: The Last Airbender"
##  [8] "8. Blue Planet II"
##  [9] "9. The Sopranos"
## [10] "10. Cosmos: A Spacetime Odyssey"
## [11] "11. Cosmos"
## [12] "12. Our Planet"
## [13] "13. Game of Thrones"
## [14] "14. Bluey"
## [15] "15. The World at War"
## [16] "16. Fullmetal Alchemist: Brotherhood"
## [17] "17. Rick and Morty"
## [18] "18. Life"
## [19] "19. The Last Dance"
## [20] "20. The Twilight Zone"
## [21] "21. The Vietnam War"
## [22] "22. Sherlock"
## [23] "23. Attack on Titan"
## [24] "24. Batman: The Animated Series"
## [25] "25. Arcane"
```

scraping the rating

```r
ratings<- page %>%
  html_nodes('span.ipc-rating-star--rating') %>%
  html_text()
ratings
```

```
##  [1] "9.5" "9.5" "9.4" "9.4" "9.3" "9.3" "9.3" "9.3" "9.2" "9.2" "9.3" "9.2"
## [13] "9.2" "9.3" "9.2" "9.1" "9.1" "9.1" "9.0" "9.0" "9.1" "9.1" "9.1" "9.0"
## [25] "9.0"
```

scraping the numbers of vote

```r
number_votes <- page %>%
  html_nodes("span.ipc-rating-star--voteCount") %>%
  html_text()
number_votes
```

```
##  [1] " (2.2M)" " (162K)" " (224K)" " (546K)" " (908K)" " (391K)" " (390K)"
##  [8] " (49K)"  " (499K)" " (131K)" " (46K)"  " (54K)"  " (2.4M)" " (34K)"
## [15] " (31K)"  " (209K)" " (628K)" " (44K)"  " (160K)" " (97K)"  " (29K)"
## [22] " (1M)"   " (563K)" " (122K)" " (318K)"
```

scraping the number of episode

```r
num_ep <- page %>%
  html_nodes('span.sc-6-ade9358-7.exckou.cli-title-metadata-item')%>%
  html_text()
num_ep
```

```
## character(0)
```

Cleaning the episode data

```r
# episode <- str_extract(num_ep, "\\d+ eps")
#  episodes <- str_remove(episode, " eps")
# episodes <- as.numeric(episodes)
# episodes
```

scraping the year release

```
year <- page %>%
  html_nodes("span.sc-5bc66c50-6.OOdsw.cli-title-metadata-item") %>%
  html_text()
year
```

```
## character(0)
```

Extract using the regex

```
 #release_years <- str_extract(year, "\\d{4}")
 #release_years <- release_years[!is.na(release_years)]
 #release_years <- as.numeric(release_years)
```

checking the length.

```
#cat("Show Titles length: ", length(title), "\n")
#cat("Show Ratings length: ", length(ratings), "\n")
#cat("Number of Votes length: ", length(number_votes), "\n")
#cat("Episode Counts length: ", length(episodes), "\n")
#cat("Release Years length: ", length(release_years), "\n")
```

```
title_list <- as.data.frame(title[1:50])
colnames(title_list)<-"ranks"
```

spliting the data frame

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

```
##      X1                                X2
## 1     1                      Breaking Bad
## 2     2                   Planet Earth II
## 3     3                      Planet Earth
## 4     4                  Band of Brothers
## 5     5                         Chernobyl
## 6     6                          The Wire
## 7     7        Avatar: The Last Airbender
## 8     8                    Blue Planet II
## 9     9                      The Sopranos
## 10   10       Cosmos: A Spacetime Odyssey
## 11   11                            Cosmos
## 12   12                        Our Planet
## 13   13                   Game of Thrones
## 14   14                             Bluey
## 15   15                  The World at War
## 16   16  Fullmetal Alchemist: Brotherhood
## 17   17                    Rick and Morty
## 18   18                              Life
## 19   19                    The Last Dance
## 20   20                 The Twilight Zone
## 21   21                   The Vietnam War
## 22   22                          Sherlock
## 23   23                   Attack on Titan
## 24   24       Batman: The Animated Series
## 25   25                            Arcane
## 26 <NA>                              <NA>
## 27 <NA>                              <NA>
## 28 <NA>                              <NA>
## 29 <NA>                              <NA>
## 30 <NA>                              <NA>
## 31 <NA>                              <NA>
## 32 <NA>                              <NA>
## 33 <NA>                              <NA>
## 34 <NA>                              <NA>
## 35 <NA>                              <NA>
## 36 <NA>                              <NA>
## 37 <NA>                              <NA>
## 38 <NA>                              <NA>
## 39 <NA>                              <NA>
## 40 <NA>                              <NA>
## 41 <NA>                              <NA>
## 42 <NA>                              <NA>
## 43 <NA>                              <NA>
## 44 <NA>                              <NA>
## 45 <NA>                              <NA>
## 46 <NA>                              <NA>
## 47 <NA>                              <NA>
## 48 <NA>                              <NA>
## 49 <NA>                              <NA>
## 50 <NA>                              <NA>
```

renaming columns

```
split_df<-split_df[-c(3,4)]
colnames(split_df)<- c("Ranks","Titles")
split_df
```

```
##    Ranks                            Titles
## 1      1                      Breaking Bad
## 2      2                   Planet Earth II
## 3      3                      Planet Earth
## 4      4                  Band of Brothers
## 5      5                         Chernobyl
## 6      6                          The Wire
## 7      7        Avatar: The Last Airbender
## 8      8                    Blue Planet II
## 9      9                      The Sopranos
## 10    10       Cosmos: A Spacetime Odyssey
## 11    11                            Cosmos
## 12    12                        Our Planet
## 13    13                   Game of Thrones
## 14    14                             Bluey
## 15    15                  The World at War
## 16    16  Fullmetal Alchemist: Brotherhood
## 17    17                    Rick and Morty
## 18    18                              Life
## 19    19                    The Last Dance
## 20    20                 The Twilight Zone
## 21    21                   The Vietnam War
## 22    22                          Sherlock
## 23    23                   Attack on Titan
## 24    24       Batman: The Animated Series
## 25    25                            Arcane
## 26  <NA>                              <NA>
## 27  <NA>                              <NA>
## 28  <NA>                              <NA>
## 29  <NA>                              <NA>
## 30  <NA>                              <NA>
## 31  <NA>                              <NA>
## 32  <NA>                              <NA>
## 33  <NA>                              <NA>
## 34  <NA>                              <NA>
## 35  <NA>                              <NA>
## 36  <NA>                              <NA>
## 37  <NA>                              <NA>
## 38  <NA>                              <NA>
## 39  <NA>                              <NA>
## 40  <NA>                              <NA>
## 41  <NA>                              <NA>
## 42  <NA>                              <NA>
## 43  <NA>                              <NA>
## 44  <NA>                              <NA>
## 45  <NA>                              <NA>
## 46  <NA>                              <NA>
## 47  <NA>                              <NA>
## 48  <NA>                              <NA>
## 49  <NA>                              <NA>
## 50  <NA>                              <NA>
```

creating csv for title and ranks

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

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

Combining them all to a data frame.

```
 # imdb_top_tv_shows <- data.frame(
 # Title = title,
 # Rating = ratings,
 # Votes = number_votes,
 # Episode = episodes,
 # Release_Year = release_years,
  #stringsAsFactors = FALSE
 # )
```

R scraping

```
library('rvest')
library('polite')

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

urlr <- "https://www.amazon.com/?&tag=phtxtabkgode-20&ref=pd_sl_73t48p1dlf_e&adgrpid=151590336221&hvpon
```

```
amazon <- read_html(urlr)
session2 <- bow(urlr,
               user_agent = "Student's Demo Educational")
session2
```

```
## <polite session> https://www.amazon.com/?&tag=phtxtabkgode-20&ref=pd_sl_73t48p1dlf_e&adgrpid=1515903
##     User-agent: Student's Demo Educational
##     robots.txt: 138 rules are defined for 5 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

```
page2 <- scrape(session2)
num_products = 31
```

Creating a data frame for storing the data.

```
data <- data.frame()
```

loop for link

url for categories

```
shirt_cat<- "https://www.amazon.com/s?k=shirt&i=fashion-mens-intl-ship&crid=6IQRNOUUJ0LB&sprefix=shirt%
pants_cat <- "https://www.amazon.com/s?k=pants&i=fashion-mens-intl-ship&crid=9U0VNEZTF2CR&sprefix=pants
shoe_cat<- "https://www.amazon.com/s?k=shoes&i=fashion-mens-intl-ship&crid=ADB2HOWLHCPK&sprefix=sho%2Cf
head_phone <-"https://www.amazon.com/s?k=headphone&i=fashion-mens-intl-ship&crid=25P9FL9QS4YNZ&sprefix=
medkit_cat<-"https://www.amazon.com/s?k=medkit&i=fashion-mens-intl-ship&crid=1HF7OZ2EVLHQY&sprefix=medk
```