RWorksheet_Cacho_Perez_Sorenio_Tupaz

2024-11-11

Extracting TV Shows Reviews

1. Each group needs to extract the top 50 tv shows in Imdb.com. It will include the rank, the title of the tv show, tv rating, the number of people who voted, the number of episodes, the year it was released. It will also include the number of user reviews and the number of critic reviews, as well as the popularity rating for each tv shows.
2. From the 50 tv shows, select at least 5 tv shows to scrape 20 user reviews that will include the reviewer's name, date of reviewed, user rating, title of the review, the numbers for "is helpful" and "is not helpful", and text reviews.
3. Create a time series graph for the tv shows released by year. Which year has the most number of tv shows released?

```
# install.packages("rvest")
# install.packages("httr")
# install.packages("polite")

library(rvest)
```

```
## Warning: package 'rvest' was built under R version 4.4.2
```

```
library(httr)
```

```
## Warning: package 'httr' was built under R version 4.4.2
```

```
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(polite)
```

```
## Warning: package 'polite' was built under R version 4.4.2
```

```
#install.packages("kableExtra")
#library(kableExtra)
#library(rmarkdown)
```

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

```
## v Setting active project to "D:/Files".
```

```
url <- 'https://www.imdb.com/chart/toptv/?ref_=nv_tvv_250&sort=rank%2Casc'
session <- bow(url,
               user_agent = "Educational")
session
```

```
## <polite session> https://www.imdb.com/chart/toptv/?ref_=nv_tvv_250&sort=rank%2Casc
##     User-agent: Educational
##     robots.txt: 35 rules are defined for 3 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_listSub <- as.data.frame(title_list[2:26])
head(title_listSub)
```

```
##      title_list[2:26]
## 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
```

```
tail(title_listSub)
```

```
##                    title_list[2:26]
## 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. The Office
```

```r
colnames(title_listSub) <- "ranks"

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

```r
split_title <- split_title[-c(3:4)]
```

```r
colnames(split_title) <- c("Rank","Title")
```

```r
str(split_title)
```

```
## 'data.frame':    25 obs. of  2 variables:
##  $ Rank : chr  "1" "2" "3" "4" ...
##  $ Title: chr  " Breaking Bad" " Planet Earth II" " Planet Earth" " Band of Brothers" ...
```

```r
head(split_title)
```

```
##   Rank             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
```

```r
rank_title <- data.frame(rank_title = split_title)
```

```r
write.csv(rank_title,file = "D:/Files/title.csv")
```

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

```r
linkList <- linkList[!is.na(linkList) & linkList != ""]
```

```r
for (i in 1:length(linkList)) {
  linkList[i] <- paste0("https://imdb.com", linkList[i])
}
```

```r
links <- as.data.frame(linkList)
names(links) <- "link"

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

```r
scrape_rankTitle <- data.frame(rank_title)
names(scrape_rankTitle) <- c("Rank", "Title")

head(scrape_rankTitle)
```

```
##   Rank             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
```

```
write.csv(scrape_rankTitle,file = "D:/Files/top250.csv")
```

```
current_row <- 1
imdbTop25 <- data.frame()

for (row in 1:25) {
  url <- links$link[current_row]
}

  if (url == "" || is.na(url)) {
    next
  }
```

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

```
ratings <- html_text(html_nodes(webpage, ".sc-d541859f-1.imUuxf"))
  if (length(ratings) == 0) {
    next
  } else {
    ratings <- ratings[1]
  }
```

```
votes <- html_text(html_nodes(webpage, 'div.sc-d541859f-3.dwhNqC'))
  if (length(votes) == 0) {
    next
  } else {
    votes <- votes[1]
  }
```

```
episodesperYear <- html_text(html_nodes(webpage, xpath = "//span[contains(text(), 'episodes')]"))
  if (length(episodesperYear) == 0) {
    next
  } else {
    episodesperYear <- episodesperYear[1]
    episodes <- sub("episodes.*", "episodes", episodesperYear)
    yearReleased <- sub(".*episodes • ", "", episodesperYear)
  }
```

```
cat("Rating for", url, "is:", ratings, "vote count is", votes, "number of episodes is", episodes, "year
```

```
## Rating for https://imdb.com/title/tt0903747/?ref_=chttvtp_t_1 is: 9.5 vote count is 2.2M number of ep
```

```
imdbTop25[current_row, 1] <- ratings
  imdbTop25[current_row, 2] <- votes
  imdbTop25[current_row, 3] <- episodes
  imdbTop25[current_row, 4] <- yearReleased
```

```
current_row <- current_row + 1
```

```
names(imdbTop25) <- c("Rating","Votes","Number of Episodes", "Year Released")
```

```
write.csv(imdbTop25,file = "D:/Files/imdbTop25.csv")
```

```
imdbTop25 <- data.frame(scrape_rankTitle, imdbTop25)
```

```
## Warning in data.frame(scrape_rankTitle, imdbTop25): row names were found from a
## short variable and have been discarded
```

```
write.csv(imdbTop25,file = "D:/Files/imdbTop250.csv")
library(kableExtra)
```

```
## Warning: package 'kableExtra' was built under R version 4.4.2
```

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

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

```
knitr::kable(imdbTop25,caption = "Extracting Rating, Votes, Number of Episodes, Year Released") %>%
  kable_classic(full_width = FALSE, html_font = "Times New Roman") %>%
  kable_styling(font_size = 13)
```

Table 1: Extracting Rating, Votes, Number of Episodes, Year Released

Rank	Title	Rating	Votes	Number.of.Episodes	Year.Released
1	Breaking Bad	9.5	2.2M	62 episodes	2008–2013
2	Planet Earth II	9.5	2.2M	62 episodes	2008–2013
3	Planet Earth	9.5	2.2M	62 episodes	2008–2013
4	Band of Brothers	9.5	2.2M	62 episodes	2008–2013
5	Chernobyl	9.5	2.2M	62 episodes	2008–2013
6	The Wire	9.5	2.2M	62 episodes	2008–2013
7	Avatar: The Last Airbender	9.5	2.2M	62 episodes	2008–2013
8	Blue Planet II	9.5	2.2M	62 episodes	2008–2013

9	The Sopranos	9.5	2.2M	62 episodes	2008–2013
10	Cosmos: A Spacetime Odyssey	9.5	2.2M	62 episodes	2008–2013
11	Cosmos	9.5	2.2M	62 episodes	2008–2013
12	Our Planet	9.5	2.2M	62 episodes	2008–2013
13	Game of Thrones	9.5	2.2M	62 episodes	2008–2013
14	Bluey	9.5	2.2M	62 episodes	2008–2013
15	The World at War	9.5	2.2M	62 episodes	2008–2013
16	Fullmetal Alchemist Brotherhood	9.5	2.2M	62 episodes	2008–2013
17	Rick and Morty	9.5	2.2M	62 episodes	2008–2013
18	Life	9.5	2.2M	62 episodes	2008–2013
19	The Last Dance	9.5	2.2M	62 episodes	2008–2013
20	The Twilight Zone	9.5	2.2M	62 episodes	2008–2013
21	The Vietnam War	9.5	2.2M	62 episodes	2008–2013
22	Sherlock	9.5	2.2M	62 episodes	2008–2013
23	Attack on Titan	9.5	2.2M	62 episodes	2008–2013
24	Batman: The Animated Series	9.5	2.2M	62 episodes	2008–2013
25	The Office	9.5	2.2M	62 episodes	2008–2013

```
library(kableExtra)
```

```
top25 <- imdbTop25[c(1:25),]

top25 <- top25 %>%
  select_if(~ !all(is.na(.)))

knitr::kable(top25, caption = "IMDB Top 25 Shows") %>%
  kable_classic(full_width = FALSE, html_font = "Times New Roman") %>%
  kable_styling(font_size = 13)
```

Table 2: IMDB Top 25 Shows

Rank	Title	Rating	Votes	Number.of.Episodes	Year.Released
1	Breaking Bad	9.5	2.2M	62 episodes	2008–2013
2	Planet Earth II	9.5	2.2M	62 episodes	2008–2013
3	Planet Earth	9.5	2.2M	62 episodes	2008–2013
4	Band of Brothers	9.5	2.2M	62 episodes	2008–2013
5	Chernobyl	9.5	2.2M	62 episodes	2008–2013
6	The Wire	9.5	2.2M	62 episodes	2008–2013
7	Avatar: The Last Airbender	9.5	2.2M	62 episodes	2008–2013
8	Blue Planet II	9.5	2.2M	62 episodes	2008–2013
9	The Sopranos	9.5	2.2M	62 episodes	2008–2013
10	Cosmos: A Spacetime Odyssey	9.5	2.2M	62 episodes	2008–2013

11	Cosmos	9.5	2.2M	62 episodes	2008–2013
12	Our Planet	9.5	2.2M	62 episodes	2008–2013
13	Game of Thrones	9.5	2.2M	62 episodes	2008–2013
14	Bluey	9.5	2.2M	62 episodes	2008–2013
15	The World at War	9.5	2.2M	62 episodes	2008–2013
16	Fullmetal Alchemist Brotherhood	9.5	2.2M	62 episodes	2008–2013
17	Rick and Morty	9.5	2.2M	62 episodes	2008–2013
18	Life	9.5	2.2M	62 episodes	2008–2013
19	The Last Dance	9.5	2.2M	62 episodes	2008–2013
20	The Twilight Zone	9.5	2.2M	62 episodes	2008–2013
21	The Vietnam War	9.5	2.2M	62 episodes	2008–2013
22	Sherlock	9.5	2.2M	62 episodes	2008–2013
23	Attack on Titan	9.5	2.2M	62 episodes	2008–2013
24	Batman: The Animated Series	9.5	2.2M	62 episodes	2008–2013
25	The Office	9.5	2.2M	62 episodes	2008–2013