Mini Project 01 - IMDB web scraping

```
library (tidyverse)
library (rvest) # scarpe data from internet
```

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

```
print(url)
```

```
[1] "https://www.imdb.com/search/title/?groups=top_100&sort=user_rating,desc"
```

```
#read html
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 .
```

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

```
titles
```

'1. The Shawshank Redemption (1994)' · '2. The Godfather (1972)' · '3. The Dark Knight (2008)' ·
'4. Schindler\'s List (1993)' · '5. The Lord of the Rings: The Return of the King (2003)' ·
'6. The Godfather Part II (1974)' · '7. 12 Angry Men (1957)' · '8. Pulp Fiction (1994)' · '9. Fight Club (1999)' ·
'10. Inception (2010)' · '11. The Lord of the Rings: The Fellowship of the Ring (2001)' ·
'12. Forrest Gump (1994)' · '13. Il buono, il brutto, il cattivo (1966)' ·
'14. The Lord of the Rings: The Two Towers (2002)' · '15. GoodFellas (1990)' · '16. The Matrix (1999)' ·
'17. One Flew Over the Cuckoo\'s Nest (1975)' · '18. The Empire Strikes Back (1980)' ·
'19. Interstellar (2014)' · '20. Se7en (1995)' · '21. The Silence of the Lambs (1991)' ·
'22. The Green Mile (1999)' · '23. Star Wars (1977)' · '24. Terminator 2: Judgment Day (1991)' ·
'25. Saving Private Ryan (1998)' · '26. Sen to Chihiro no kamikakushi (2001)' · '27. La vita è bella (1997)' ·
'28. Cidade de Deus (2002)' · '29. It\'s a Wonderful Life (1946)' · '30. Shichinin no samurai (1954)' ·
'31. Seppuku (1962)' · '32. Whiplash (2014)' · '33. Gladiator (2000)' · '34. Gisaengchung (2019)' ·
'35. The Departed (2006)' · '36. Léon (1994)' · '37. Apocalypse Now (1979)' · '38. The Prestige (2006)' ·
'39. Alien (1979)' · '40. Back to the Future (1985)' · '41. The Lion King (1994)' ·
'42. The Usual Suspects (1995)' · '43. American History X (1998)' · '44. The Intouchables (2011)' ·
'45. The Pianist (2002)' · '46. Once Upon a Time in the West (1968)' · '47. Psycho (1960)' ·
'48. Casablanca (1942)' · '49. Hotaru no haka (1988)' · '50. Rear Window (1954)'

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

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

```
num_votes
```

'Votes: 2,686,023 | Gross: $28.34M | Top 250: #1'

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

A data.frame: 6 × 3

	title	rating	num_vote
	<chr>	<dbl>	<chr>
1	1. The Shawshank Redemption (1994)	9.3	Votes: 2,686,023 | Gross: $28.34M | Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 2,686,023 | Gross: $28.34M | Top 250: #1
3	3. The Dark Knight (2008)	9.0	Votes: 2,686,023 | Gross: $28.34M | Top 250: #1
4	4. Schindler's List (1993)	9.0	Votes: 2,686,023 | Gross: $28.34M | Top 250: #1
5	5. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 2,686,023 | Gross: $28.34M | Top 250: #1
6	6. The Godfather Part II (1974)	9.0	Votes: 2,686,023 | Gross: $28.34M | Top 250: #1