Mini Project01 - IMDB web scraping

```
library(tidyverse)
library(rvest) # scrape 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[1:10]
```

```
'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. The Godfather Part II (1974)' ·
'6. Schindler\'s List (1993)' · '7. 12 Angry Men (1957)' · '8. Pulp Fiction (1994)' ·
'9. The Lord of the Rings: The Fellowship of the Ring (2001)' · '10. Fight Club (1999)'
```

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

```
ratings[1:10]
```

```
9.3 · 9.2 · 9 · 9 · 9 · 9 · 9 · 8.9 · 8.8 · 8.8
```

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

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

```
df
```

A data.frame: 50 × 3

title	rating	num_vote
<chr>	<dbl>	<chr>
1. The Shawshank Redemption (1994)	9.3	Votes: 2,680,735 | Gross: $28.34M | Top 250: #1
2. The Godfather (1972)	9.2	Votes: 1,858,518 | Gross: $134.97M | Top 250: #2
3. The Dark Knight (2008)	9.0	Votes: 2,654,049 | Gross: $534.86M | Top 250: #3
4. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,847,326 | Gross: $377.85M | Top 250: #7
5. The Godfather Part II (1974)	9.0	Votes: 1,272,126 | Gross: $57.30M | Top 250: #4
6. Schindler's List (1993)	9.0	Votes: 1,356,570 | Gross: $96.90M | Top 250: #6
7. 12 Angry Men (1957)	9.0	Votes: 791,915 | Gross: $4.36M | Top 250: #5
8. Pulp Fiction (1994)	8.9	Votes: 2,056,022 | Gross: $107.93M | Top 250: #8
9. The Lord of the Rings: The Fellowship of the Ring (2001)	8.8	Votes: 1,876,709 | Gross: $315.54M | Top 250: #9
10. Fight Club (1999)	8.8	Votes: 2,126,278 | Gross: $37.03M | Top 250: #12
11. Inception (2010)	8.8	Votes: 2,353,576 | Gross: $292.58M | Top 250: #14
12. Forrest Gump (1994)	8.8	Votes: 2,080,020 | Gross: $330.25M | Top 250: #11
13. The Lord of the Rings: The Two Towers (2002)	8.8	Votes: 1,668,134 | Gross: $342.55M | Top 250: #13
14. Il buono, il brutto, il cattivo (1966)	8.8	Votes: 763,089 | Gross: $6.10M | Top 250: #10
15. Goodfellas (1990)	8.7	Votes: 1,162,901 | Gross: $46.84M | Top 250: #17
16. The Matrix (1999)	8.7	Votes: 1,914,276 | Gross: $171.48M | Top 250: #16
17. One Flew Over the Cuckoo's Nest (1975)	8.7	Votes: 1,009,380 | Gross: $112.00M | Top 250: #18
18. The Empire Strikes Back (1980)	8.7	Votes: 1,293,710 | Gross: $290.48M | Top 250: #15
19. It's a Wonderful Life (1946)	8.6	Votes: 463,688 | Top 250: #21
20. Interstellar (2014)	8.6	Votes: 1,832,288 | Gross: $188.02M | Top 250: #26
21. Se7en (1995)	8.6	Votes: 1,653,807 | Gross: $100.13M | Top 250: #19
22. The Green Mile (1999)	8.6	Votes: 1,303,259 | Gross: $136.80M | Top 250: #27
23. Star Wars (1977)	8.6	Votes: 1,366,304 | Gross: $322.74M | Top 250: #28
24. The Silence of the Lambs (1991)	8.6	Votes: 1,434,082 | Gross: $130.74M | Top 250: #22
25. Terminator 2: Judgment Day (1991)	8.6	Votes: 1,100,872 | Gross: $204.84M | Top 250: #29
26. Saving Private Ryan (1998)	8.6	Votes: 1,392,940 | Gross: $216.54M | Top 250: #24
27. Cidade de Deus (2002)	8.6	Votes: 758,380 | Gross: $7.56M | Top 250: #23
28. Sen to Chihiro no kamikakushi (2001)	8.6	Votes: 765,308 | Gross: $10.06M | Top 250: #31
29. La vita è bella (1997)	8.6	Votes: 696,632 | Gross: $57.60M | Top 250: #25
30. Shichinin no samurai (1954)	8.6	Votes: 347,298 | Gross: $0.27M | Top 250: #20
31. Seppuku (1962)	8.6	Votes: 58,045 | Top 250: #44
32. Whiplash (2014)	8.5	Votes: 863,847 | Gross: $13.09M | Top 250: #42

33. Gladiator (2000)	8.5	Votes: 1,502,031 | Gross: $187.71M | Top 250: #37
34. Gisaengchung (2019)	8.5	Votes: 805,181 | Gross: $53.37M | Top 250: #34
35. Back to the Future (1985)	8.5	Votes: 1,207,267 | Gross: $210.61M | Top 250: #30
36. Léon (1994)	8.5	Votes: 1,162,715 | Gross: $19.50M | Top 250: #35
37. Alien (1979)	8.5	Votes: 884,755 | Gross: $78.90M | Top 250: #51
38. The Departed (2006)	8.5	Votes: 1,327,021 | Gross: $132.38M | Top 250: #39
39. The Prestige (2006)	8.5	Votes: 1,334,865 | Gross: $53.09M | Top 250: #41
40. American History X (1998)	8.5	Votes: 1,124,343 | Gross: $6.72M | Top 250: #38
41. Apocalypse Now (1979)	8.5	Votes: 669,379 | Gross: $83.47M | Top 250: #53
42. Rear Window (1954)	8.5	Votes: 493,576 | Gross: $36.76M | Top 250: #49
43. The Usual Suspects (1995)	8.5	Votes: 1,087,077 | Gross: $23.34M | Top 250: #40
44. The Lion King (1994)	8.5	Votes: 1,059,873 | Gross: $422.78M | Top 250: #36
45. Once Upon a Time in the West (1968)	8.5	Votes: 331,299 | Gross: $5.32M | Top 250: #48
46. The Intouchables (2011)	8.5	Votes: 860,475 | Gross: $13.18M | Top 250: #46
47. The Pianist (2002)	8.5	Votes: 833,918 | Gross: $32.57M | Top 250: #33
48. Casablanca (1942)	8.5	Votes: 573,677 | Gross: $1.02M | Top 250: #43
49. Psycho (1960)	8.5	Votes: 673,868 | Gross: $32.00M | Top 250: #32
50. Hotaru no haka (1988)	8.5	Votes: 279,114 | Top 250: #45

Mini Project02 - SpecPhone Database

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

```
Warning message in system("timedatectl", intern = TRUE):
“running command 'timedatectl' had status 1”
Warning message:
“Failed to locate timezone database”
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

✓ ggplot2 3.3.5     ✓ purrr   0.3.4
✓ tibble  3.1.5     ✓ dplyr   1.0.7
✓ tidyr   1.1.4     ✓ stringr 1.4.0
✓ readr   2.0.2     ✓ forcats 0.5.1

── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter()  masks stats::filter()
✖ purrr::flatten() masks jsonlite::flatten()
✖ dplyr::lag()     masks stats::lag()


Attaching package: ‘rvest’
```

```
url <- "https://specphone.com/ZTE-nubia-Red-Magic-8-Pro-.html"
```

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

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

```
data.frame(attibute = att, value = value)
```

A data.frame: 32 × 2

attibute	value
<chr>	<chr>
วันเปิดตัว	ธันวาคม 2565
วันวางจำหน่าย	ยังไม่วางจำหน่าย
ขนาด	164.00 x 76.40 x 8.90 มม.
น้ำหนัก	230 กรัม
วัสดุ	Glass front, glass back, aluminum frame
SIM	รองรับ 2 ซิมการ์ด (nano sim, nano sim)
Technology	HSPA, LTE-A, 5G
2G	850/900/1800/1900
3G	850/900/1900/2100
4G	850/900/1900/2100/2600
5G	2100/2600/3500/4700
ความเร็ว	HSPA, LTE-A, 5G
ประเภท	AMOLED
ขนาดหน้าจอ	6.80 นิ้ว
ความละเอียด	1116 x 2480 pixels
ระบบปฏิบัติการ	Android 13
ชิปประมวลผล	Qualcomm Snapdragon 8 Gen 2 SM8550 3.2 GHz
ชิปกราฟิก	Adreno 740
หน่วยความจำ	12 GB
ความจุ	256 GB
Memory Card	ไม่รองรับ
กล้องหลัก	ตัวที่ 1: 50 MP, f/1.8, (wide), 1/1.57 ตัวที่ 2: 8 MP, f/2.2, 120˚, 13mm (ultrawide), 1/4.0 ตัวที่ 3: 2 MP, f/2.4, (macro)
ความละเอียดวีดีโอ	8K@30fps, 4K@30/60fps, 1080p@30/60/120/240fps
กล้องหน้า	ตัวที่ 1: 16 MP, (wide), under display
Bluetooth	5.3, A2DP, LE
Wi-Fi	802.11 a/b/g/n/ac/6e, dua
USB	Type-C
GPS	GPS (L1+L5), GLONASS, BDS
NFC	รอบรับ
ความจุ	5,000 mAh
ประเภท	Non-removable Li-Po Batt
Fast Charging	รองรับ (165W)

```r
# All Samsung Smartphones
samsung_url <- read_html("https://specphone.com/brand/Samsung")
```

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

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

```r
result <- data.frame()

for (link in full_links[1:10]) {
    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(attibute = ss_topic, value = ss_detail)

    result <- bind_rows(result, tmp)
    print("pregress...")
}
```

```
[1] "pregress..."
[1] "pregress..."
[1] "pregress..."
[1] "pregress..."
[1] "pregress..."
[1] "pregress..."
[1] "pregress..."
[1] "pregress..."
[1] "pregress..."
[1] "pregress..."
```

```
write_csv(result, "result_ss_phone.csv")
```

```
print(head(result))
```

```
     attibute                                    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)
```