Mini Project 01 - IMDB we scraping

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

```
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() # covert factor to numeric
```

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

```
# build a dataset
df <- data.frame(
    Title = titles,
    Rating = ratings,
    Vote = num_votes
)
head(df)
```

A data.frame: 6 × 3

	Title	Rating	Vote
	<chr>	<dbl>	<chr>
1	1. The Shawshank Redemption (1994)	9.3	Votes: 2,666,134 | Gross: $28.34M | Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,847,559 | Gross: $134.97M | Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,639,082 | Gross: $534.86M | Top 250: #3
4	4. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,837,942 | Gross: $377.85M | Top 250: #7
5	5. Schindler's List (1993)	9.0	Votes: 1,349,956 | Gross: $96.90M | Top 250: #6
6	6. The Godfather Part II (1974)	9.0	Votes: 1,265,327 | Gross: $57.30M | Top 250: #4

Mini Project 02 - Specphone Phone Database

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

```
url <- read_html("https://specphone.com/Samsung-Galaxy-Note-20-Ultra-5G.html")
```

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

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

```
data.frame(Attribute = att,
           Value = value)
```

A data.frame: 34 × 2

Attribute	Value
<chr>	<chr>
วันเปิดตัว	**สิงหาคม** 2563
วันวางจำหน่าย	**สิงหาคม** 2563, **วางจำหน่ายแล้ว**
ขนาด	164.80 x 77.20 x 8.10 **มม.**
น้ำหนัก	**ไม่รองรับ**
วัสดุ	Gorilla Glass 7, aluminum frame
SIM	**รองรับ** 2 **ซิมการ์ด** (nano sim, nano sim)
Technology	5G,LTE,HSPA+
2G	850/900/1800/1900
3G	850/900/1900/2100
4G	850/900/1900/2100/2600
5G	850/900/1900/2100/2600
ความเร็ว	5G,LTE,HSPA+
ประเภท	Dynamic AMOLED 2X Infinity-O Display
ขนาดหน้าจอ	6.90 **นิ้ว**
ความละเอียด	3000 x 1440 pixels
ฟีเจอร์เพิ่มเติม	**หน้าจอ** refresh rate 120Hz , HDR10+ **รองรับ** Samsung DeX **จอกันรอย** Gorilla Glass 7 **กันน้ำกันฝุ่น มาตรฐาน** IP68
ระบบปฏิบัติการ	Android 10
ชิปประมวลผล	Samsung Exynos 990 2.73 GHz
ชิปกราฟิก	Mali G77 MP11
หน่วยความจำ	12 GB
ความจุ	256/512 GB
Memory Card	3.2, Type- (256)
กล้องหลัก	**ตัวที่** 1: 108 MP, f/1.8, 26mm (wide), 1/1.33 **ตัวที่** 2: 12 MP, f/3.0, 103mm (periscope telephoto), 1.0µm, PDAF, OIS, 5x optical zoom, 50x hybrid zoom **ตัวที่** 3: 12 MP, f/2.2, 13mm (ultrawide), 1/2.55
ความละเอียดวีดีโอ	4K@30/60fps, 1080p@30fps
กล้องหน้า	**ตัวที่** 1: 10 MP, f/2.2, 26mm (wide), 1/3.2
Bluetooth	5.0, A2DP, LE, aptX
Wi-Fi	802.11 a/b/g/n/ac/6
USB	addnew
GPS	GPS, Galileo, Glonass, Be
NFC	**รอบรับ**

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

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

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

```
# Dataframe
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(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")
```