Mini-Project 2: Scrape your data

Math/CS 215: Intro to Data Science

Sarah Onstad-Hawes

Link to Data https://www.imdb.com/list/ls064721857/

Load Packages

```
library(tidyverse)
library(robotstxt)
library(rvest)
```

Creation of Pages

```
paths_allowed("https://www.imdb.com/list/ls064721857/")
```

```
## [1] TRUE
```

```
first_page <-("https://www.imdb.com/list/ls064721857/")
second_page <- ("https://www.imdb.com/list/ls064721857/?sort=list_order,asc&st_dt=&mode=detail&page=2")
third_page <- ("https://www.imdb.com/list/ls064721857/?sort=list_order,asc&st_dt=&mode=detail&page=3")

page1 <- read_html(first_page)
page2 <- read_html(second_page)
page3 <- read_html(third_page)
```

Creation of Movie Title Variable

```
Movie_Title1 <- page1 %>%
  html_nodes(".lister-item-header a") %>%
  html_text()

Movie_Title2 <- page2 %>%
  html_nodes(".lister-item-header a") %>%
  html_text()

Movie_Title3 <- page3 %>%
  html_nodes(".lister-item-header a") %>%
  html_text()

head(Movie_Title1, 3)
```

```
## [1] "12 Angry Men"
## [2] "The Green Mile"
## [3] "The Lord of the Rings: The Return of the King"
```

```
head(Movie_Title2, 3)
```

```
## [1] "First Blood"  "Alien"        "Spider-Man 2"
```

```
head(Movie_Title3, 3)
```

```
## [1] "The Dark Knight Rises"
## [2] "Big Fish"
## [3] "Mission: Impossible - Ghost Protocol"
```

Creation of Rank Variable

```
Rank1 <- page1 %>%
  html_nodes(".text-primary") %>%
  html_text() %>%
  as.numeric()

Rank2 <- page2 %>%
  html_nodes(".text-primary") %>%
  html_text() %>%
  as.numeric()

Rank3 <- page3 %>%
  html_nodes(".text-primary") %>%
  html_text() %>%
  as.numeric()

head(Rank1, 3)
```

```
## [1] 1 2 3
```

```
head(Rank2, 3)
```

```
## [1] 101 102 103
```

```
head(Rank3, 3)
```

```
## [1] 201 202 203
```

Creation of Movie Length Variable (in minutes)

```
Movie_Length1 <- page1 %>%
  html_nodes(".runtime")%>%
  html_text() %>%
  str_remove("min") %>%
  as.numeric()

Movie_Length2 <- page2 %>%
  html_nodes(".runtime")%>%
  html_text() %>%
  str_remove("min") %>%
  as.numeric()

Movie_Length3 <- page3 %>%
  html_nodes(".runtime")%>%
  html_text() %>%
  str_remove("min") %>%
  as.numeric()

head(Movie_Length1, 3)
```

```
## [1]  96 189 201
```

```
head(Movie_Length2, 3)
```

```
## [1]  93 117 127
```

```
head(Movie_Length3, 3)
```

```
## [1] 164 125 132
```

Creation of Movie Genre Variable

```
Movie_Genre1 <- page1 %>%
  html_nodes(".genre")%>%
  html_text() %>%
  str_remove("\n")

Movie_Genre2 <- page2 %>%
  html_nodes(".genre")%>%
  html_text() %>%
  str_remove("\n")

Movie_Genre3 <- page3 %>%
  html_nodes(".genre")%>%
  html_text() %>%
  str_remove("\n")

head(Movie_Genre1, 3)
```

```
## [1] "Crime, Drama            "
## [2] "Crime, Drama, Fantasy            "
## [3] "Action, Adventure, Drama            "
```

```
head(Movie_Genre2, 3)
```

```
## [1] "Action, Adventure            "
## [2] "Horror, Sci-Fi            "
## [3] "Action, Adventure, Sci-Fi            "
```

```
head(Movie_Genre3, 3)
```

```
## [1] "Action, Crime            "
## [2] "Adventure, Drama, Fantasy            "
## [3] "Action, Adventure, Thriller            "
```

Creation of Movie Rating Variable

```
Movie_Rating1 <- page1 %>%
  html_nodes(".certificate")%>%
  html_text()

Movie_Rating2 <- page2 %>%
  html_nodes(".certificate")%>%
  html_text()

Movie_Rating3 <- page3 %>%
  html_nodes(".certificate")%>%
  html_text()

head(Movie_Rating1, 3)
```

```
## [1] "Approved" "R"        "PG-13"
```

```
head(Movie_Rating2, 3)
```

```
## [1] "R"     "R"     "PG-13"
```

```
head(Movie_Rating3, 3)
```

```
## [1] "PG-13" "PG-13" "PG-13"
```

Creation of Star Rating Variable (out of 10)

```
Movie_Critic_Score1 <- page1 %>%
  html_nodes(".ipl-rating-star.small .ipl-rating-star__rating")%>%
  html_text() %>%
  as.numeric()

Movie_Critic_Score2 <- page2 %>%
  html_nodes(".ipl-rating-star.small .ipl-rating-star__rating")%>%
  html_text() %>%
  as.numeric()

Movie_Critic_Score3 <- page3 %>%
  html_nodes(".ipl-rating-star.small .ipl-rating-star__rating")%>%
  html_text() %>%
  as.numeric()

head(Movie_Critic_Score1, 3)
```

```
## [1] 9.0 8.6 8.9
```

```
head(Movie_Critic_Score2, 3)
```

```
## [1] 7.7 8.4 7.3
```

```
head(Movie_Critic_Score3, 3)
```

```
## [1] 8.4 8.0 7.4
```

Combine these vectors into 3 single data frame

```
Top_Movies1 <- tibble(Title=Movie_Title1, Rank=Rank1,
                      Length=Movie_Length1, Genre=Movie_Genre1,
                      Rating=Movie_Rating1, Starscore=Movie_Critic_Score1)

Top_Movies2 <- tibble(Title=Movie_Title2, Rank=Rank2,
                      Length=Movie_Length2, Genre=Movie_Genre2,
                      Rating=Movie_Rating2, Starscore=Movie_Critic_Score2)

Top_Movies3 <- tibble(Title=Movie_Title3, Rank=Rank3,
                      Length=Movie_Length3, Genre=Movie_Genre3,
                      Rating=Movie_Rating3, Starscore=Movie_Critic_Score3)
```

Combine into 1 single data frame

```
Top_Movies <- full_join(Top_Movies1, Top_Movies2)
Top_Movies <- full_join(Top_Movies, Top_Movies3)
Top_Movies
```

```
## # A tibble: 220 x 6
```

```
##    Title                     Rank Length Genre                  Rating Starscore
##    <chr>                    <dbl>  <dbl> <chr>                  <chr>      <dbl>
##  1 12 Angry Men                 1     96 "Crime, Drama        ~ Appro~       9
##  2 The Green Mile               2    189 "Crime, Drama, Fantas~ R            8.6
##  3 The Lord of the Rings: ~     3    201 "Action, Adventure, D~ PG-13        8.9
##  4 The Shawshank Redemption     4    142 "Drama            "    R            9.3
##  5 Memento                      5    113 "Mystery, Thriller   ~ R            8.4
##  6 Se7en                        6    127 "Crime, Drama, Myster~ R            8.6
##  7 The Dark Knight              7    152 "Action, Crime, Drama~ PG-13        9
##  8 Trainspotting                8     93 "Drama            "    R            8.1
##  9 Inception                    9    148 "Action, Adventure, S~ PG-13        8.8
## 10 Once Upon a Time in the~    10    165 "Western            "  PG-13        8.5
## # ... with 210 more rows
```

Export Data Frame

```
write.csv(Top_Movies, "/home/onstadsa/Math 215 - Fall 2021/Project 2/Top_Movies.csv")
view(Top_Movies)
```

4. Why you chose this data set, and what makes it interesting to you (10 pts)

I chose this data set because I really enjoy watching movies and there are so many different ways to rank the best movies. Everyone has different opinions on what the best movie/list of top movies are and I was interested in exploring if the length, genre, rating, and critic score have a significant effect on the ranking of the list of a users top 220 movies. It was also interesting to me that this persons rankings included a lot of movies I haven't seen before.