Mini Project 01 - IMDb Web Scraping

```
library(tidyverse)
```

```
library(rvest)
```

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

```
print(urllink)
```

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

```
imdb <- read_html(urllink)
```

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

```
# rating
rating <- imdb %>%
    html_nodes("div.inline-block.ratings-imdb-rating") %>%
    html_text2()
```

```
rating[1:10]
```

'9.3' · '9.2' · '9.0' · '9.0' · '9.0' · '9.0' · '9.0' · '8.9' · '8.8' · '8.8'

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

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

```
df
```

A data.frame: 50 × 3

title	rating	num_vote
<chr>	<chr>	<chr>
1. The Shawshank Redemption (1994)	9.3	Votes: 2,663,685 | Gross: $28.34M | Top 250: #1
2. The Godfather (1972)	9.2	Votes: 1,845,916 | Gross: $134.97M | Top 250: #2
3. The Dark Knight (2008)	9.0	Votes: 2,636,656 | Gross: $534.86M | Top 250: #3
4. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,836,488 | Gross: $377.85M | Top 250: #7
5. Schindler's List (1993)	9.0	Votes: 1,348,865 | Gross: $96.90M | Top 250: #6
6. The Godfather Part II (1974)	9.0	Votes: 1,264,373 | Gross: $57.30M | Top 250: #4
7. 12 Angry Men (1957)	9.0	Votes: 786,621 | Gross: $4.36M | Top 250: #5
8. Pulp Fiction (1994)	8.9	Votes: 2,038,849 | Gross: $107.93M | Top 250: #8
9. Inception (2010)	8.8	Votes: 2,336,428 | Gross: $292.58M | Top 250: #14
10. The Lord of the Rings: The Two Towers (2002)	8.8	Votes: 1,658,290 | Gross: $342.55M | Top 250: #13
11. Fight Club (1999)	8.8	Votes: 2,109,070 | Gross: $37.03M | Top 250: #12
12. The Lord of the Rings: The Fellowship of the Ring (2001)	8.8	Votes: 1,865,553 | Gross: $315.54M | Top 250: #9
13. Forrest Gump (1994)	8.8	Votes: 2,064,488 | Gross: $330.25M | Top 250: #11
14. Il buono, il brutto, il cattivo (1966)	8.8	Votes: 759,186 | Gross: $6.10M | Top 250: #10
15. The Matrix (1999)	8.7	Votes: 1,903,244 | Gross: $171.48M | Top 250: #16
16. Goodfellas (1990)	8.7	Votes: 1,154,405 | Gross: $46.84M | Top 250: #17
17. The Empire Strikes Back (1980)	8.7	Votes: 1,286,372 | Gross: $290.48M | Top 250: #15
18. One Flew Over the Cuckoo's Nest (1975)	8.7	Votes: 1,004,542 | Gross: $112.00M | Top 250: #18
19. Interstellar (2014)	8.6	Votes: 1,810,217 | Gross: $188.02M | Top 250: #26
20. Cidade de Deus (2002)	8.6	Votes: 754,904 | Gross: $7.56M | Top 250: #23
21. Sen to Chihiro no kamikakushi (2001)	8.6	Votes: 758,959 | Gross: $10.06M | Top 250: #31
22. Saving Private Ryan (1998)	8.6	Votes: 1,384,460 | Gross: $216.54M | Top 250: #24
23. The Green Mile (1999)	8.6	Votes: 1,294,369 | Gross: $136.80M | Top 250: #27
24. La vita è bella (1997)	8.6	Votes: 692,575 | Gross: $57.60M | Top 250: #25
25. Se7en (1995)	8.6	Votes: 1,641,848 | Gross: $100.13M | Top 250: #19
26. Terminator 2: Judgment Day (1991)	8.6	Votes: 1,094,495 | Gross: $204.84M | Top 250: #29
27. The Silence of the Lambs (1991)	8.6	Votes: 1,425,169 | Gross: $130.74M | Top 250: #22
28. Star Wars (1977)	8.6	Votes: 1,358,938 | Gross: $322.74M | Top 250: #28

29. Seppuku (1962)	8.6	Votes: 57,283 | Top 250: #44
30. Shichinin no samurai (1954)	8.6	Votes: 345,685 | Gross: $0.27M | Top 250: #20
31. It's a Wonderful Life (1946)	8.6	Votes: 454,444 | Top 250: #21
32. Gisaengchung (2019)	8.5	Votes: 792,465 | Gross: $53.37M | Top 250: #34
33. Whiplash (2014)	8.5	Votes: 852,469 | Gross: $13.09M | Top 250: #42
34. The Intouchables (2011)	8.5	Votes: 854,236 | Gross: $13.18M | Top 250: #45
35. The Prestige (2006)	8.5	Votes: 1,326,821 | Gross: $53.09M | Top 250: #41
36. The Departed (2006)	8.5	Votes: 1,318,993 | Gross: $132.38M | Top 250: #39
37. The Pianist (2002)	8.5	Votes: 828,267 | Gross: $32.57M | Top 250: #33
38. Gladiator (2000)	8.5	Votes: 1,492,752 | Gross: $187.71M | Top 250: #37
39. American History X (1998)	8.5	Votes: 1,118,871 | Gross: $6.72M | Top 250: #38
40. The Usual Suspects (1995)	8.5	Votes: 1,081,802 | Gross: $23.34M | Top 250: #40
41. Léon (1994)	8.5	Votes: 1,155,831 | Gross: $19.50M | Top 250: #35
42. The Lion King (1994)	8.5	Votes: 1,053,211 | Gross: $422.78M | Top 250: #36
43. Nuovo Cinema Paradiso (1988)	8.5	Votes: 261,042 | Gross: $11.99M | Top 250: #52
44. Hotaru no haka (1988)	8.5	Votes: 276,795 | Top 250: #46
45. Back to the Future (1985)	8.5	Votes: 1,198,155 | Gross: $210.61M | Top 250: #30
46. Apocalypse Now (1979)	8.5	Votes: 665,857 | Gross: $83.47M | Top 250: #53
47. Alien (1979)	8.5	Votes: 879,306 | Gross: $78.90M | Top 250: #50
48. Once Upon a Time in the West (1968)	8.5	Votes: 329,241 | Gross: $5.32M | Top 250: #48
49. Psycho (1960)	8.5	Votes: 670,363 | Gross: $32.00M | Top 250: #32
50. Rear Window (1954)	8.5	Votes: 491,129 | Gross: $36.76M | Top 250: #49

Mini Project 02 - Spec Phone Database

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

```
url <- "https://specphone.com/Samsung-Galaxy-A04.html"
```

```
att <- url %>%
    read_html %>%
    html_nodes("div.topic") %>%
    html_text2

details <- url %>%
    read_html %>%
    html_nodes("div.detail") %>%
    html_text2
```

```
df <- data.frame(
    attribute = att,
    value = details
)
```

```
df
```

A data.frame: 31 × 2

attribute	value
<chr>	<chr>
วันเปิดตัว	ตุลาคม 2565
วันวางจำหน่าย	ยังไม่วางจำหน่าย
ขนาด	164.40 x 76.30 x 9.10 มม.
น้ำหนัก	192 กรัม
วัสดุ	Glass front, plastic back, plastic frame
SIM	รองรับ 2 ซิมการ์ด (nano sim, nano sim)
Technology	HSPA 42.2/5.76 Mbps, LTE-A
2G	850/900/1800/1900
3G	850/900/1900/2100
4G	850/900/1900/2100/2600
5G	-
ความเร็ว	HSPA 42.2/5.76 Mbps, LTE-A
ประเภท	PLS LCD
ขนาดหน้าจอ	6.50 นิ้ว
ความละเอียด	720 x 1600 pixels
ระบบปฏิบัติการ	Android 12
ชิปประมวลผล	Spreadtrum Unisoc SC9863A 1.6 GHz
ชิปกราฟิก	PowerVR GE8322
หน่วยความจำ	3 GB
ความจุ	32 GB
Memory Card	microSD (1)
กล้องหลัก	ตัวที่ 1: 50 MP, f/1.8, (wide), AF ตัวที่ 2: 2 MP, f/2.4, (depth)
ความละเอียดวีดีโอ	1080p@30fps
กล้องหน้า	ตัวที่ 1: 5 MP, f/2.2
Bluetooth	5.0, A2DP, LE
Wi-Fi	802.11 a/b/g/n/ac, dual-b
USB	Type-C
GPS	GLONASS, GALILEO, BDS
NFC	ไม่รองรับ
ความจุ	5,000 mAh
ประเภท	Non-removable Li-Po Batt

```
# All samsung smartphones
urlss <- read_html("https://specphone.com/brand/Samsung")
```

```
# links to all samsung smartphone
links <- urlss %>%
    html_nodes("li.mobile-brand-item a") %>%
    html_attr("href")
```

```
full_links <- paste0("http://specphone.com",links)
```

```
full_links
```

'http://specphone.com/Samsung-Galaxy-M13.html' · 'http://specphone.com/Samsung-Galaxy-A23.html' ·
'http://specphone.com/Samsung-Galaxy-A13.html' · 'http://specphone.com/Samsung-Galaxy-M32-5G.html' ·
'http://specphone.com/Samsung-Galaxy-A12-Nacho.html' ·
'http://specphone.com/Samsung-Galaxy-Pocket-Neo.html' · 'http://specphone.com/Samsung-Galaxy-Young.html' ·
'http://specphone.com/Samsung-Galaxy-J1-Mini.html' ·
'http://specphone.com/Samsung-Galaxy-A01-Core-1-16GB.html' ·
'http://specphone.com/Samsung-Galaxy-V-PLUS.html' · 'http://specphone.com/Samsung-Galaxy-Young-2.html' ·
'http://specphone.com/Samsung-Galaxy-M02.html' · 'http://specphone.com/Samsung-Galaxy-A11.html' ·
'http://specphone.com/Samsung-Galaxy-J2-Pro-2018.html' ·
'http://specphone.com/Samsung-Galaxy-A12-2021.html' ·
'http://specphone.com/Samsung-Galaxy-A21s-3-32GB.html' · 'http://specphone.com/Samsung-Galaxy-J5.html' ·
'http://specphone.com/Samsung-Galaxy-J4.html' · 'http://specphone.com/Samsung-Galaxy-Core-2-Duos.html' ·
'http://specphone.com/Samsung-Galaxy-Ace-Plus.html' · 'http://specphone.com/Samsung-Galaxy-A20.html' ·
'http://specphone.com/Samsung-Galaxy-Chat.html' · 'http://specphone.com/Samsung-Galaxy-Gio.html' ·
'http://specphone.com/Samsung-Galaxy-Tab-A7-Lite-LTE.html' ·
'http://specphone.com/Samsung-Galaxy-Tab-A-10.5WIFI.html' ·
'http://specphone.com/Samsung-Galaxy-Alpha.html' · 'http://specphone.com/Samsung-Galaxy-S3-Slim.html' ·
'http://specphone.com/Samsung-Galaxy-S4-zoom.html' · 'http://specphone.com/Samsung-Galaxy-Xcover-2.html' ·
'http://specphone.com/Samsung-Galaxy-Tab-8.9-3G-16GB.html' ·
'http://specphone.com/Samsung-Galaxy-Tab-A8-LTE-2021.html' ·
'http://specphone.com/Samsung-Galaxy-A8-2018.html' ·
'http://specphone.com/Samsung-Galaxy-Tab4-8.0-wifi.html' ·
'http://specphone.com/Samsung-Galaxy-M33-5G.html' · 'http://specphone.com/Samsung-Galaxy-A50.html' ·
'http://specphone.com/Samsung-Galaxy-E7.html' · 'http://specphone.com/Samsung-Galaxy-S6.html' ·
'http://specphone.com/Samsung-Galaxy-S20-FE.html' · 'http://specphone.com/Samsung-Galaxy-Tab-S4-WIFI.html' ·
'http://specphone.com/Samsung-Galaxy-S7.html' · 'http://specphone.com/Samsung-Galaxy-Note-5-Exynos.html' ·
'http://specphone.com/Samsung-Galaxy-TabPRO-12.2-LTE.html' ·
'http://specphone.com/Samsung-Galaxy-S4-Active.html' ·
'http://specphone.com/Samsung-Galaxy-Tab-Active-3.html' ·
'http://specphone.com/Samsung-Galaxy-Tab-S3-9.7.html' · 'http://specphone.com/Samsung-Galaxy-S6-edge.html' ·
'http://specphone.com/Samsung-Galaxy-Note-4-Exynos.html' ·
'http://specphone.com/Samsung-Galaxy-Round.html' ·
'http://specphone.com/Samsung-Galaxy-Note-20-Ultra-5G.html' · 'http://specphone.com/Samsung-ATIV-Q.html' ·
'http://specphone.com/Samsung-ATIV-Smart-PC-PRO.html' ·
'http://specphone.com/Samsung-Galaxy-S22-Ultra12-128GB.html' ·
'http://specphone.com/Samsung-Galaxy-Z-Flip-5G.html' · 'http://specphone.com/Samsung-Galaxy-Z-Flip.html' ·
'http://specphone.com/Samsung-Galaxy-Tab-S8-Ultra-5G.html' ·
'http://specphone.com/Samsung-Galaxy-S21-Ultra-16-512GB.html' ·
'http://specphone.com/Samsung-Galaxy-S10-Plus-Ram-12GB.html' ·
'http://specphone.com/Samsung-Galaxy-Z-Fold-3.html' · 'http://specphone.com/Samsung-Galaxy-Z-Fold4.html' ·
'http://specphone.com/Samsung-Galaxy-Z-Fold-2-5G.html'

```
result <- data.frame()

for (link in full_links[1:5]) {
    ss_topic <- link %>%
    read_html %>%
    html_nodes("div.topic") %>%
    html_text2()

    ss_detail <- link %>%
    read_html %>%
    html_nodes("div.detail") %>%
    html_text2()

    tmp <- data.frame(attribute = ss_topic,
                      value = ss_detail)
        result <- bind_rows(result, tmp)
        print("Progress ...")
}
```

```
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
```

```
print(head(result),3)
```

```
    attribute                                        value
1     วันเปิดตัว                                 มิถุนายน 2565
2 วันวางจำหน่าย                                ยังไม่วางจำหน่าย
3        ขนาด                 165.40 x 76.90 x 8.40 มม.
4       น้ำหนัก                                      192 กรัม
5         วัสดุ Glass front, plastic back, plastic frame
6         SIM      รองรับ 2 ซิมการ์ด (nano sim, nano sim)
```

```
# write csv
write_csv(result, "result_ss_phone_csv")
```