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/"

 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
```

```
##  [1] "IMDb Charts"
##  [2] "1. Breaking Bad"
##  [3] "2. Planet Earth II"
##  [4] "3. Planet Earth"
##  [5] "4. Band of Brothers"
##  [6] "5. Chernobyl"
##  [7] "6. The Wire"
```

```
##  [8] "7. Avatar: The Last Airbender"
##  [9] "8. Blue Planet II"
## [10] "9. The Sopranos"
## [11] "10. Cosmos: A Spacetime Odyssey"
## [12] "11. Cosmos"
## [13] "12. Our Planet"
## [14] "13. Game of Thrones"
## [15] "14. Bluey"
## [16] "15. The World at War"
## [17] "16. Fullmetal Alchemist: Brotherhood"
## [18] "17. Rick and Morty"
## [19] "18. Life"
## [20] "19. The Last Dance"
## [21] "20. The Twilight Zone"
## [22] "21. The Vietnam War"
## [23] "22. Sherlock"
## [24] "23. Attack on Titan"
## [25] "24. Batman: The Animated Series"
## [26] "25. Arcane"
## [27] "Recently viewed"
```

scraping the rating

```
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

```
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)" " (33K)"
## [15] " (31K)"  " (209K)" " (627K)" " (44K)"  " (160K)" " (97K)"  " (29K)"
## [22] " (1M)"   " (562K)" " (122K)" " (308K)"
```

scraping the number of episode

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

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

Cleaning the episode data

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

scraping the year

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

```
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      IMDb Charts                     IMDb Charts
## 2                1                    Breaking Bad
## 3                2                 Planet Earth II
## 4                3                    Planet Earth
## 5                4                Band of Brothers
## 6                5                       Chernobyl
## 7                6                       The Wire
## 8                7       Avatar: The Last Airbender
## 9                8                  Blue Planet II
## 10               9                    The Sopranos
## 11              10     Cosmos: A Spacetime Odyssey
## 12              11                          Cosmos
## 13              12                      Our Planet
## 14              13                Game of Thrones
## 15              14                           Bluey
## 16              15               The World at War
## 17              16 Fullmetal Alchemist: Brotherhood
## 18              17                 Rick and Morty
## 19              18                           Life
## 20              19                The Last Dance
## 21              20              The Twilight Zone
## 22              21                The Vietnam War
## 23              22                       Sherlock
## 24              23                Attack on Titan
## 25              24     Batman: The Animated Series
## 26              25                         Arcane
## 27 Recently viewed                Recently viewed
## 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      IMDb Charts                     IMDb Charts
## 2                1                   Breaking Bad
## 3                2                 Planet Earth II
## 4                3                    Planet Earth
## 5                4                Band of Brothers
## 6                5                       Chernobyl
## 7                6                        The Wire
## 8                7      Avatar: The Last Airbender
## 9                8                 Blue Planet II
## 10               9                   The Sopranos
## 11              10     Cosmos: A Spacetime Odyssey
## 12              11                          Cosmos
## 13              12                      Our Planet
## 14              13                Game of Thrones
## 15              14                           Bluey
## 16              15               The World at War
## 17              16 Fullmetal Alchemist: Brotherhood
## 18              17                 Rick and Morty
## 19              18                           Life
## 20              19                The Last Dance
## 21              20              The Twilight Zone
## 22              21                The Vietnam War
## 23              22                       Sherlock
## 24              23                Attack on Titan
## 25              24     Batman: The Animated Series
## 26              25                          Arcane
## 27 Recently viewed                Recently viewed
## 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

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

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