Mini Project 01 - IMDB Web Scraping

Load related library

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(rvest)
```

```
##
## Attaching package: 'rvest'
##
## The following object is masked from 'package:readr':
##
##     guess_encoding
```

Read Data From Website

```
url <- "https://www.imdb.com/search/title/?groups=top_100&sort=user_rating,desc"
#read html
imdb <- read_html(url)

#movie title
titles <- imdb %>%
  html_nodes("h3.lister-item-header") %>%
  html_text2()

#rating
ratings <- imdb %>%
  html_nodes("div.ratings-imdb-rating") %>%
  html_text2() %>%
  as.numeric()

#number of votes
num_votes <- imdb %>%
  html_nodes("p.sort-num_votes-visible") %>%
  html_text2()
```

Build Dataset and Show

```
#build a dataset
df <- data.frame(
  title = titles,
  rating = ratings,
  num_vote = num_votes
)

head(df)
```

```
##                                title rating
## 1 1. The Shawshank Redemption (1994)    9.3
## 2            2. The Godfather (1972)    9.2
## 3         3. Schindler's List (1993)    9.0
## 4          4. The Dark Knight (2008)    9.0
## 5             5. 12 Angry Men (1957)    9.0
## 6    6. The Godfather Part II (1974)    9.0
##                                                num_vote
## 1  Votes: 2,695,944 | Gross: $28.34M | Top 250: #1
## 2 Votes: 1,870,725 | Gross: $134.97M | Top 250: #2
## 3  Votes: 1,363,017 | Gross: $96.90M | Top 250: #6
## 4 Votes: 2,669,683 | Gross: $534.86M | Top 250: #3
## 5     Votes: 796,089 | Gross: $4.36M | Top 250: #5
## 6  Votes: 1,278,747 | Gross: $57.30M | Top 250: #4
```