Web Scraping

```
#install.packages("tidyverse")
#installed.packages("rvest")
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) # scrap data from internet
```

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

Movie from IMDB

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

```
## {html_document}
## <html xmlns:og="http://ogp.me/ns#" xmlns:fb="http://www.facebook.com/2008/fbml">
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-8 ...
## [2] <body id="styleguide-v2" class="fixed">\n            <img height="1" widt ...
```

title

```
titles = imdb %>%
  html_nodes("h3.lister-item-header") %>%
  html_text2()
head(titles)
```

```
## [1] "1. The Shawshank Redemption (1994)"
## [2] "2. The Godfather (1972)"
## [3] "3. The Dark Knight (2008)"
## [4] "4. Schindler's List (1993)"
## [5] "5. The Lord of the Rings: The Return of the King (2003)"
## [6] "6. The Godfather Part II (1974)"
```

rating (score)

```
ratings = imdb %>%
  html_nodes("div.ratings-imdb-rating") %>%
  html_text2() %>%
  as.numeric()
ratings[1:10]
```

```
##  [1] 9.3 9.2 9.0 9.0 9.0 9.0 9.0 8.9 8.8 8.8
```

amount of vote

```
vote = imdb %>%
  html_nodes("p.sort-num_votes-visible") %>%
  html_text2()
head(vote)
```

```
## [1] "Votes: 2,702,990 | Gross: $28.34M | Top 250: #1"
## [2] "Votes: 1,876,675 | Gross: $134.97M | Top 250: #2"
## [3] "Votes: 2,676,820 | Gross: $534.86M | Top 250: #3"
## [4] "Votes: 1,366,267 | Gross: $96.90M | Top 250: #6"
## [5] "Votes: 1,861,148 | Gross: $377.85M | Top 250: #7"
## [6] "Votes: 1,282,003 | Gross: $57.30M | Top 250: #4"
```

Movie top 50

```
df = data.frame(
  titles,
  ratings,
  vote
)
df
```

```
##                                                          titles ratings
## 1                           1. The Shawshank Redemption (1994)     9.3
## 2                                      2. The Godfather (1972)     9.2
## 3                                    3. The Dark Knight (2008)     9.0
## 4                                   4. Schindler’s List (1993)     9.0
## 5      5. The Lord of the Rings: The Return of the King (2003)     9.0
## 6                              6. The Godfather Part II (1974)     9.0
## 7                                       7. 12 Angry Men (1957)     9.0
## 8                                       8. Pulp Fiction (1994)     8.9
## 9  9. The Lord of the Rings: The Fellowship of the Ring (2001)     8.8
## 10                                        10. Inception (2010)     8.8
## 11                                       11. Fight Club (1999)     8.8
## 12                                     12. Forrest Gump (1994)     8.8
## 13            13. The Lord of the Rings: The Two Towers (2002)     8.8
## 14                  14. Il buono, il brutto, il cattivo (1966)     8.8
## 15                                     15. The Matrix (1999)     8.7
## 16                  16. One Flew Over the Cuckoo’s Nest (1975)     8.7
## 17                                     17. GoodFellas (1990)     8.7
## 18                          18. The Empire Strikes Back (1980)     8.7
## 19                                   19. Interstellar (2014)     8.6
## 20                                          20. Se7en (1995)     8.6
## 21                         21. The Silence of the Lambs (1991)     8.6
## 22                                   22. The Green Mile (1999)     8.6
## 23                                      23. Star Wars (1977)     8.6
## 24                              24. Saving Private Ryan (1998)     8.6
## 25                       25. Terminator 2: Judgment Day (1991)     8.6
## 26                    26. Sen to Chihiro no kamikakushi (2001)     8.6
## 27                                  27. La vita è bella (1997)     8.6
## 28                                   28. Cidade de Deus (2002)     8.6
## 29                            29. It’s a Wonderful Life (1946)     8.6
## 30                             30. Shichinin no samurai (1954)     8.6
## 31                                        31. Seppuku (1962)     8.6
## 32                                       32. Whiplash (2014)     8.5
## 33                                      33. Gladiator (2000)     8.5
## 34                                   34. Gisaengchung (2019)     8.5
## 35                                   35. The Departed (2006)     8.5
## 36                               36. Back to the Future (1985)     8.5
## 37                                   37. The Prestige (2006)     8.5
## 38                                   38. Apocalypse Now (1979)     8.5
## 39                                           39. Léon (1994)     8.5
## 40                                          40. Alien (1979)     8.5
## 41                               41. The Usual Suspects (1995)     8.5
## 42                                    42. The Lion King (1994)     8.5
## 43                               43. American History X (1998)     8.5
## 44                     44. Once Upon a Time in the West (1968)     8.5
## 45                                      45. The Pianist (2002)     8.5
## 46                                 46. The Intouchables (2011)     8.5
## 47                                     47. Casablanca (1942)     8.5
## 48                                         48. Psycho (1960)     8.5
## 49                                   49. Hotaru no haka (1988)     8.5
## 50                                      50. Rear Window (1954)     8.5
##                                                  vote
## 1    Votes: 2,702,990 | Gross: $28.34M | Top 250: #1
## 2   Votes: 1,876,675 | Gross: $134.97M | Top 250: #2
## 3   Votes: 2,676,820 | Gross: $534.86M | Top 250: #3
## 4    Votes: 1,366,267 | Gross: $96.90M | Top 250: #6
## 5   Votes: 1,861,148 | Gross: $377.85M | Top 250: #7
## 6    Votes: 1,282,003 | Gross: $57.30M | Top 250: #4
## 7       Votes: 798,527 | Gross: $4.36M | Top 250: #5
## 8   Votes: 2,075,085 | Gross: $107.93M | Top 250: #8
## 9   Votes: 1,890,524 | Gross: $315.54M | Top 250: #9
## 10 Votes: 2,374,814 | Gross: $292.58M | Top 250: #14
## 11  Votes: 2,147,865 | Gross: $37.03M | Top 250: #12
## 12 Votes: 2,099,931 | Gross: $330.25M | Top 250: #11
## 13 Votes: 1,680,515 | Gross: $342.55M | Top 250: #13
## 14     Votes: 767,669 | Gross: $6.10M | Top 250: #10
## 15 Votes: 1,928,945 | Gross: $171.48M | Top 250: #16
## 16 Votes: 1,015,135 | Gross: $112.00M | Top 250: #18
## 17  Votes: 1,172,520 | Gross: $46.84M | Top 250: #17
## 18 Votes: 1,302,663 | Gross: $290.48M | Top 250: #15
## 19 Votes: 1,860,847 | Gross: $188.02M | Top 250: #25
## 20 Votes: 1,668,946 | Gross: $100.13M | Top 250: #19
## 21 Votes: 1,445,285 | Gross: $130.74M | Top 250: #22
## 22 Votes: 1,313,987 | Gross: $136.80M | Top 250: #27
## 23 Votes: 1,375,080 | Gross: $322.74M | Top 250: #28
## 24 Votes: 1,403,164 | Gross: $216.54M | Top 250: #23
## 25 Votes: 1,108,365 | Gross: $204.84M | Top 250: #29
## 26    Votes: 773,251 | Gross: $10.06M | Top 250: #31
## 27    Votes: 701,613 | Gross: $57.60M | Top 250: #26
## 28     Votes: 762,232 | Gross: $7.56M | Top 250: #24
## 29                     Votes: 466,751 | Top 250: #21
## 30     Votes: 348,751 | Gross: $0.27M | Top 250: #20
## 31                      Votes: 58,982 | Top 250: #44
## 32    Votes: 878,845 | Gross: $13.09M | Top 250: #42
## 33 Votes: 1,513,406 | Gross: $187.71M | Top 250: #37
## 34    Votes: 823,006 | Gross: $53.37M | Top 250: #34
## 35 Votes: 1,336,437 | Gross: $132.38M | Top 250: #39
## 36 Votes: 1,217,702 | Gross: $210.61M | Top 250: #30
## 37  Votes: 1,345,095 | Gross: $53.09M | Top 250: #41
## 38    Votes: 673,655 | Gross: $83.47M | Top 250: #53
## 39  Votes: 1,172,099 | Gross: $19.50M | Top 250: #35
## 40    Votes: 890,783 | Gross: $78.90M | Top 250: #51
## 41  Votes: 1,092,949 | Gross: $23.34M | Top 250: #40
## 42 Votes: 1,068,403 | Gross: $422.78M | Top 250: #36
## 43   Votes: 1,130,653 | Gross: $6.72M | Top 250: #38
## 44     Votes: 333,072 | Gross: $5.32M | Top 250: #48
## 45    Votes: 841,219 | Gross: $32.57M | Top 250: #32
## 46    Votes: 867,682 | Gross: $13.18M | Top 250: #46
## 47     Votes: 576,441 | Gross: $1.02M | Top 250: #43
## 48    Votes: 677,824 | Gross: $32.00M | Top 250: #33
## 49                     Votes: 281,435 | Top 250: #45
## 50    Votes: 495,986 | Gross: $36.76M | Top 250: #49
```