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,desc"
imdb <- read_html(url)
```

```
#movies title
titles <- imdb %>%
    html_nodes("h3.lister-item-header") %>%
    html_text2()
```

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

```
#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, number_of_vote = num_votes)
head(df)
```

A data.frame: 6 × 3

	title	rating	number_of_vote
	<chr>	<dbl>	<chr>
1	1. The Shawshank Redemption (1994)	9.3	Votes: 2,705,242 | Gross: $28.34M | Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,878,458 | Gross: $134.97M | Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,678,896 | Gross: $534.86M | Top 250: #3
4	4. The Godfather Part II (1974)	9.0	Votes: 1,282,919 | Gross: $57.30M | Top 250: #4
5	5. Schindler's List (1993)	9.0	Votes: 1,367,209 | Gross: $96.90M | Top 250: #6
6	6. 12 Angry Men (1957)	9.0	Votes: 799,166 | Gross: $4.36M | Top 250: #5

Mini Project 02 - Specphone phone database

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

```
url <- read_html("https://specphone.com/Samsung-Galaxy-A04.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: 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
ss_url <- read_html("https://specphone.com/brand/Samsung")
```

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

```
#http://specphone.com/link..
full_link <- paste0("http://specphone.com", links[1:5])
```

```
result <- data.frame()

for (link in full_link[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("Progess...")

}

print(result)
```

```
[1] "Progess..."
[1] "Progess..."
[1] "Progess..."
[1] "Progess..."
[1] "Progess..."
         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

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