RWorksheet#5_group.R,

Noblezada, Camasa, Cabia

2024-11-06

```
library(polite)
library(kableExtra)
library(rmarkdown)

url <- 'https://www.amazon.com/ref=nav_logo'

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

```
## <polite session> https://www.amazon.com/ref=nav_logo
##     User-agent: Educational
##     robots.txt: 137 rules are defined for 4 bots
##    Crawl delay: 5 sec
##   The path is scrapable for this user-agent
```

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

```
## No encoding supplied: defaulting to UTF-8.
```

```
library(polite)
library(kableExtra)
library(rmarkdown)
library(rvest)

url <- 'https://www.imdb.com/search/title/?title_type=tv_series&sort=num_votes,desc'

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

```
## <polite session> https://www.imdb.com/search/title/?title_type=tv_series&sort=num_votes,desc
##     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

title_list_sub <- as.data.frame(title_list[1:50])
```

```r
head(title_list_sub)
```

```
##      title_list[1:50]
## 1  1. Game of Thrones
## 2     2. Breaking Bad
## 3  3. Stranger Things
## 4          4. Friends
## 5 5. The Walking Dead
## 6         6. Sherlock
```

```r
tail(title_list_sub)
```

```
##    title_list[1:50]
## 45             <NA>
## 46             <NA>
## 47             <NA>
## 48             <NA>
## 49             <NA>
## 50             <NA>
```

```r
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))

# deleting columns 3 and 4 since it duplicated the columns
split_df <- split_df[-c(3:4)]

#renaming column 1 and 2
colnames(split_df) <- c("ranks","title")

# structure of splif_df
str(split_df)
```

```
## 'data.frame':    50 obs. of  2 variables:
##  $ ranks: chr  "1" "2" "3" "4" ...
##  $ title: chr  " Game of Thrones" " Breaking Bad" " Stranger Things" " Friends" ...
```

```r
head(split_df)
```

```
##   ranks             title
## 1     1   Game of Thrones
## 2     2      Breaking Bad
## 3     3   Stranger Things
## 4     4           Friends
## 5     5  The Walking Dead
## 6     6          Sherlock
```

```r
split_df
```

```
##              ranks                  title
## 1                1        Game of Thrones
## 2                2           Breaking Bad
## 3                3        Stranger Things
## 4                4                Friends
## 5                5       The Walking Dead
## 6                6               Sherlock
```

```
## 7                7    The Big Bang Theory
## 8                8                 Dexter
## 9                9  How I Met Your Mother
## 10              10             The Office
## 11              11               The Boys
## 12              12       Better Call Saul
## 13              13         Peaky Blinders
## 14              14         True Detective
## 15              15           Black Mirror
## 16              16         Rick and Morty
## 17              17                   Lost
## 18              18        The Mandalorian
## 19              19           Prison Break
## 20              20                Vikings
## 21              21            The Witcher
## 22              22         The Last of Us
## 23              23             Squid Game
## 24              24        Attack on Titan
## 25              25            Money Heist
## 26 Recently viewed        Recently viewed
## 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>
```