Mini Project 01 - IMDB Web Scraping

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

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

```
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/fb
[1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-
[2] <body id="styleguide-v2" class="fixed">\n            <img height="1" wid
```

```
# 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. 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)'

```
# 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

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

```
# 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: 2,670,356 | Gross: $28.34M | Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,850,672 | Gross: $134.97M | Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,643,204 | Gross: $534.86M | Top 250: #3
4	4. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,840,524 | Gross: $377.85M | Top 250: #7
5	5. Schindler's List (1993)	9.0	Votes: 1,351,931 | Gross: $96.90M | Top 250: #6
6	6. The Godfather Part II (1974)	9.0	Votes: 1,267,268 | Gross: $57.30M | Top 250: #4

Mini Project 02 - Specphone Phone Database

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

```
url <- read_html("https://specphone.com/Huawei-Mate-50-Pro.html")
```

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

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

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

A data.frame: 33 × 2

attribute	value
<chr>	<chr>
วันเปิดตัว	พฤศจิกายน 2565
วันวางจำหน่าย	พฤศจิกายน 2565, ยังไม่วางจำหน่าย
ขนาด	162.10 x 75.50 x 8.50 มม.
น้ำหนัก	205 กรัม
วัสดุ	Glass front, glass back or eco leather back, aluminum frame
SIM	รองรับ 2 ซิมการ์ด (nano sim, nano sim)
Technology	HSPA, LTE-A
2G	850/900/1800/1900
3G	850/900/1900/2100
4G	850/900/1900/2100/2600
5G	-
ความเร็ว	HSPA, LTE-A
ประเภท	OLED
ขนาดหน้าจอ	6.74 นิ้ว
ความละเอียด	1212 x 2616 pixels
ระบบปฏิบัติการ	EMUI 13
ชิปประมวลผล	Qualcomm Snapdragon 8+ Gen 1 SM8475 3.19 GHz
ชิปกราฟิก	Adreno 730
หน่วยความจำ	8 GB
ความจุ	256 GB
Memory Card	microSD (256)
กล้องหลัก	ตัวที่ 1: 50 MP, f/1.4-f/4.0, 24mm (wide), PDAF, Laser AF, OIS ตัวที่ 2: 64 MP, f/3.5, 90mm (periscope telephoto), PDAF, OIS, 3.5x optical zoom ตัวที่ 3: 13 MP, f/2.2, 13mm, 120˚ (ultrawide), PDAF
ความละเอียดวีดีโอ	4K@30/60fps, 1080p@30/60/120/240/480fps, 720p@960fps, 720p@3840fps, HDR, gyro-EIS
กล้องหน้า	ตัวที่ 1: 13 MP, f/2.4, 18mm (ultrawide)
Bluetooth	5.2, A2DP, LE
Wi-Fi	802.11 a/b/g/n/ac/6, dual
USB	Type-C
GPS	with dual-band A-GPS, GLO
NFC	รอบรับ
ความจุ	4,700 mAh
ประเภท	Non-removable Li-Po Batt
Wireless Charging	รองรับ
Fast Charging	รองรับ (66W)

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

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

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

```
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(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 ..."
[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")
```