Mini Project 01 - IMDB Scraping

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

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

```r
print(url)
```

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

```r
# read html
imdb <- read_html(url)
```

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

```r
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 .
```

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

```r
titles[1:10]
```

'1. The Shawshank Redemption (1994)' · '2. The Godfather (1972)' · '3. The Dark Knight (2008)' ·
'4. The Godfather Part II (1974)' · '5. Schindler\'s List (1993)' · '6. 12 Angry Men (1957)' ·
'7. The Lord of the Rings: The Return of the King (2003)' · '8. Pulp Fiction (1994)' ·
'9. The Lord of the Rings: The Fellowship of the Ring (2001)' · '10. Inception (2010)'

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

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

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

```r
num_votes[]
```

'Votes: 2,706,500 | Gross: $28.34M | Top 250: #1' · 'Votes: 1,879,480 | Gross: $134.97M | Top 250: #2' ·
'Votes: 2,679,924 | Gross: $534.86M | Top 250: #3' · 'Votes: 1,283,440 | Gross: $57.30M | Top 250: #4' ·
'Votes: 1,367,747 | Gross: $96.90M | Top 250: #6' · 'Votes: 799,551 | Gross: $4.36M | Top 250: #5' ·
'Votes: 1,863,125 | Gross: $377.85M | Top 250: #7' · 'Votes: 2,077,873 | Gross: $107.93M | Top 250: #8' ·
'Votes: 1,892,594 | Gross: $315.54M | Top 250: #9' · 'Votes: 2,377,927 | Gross: $292.58M | Top 250: #14' ·
'Votes: 2,150,976 | Gross: $37.03M | Top 250: #12' · 'Votes: 2,102,851 | Gross: $330.25M | Top 250: #11' ·
'Votes: 1,682,334 | Gross: $342.55M | Top 250: #13' · 'Votes: 768,437 | Gross: $6.10M | Top 250: #10' ·
'Votes: 1,173,959 | Gross: $46.84M | Top 250: #17' · 'Votes: 1,016,117 | Gross: $112.00M | Top 250: #18' ·
'Votes: 1,931,207 | Gross: $171.48M | Top 250: #16' · 'Votes: 1,303,985 | Gross: $290.48M | Top 250: #15' ·
'Votes: 1,864,885 | Gross: $188.02M | Top 250: #25' · 'Votes: 1,447,084 | Gross: $130.74M | Top 250: #22' ·
'Votes: 1,315,690 | Gross: $136.80M | Top 250: #27' · 'Votes: 1,671,211 | Gross: $100.13M | Top 250: #19' ·
'Votes: 1,376,379 | Gross: $322.74M | Top 250: #28' · 'Votes: 1,109,521 | Gross: $204.84M | Top 250: #29' ·
'Votes: 774,477 | Gross: $10.06M | Top 250: #31' · 'Votes: 1,404,690 | Gross: $216.54M | Top 250: #23' ·
'Votes: 762,928 | Gross: $7.56M | Top 250: #24' · 'Votes: 702,481 | Gross: $57.60M | Top 250: #26' ·
'Votes: 467,170 | Top 250: #21' · 'Votes: 349,061 | Gross: $0.27M | Top 250: #20' · 'Votes: 59,150 | Top 250: #44' ·
'Votes: 1,337,927 | Gross: $132.38M | Top 250: #39' · 'Votes: 881,060 | Gross: $13.09M | Top 250: #42' ·
'Votes: 825,493 | Gross: $53.37M | Top 250: #34' · 'Votes: 1,515,080 | Gross: $187.71M | Top 250: #37' ·
'Votes: 1,219,292 | Gross: $210.61M | Top 250: #30' · 'Votes: 1,346,541 | Gross: $53.09M | Top 250: #41' ·
'Votes: 674,359 | Gross: $83.47M | Top 250: #53' · 'Votes: 891,741 | Gross: $78.90M | Top 250: #51' ·
'Votes: 1,173,433 | Gross: $19.50M | Top 250: #35' · 'Votes: 1,069,696 | Gross: $422.78M | Top 250: #36' ·
'Votes: 1,093,895 | Gross: $23.34M | Top 250: #40' · 'Votes: 1,131,687 | Gross: $6.72M | Top 250: #38' ·
'Votes: 842,331 | Gross: $32.57M | Top 250: #32' · 'Votes: 576,971 | Gross: $1.02M | Top 250: #43' ·
'Votes: 868,841 | Gross: $13.18M | Top 250: #46' · 'Votes: 678,546 | Gross: $32.00M | Top 250: #33' ·
'Votes: 333,397 | Gross: $5.32M | Top 250: #48' · 'Votes: 281,884 | Top 250: #45' ·
'Votes: 264,540 | Gross: $11.99M | Top 250: #50'

```r
# 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: 2,706,500	Gross: $28.34M	Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,879,480	Gross: $134.97M	Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,679,924	Gross: $534.86M	Top 250: #3
4	4. The Godfather Part II (1974)	9.0	Votes: 1,283,440	Gross: $57.30M	Top 250: #4
5	5. Schindler's List (1993)	9.0	Votes: 1,367,747	Gross: $96.90M	Top 250: #6
6	6. 12 Angry Men (1957)	9.0	Votes: 799,551	Gross: $4.36M	Top 250: #5

Mini Project 02 - SpecPhone Database

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

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

```r
specphone <- read_html(url)

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

```r
values <- specphone %>%
    html_nodes("div.detail") %>%
    html_text2()
```

```r
Galaxy <- data.frame(
    attribute = att,
    value = values
)

Galaxy[]
```

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 smartphone
samsung_url <- read_html("https://specphone.com/brand/Samsung")
```

```
# link to Samsung Smart phone
link <- samsung_url %>%
    html_nodes("li.mobile-brand-item a") %>%
    html_attr("href")
```

link

'/Samsung-Galaxy-M13.html' · '/Samsung-Galaxy-A23.html' · '/Samsung-Galaxy-A13.html' ·
'/Samsung-Galaxy-M32-5G.html' · '/Samsung-Galaxy-A12-Nacho.html' · '/Samsung-Galaxy-Pocket-Neo.html' ·
'/Samsung-Galaxy-Young.html' · '/Samsung-Galaxy-J1-Mini.html' · '/Samsung-Galaxy-A01-Core-1-16GB.html' ·
'/Samsung-Galaxy-V-PLUS.html' · '/Samsung-Galaxy-Young-2.html' · '/Samsung-Galaxy-M02.html' ·
'/Samsung-Galaxy-A11.html' · '/Samsung-Galaxy-J2-Pro-2018.html' · '/Samsung-Galaxy-A12-2021.html' ·
'/Samsung-Galaxy-A21s-3-32GB.html' · '/Samsung-Galaxy-J5.html' · '/Samsung-Galaxy-J4.html' ·
'/Samsung-Galaxy-Core-2-Duos.html' · '/Samsung-Galaxy-Ace-Plus.html' · '/Samsung-Galaxy-A20.html' ·
'/Samsung-Galaxy-Chat.html' · '/Samsung-Galaxy-Gio.html' · '/Samsung-Galaxy-Tab-A7-Lite-LTE.html' ·
'/Samsung-Galaxy-Tab-A-10.5WIFI.html' · '/Samsung-Galaxy-Alpha.html' · '/Samsung-Galaxy-S3-Slim.html' ·
'/Samsung-Galaxy-S4-zoom.html' · '/Samsung-Galaxy-Xcover-2.html' · '/Samsung-Galaxy-Tab-8.9-3G-16GB.html' ·
'/Samsung-Galaxy-Tab-A8-LTE-2021.html' · '/Samsung-Galaxy-A8-2018.html' ·
'/Samsung-Galaxy-Tab4-8.0-wifi.html' · '/Samsung-Galaxy-M33-5G.html' · '/Samsung-Galaxy-A50.html' ·
'/Samsung-Galaxy-E7.html' · '/Samsung-Galaxy-S6.html' · '/Samsung-Galaxy-S20-FE.html' ·
'/Samsung-Galaxy-Tab-S4-WIFI.html' · '/Samsung-Galaxy-S7.html' · '/Samsung-Galaxy-Note-5-Exynos.html' ·
'/Samsung-Galaxy-TabPRO-12.2-LTE.html' · '/Samsung-Galaxy-S4-Active.html' ·
'/Samsung-Galaxy-Tab-Active-3.html' · '/Samsung-Galaxy-Tab-S3-9.7.html' · '/Samsung-Galaxy-S6-edge.html' ·
'/Samsung-Galaxy-Note-4-Exynos.html' · '/Samsung-Galaxy-Round.html' ·
'/Samsung-Galaxy-Note-20-Ultra-5G.html' · '/Samsung-ATIV-Q.html' · '/Samsung-ATIV-Smart-PC-PRO.html' ·
'/Samsung-Galaxy-S23-Ultra-5G8-256GB.html' · '/Samsung-Galaxy-S22-Ultra12-128GB.html' ·
'/Samsung-Galaxy-Z-Flip-5G.html' · '/Samsung-Galaxy-Z-Flip.html' · '/Samsung-Galaxy-Tab-S8-Ultra-5G.html' ·
'/Samsung-Galaxy-S21-Ultra-16-512GB.html' · '/Samsung-Galaxy-S23-Ultra-5G.html' ·
'/Samsung-Galaxy-S10-Plus-Ram-12GB.html' · '/Samsung-Galaxy-Z-Fold-3.html'

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

'https://specphone.com/Samsung-Galaxy-M13.html' · 'https://specphone.com/Samsung-Galaxy-A23.html' ·
'https://specphone.com/Samsung-Galaxy-A13.html' · 'https://specphone.com/Samsung-Galaxy-M32-5G.html' ·
'https://specphone.com/Samsung-Galaxy-A12-Nacho.html' ·
'https://specphone.com/Samsung-Galaxy-Pocket-Neo.html' ·
'https://specphone.com/Samsung-Galaxy-Young.html' · 'https://specphone.com/Samsung-Galaxy-J1-Mini.html' ·
'https://specphone.com/Samsung-Galaxy-A01-Core-1-16GB.html' ·
'https://specphone.com/Samsung-Galaxy-V-PLUS.html' · 'https://specphone.com/Samsung-Galaxy-Young-2.html' ·
'https://specphone.com/Samsung-Galaxy-M02.html' · 'https://specphone.com/Samsung-Galaxy-A11.html' ·
'https://specphone.com/Samsung-Galaxy-J2-Pro-2018.html' ·
'https://specphone.com/Samsung-Galaxy-A12-2021.html' ·
'https://specphone.com/Samsung-Galaxy-A21s-3-32GB.html' · 'https://specphone.com/Samsung-Galaxy-J5.html' ·
'https://specphone.com/Samsung-Galaxy-J4.html' · 'https://specphone.com/Samsung-Galaxy-Core-2-Duos.html' ·
'https://specphone.com/Samsung-Galaxy-Ace-Plus.html' · 'https://specphone.com/Samsung-Galaxy-A20.html' ·
'https://specphone.com/Samsung-Galaxy-Chat.html' · 'https://specphone.com/Samsung-Galaxy-Gio.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-A7-Lite-LTE.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-A-10.5WIFI.html' ·
'https://specphone.com/Samsung-Galaxy-Alpha.html' · 'https://specphone.com/Samsung-Galaxy-S3-Slim.html' ·
'https://specphone.com/Samsung-Galaxy-S4-zoom.html' ·
'https://specphone.com/Samsung-Galaxy-Xcover-2.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-8.9-3G-16GB.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-A8-LTE-2021.html' ·
'https://specphone.com/Samsung-Galaxy-A8-2018.html' ·
'https://specphone.com/Samsung-Galaxy-Tab4-8.0-wifi.html' ·
'https://specphone.com/Samsung-Galaxy-M33-5G.html' · 'https://specphone.com/Samsung-Galaxy-A50.html' ·
'https://specphone.com/Samsung-Galaxy-E7.html' · 'https://specphone.com/Samsung-Galaxy-S6.html' ·
'https://specphone.com/Samsung-Galaxy-S20-FE.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-S4-WIFI.html' · 'https://specphone.com/Samsung-Galaxy-S7.html' ·
'https://specphone.com/Samsung-Galaxy-Note-5-Exynos.html' ·
'https://specphone.com/Samsung-Galaxy-TabPRO-12.2-LTE.html' ·
'https://specphone.com/Samsung-Galaxy-S4-Active.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-Active-3.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-S3-9.7.html' ·
'https://specphone.com/Samsung-Galaxy-S6-edge.html' ·
'https://specphone.com/Samsung-Galaxy-Note-4-Exynos.html' ·
'https://specphone.com/Samsung-Galaxy-Round.html' ·
'https://specphone.com/Samsung-Galaxy-Note-20-Ultra-5G.html' · 'https://specphone.com/Samsung-ATIV-Q.html' ·
'https://specphone.com/Samsung-ATIV-Smart-PC-PRO.html' ·
'https://specphone.com/Samsung-Galaxy-S23-Ultra-5G8-256GB.html' ·
'https://specphone.com/Samsung-Galaxy-S22-Ultra12-128GB.html' ·
'https://specphone.com/Samsung-Galaxy-Z-Flip-5G.html' · 'https://specphone.com/Samsung-Galaxy-Z-Flip.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-S8-Ultra-5G.html' ·
'https://specphone.com/Samsung-Galaxy-S21-Ultra-16-512GB.html' ·
'https://specphone.com/Samsung-Galaxy-S23-Ultra-5G.html' ·
'https://specphone.com/Samsung-Galaxy-S10-Plus-Ram-12GB.html' ·
'https://specphone.com/Samsung-Galaxy-Z-Fold-3.html'

```r
# For loop
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...")
}

print(result)
```

```
[1] "Progress..."
[1] "Progress..."
[1] "Progress..."
[1] "Progress..."
[1] "Progress..."
           attribute
1          วันเปิดตัว
2      วันวางจำหน่าย
3              ขนาด
4            น้ำหนัก
5              วัสดุ
6                SIM
7         Technology
8                 2G
9                 3G
10                4G
11                5G
12         ความเร็ว
13           ประเภท
14       ขนาดหน้าจอ
```

```r
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")
```