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. 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: 2,665,057 | Gross: $28.34M | Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,846,871 | Gross: $134.97M | Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,638,048 | Gross: $534.86M | Top 250: #3
4	4. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,837,315 | Gross: $377.85M | Top 250: #7
5	5. Schindler's List (1993)	9.0	Votes: 1,349,472 | Gross: $96.90M | Top 250: #6
6	6. The Godfather Part II (1974)	9.0	Votes: 1,264,929 | Gross: $57.30M | Top 250: #4

Mini Project02 - Specphone Phone Database

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

```
url <- read_html("https://specphone.com/Samsung-Galaxy-Z-Fold4.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: 32 × 2

attribute	value
<chr>	<chr>
วันเปิดตัว	**สิงหาคม** 2565
วันวางจำหน่าย	**สิงหาคม** 2565, **วางจำหน่ายแล้ว**
ขนาด	155.10 x 130.10 x 6.30 **มม.**
น้ำหนัก	263 **กรัม**
วัสดุ	Plastic front (opened), glass back (Gorilla Glass Victus+), aluminum frame
SIM	**รองรับ** 2 **ซิมการ์ด** (nano sim, nano sim)
Technology	HSPA 42.2/5.76 Mbps, LTE-A (7CA) Cat20 2000/200 Mbps, 5G
2G	850/900/1800/1900
3G	850/900/1900/2100
4G	850/900/1900/2100/2600
5G	2100/2600/3500/4700
ความเร็ว	HSPA 42.2/5.76 Mbps, LTE-A (7CA) Cat20 2000/200 Mbps, 5G
ประเภท	Foldable Dynamic AMOLED 2X
ขนาดหน้าจอ	7.60 **นิ้ว**
ความละเอียด	1812 x 2716 pixels
ระบบปฏิบัติการ	Android 12
ชิปประมวลผล	Qualcomm Snapdragon 8+ Gen 1 SM8475 3.19 GHz
ชิปกราฟิก	Adreno 730
หน่วยความจำ	12 GB
ความจุ	256 GB
Memory Card	**ไม่รองรับ**
กล้องหลัก	**ตัวที่** 1: 50 MP, f/1.8, 24mm (wide), 1.0µm, Dual Pixel PDAF, OIS **ตัวที่** 2: 10 MP, f/2.4, 67mm (telephoto), 1.0µm, PDAF, OIS, 3x optical zoom **ตัวที่** 3: 12 MP, f/2.2, 123˚, 12mm (ultrawide), 1.12µm
ความละเอียดวีดีโอ	4K@60fps, 1080p@60/240fps (gyro-EIS), 720p@960fps (gyro-EIS), HDR10+
กล้องหน้า	**ตัวที่** 1: 4 MP, f/1.8, 26mm (wide), 2.0µm, under display
Bluetooth	5.2, A2DP, LE, aptX HD
Wi-Fi	802.11 a/b/g/n/ac/6e, dua
USB	Type-C
GPS	A-GPS, GLONASS, GALILEO,
NFC	**รอบรับ**
[illegible]	4,400 mAh

```
# 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") %>%
    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..."
         attribute
1          วันเปิดตัว
2      วันวางจำหน่าย
3              ขนาด
4             น้ำหนัก
5               วัสดุ
6               SIM
7       Technology
8                2G
```

```
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")
```