Mini Project 01 - IMDB web scraping

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

```
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. Schindler\'s List (1993)' · '5. The Godfather Part II (1974)' · '6. 12 Angry Men (1957)' · '7. The Lord of the Rings: The Return of the King (2003)' · '8. Pulp Fiction (1994)' · '9. Inception (2010)' · '10. Fight Club (1999)'

```
# 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: 2,690,554 | Gross: $28.34M | Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,866,345 | Gross: $134.97M | Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,664,237 | Gross: $534.86M | Top 250: #3
4	4. Schindler's List (1993)	9.0	Votes: 1,360,761 | Gross: $96.90M | Top 250: #6
5	5. The Godfather Part II (1974)	9.0	Votes: 1,276,541 | Gross: $57.30M | Top 250: #4
6	6. 12 Angry Men (1957)	9.0	Votes: 794,876 | Gross: $4.36M | Top 250: #5

Mini Project 02 - Specphone Phone Database

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

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

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

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

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

A data.frame: 31 × 2

attributes	value
<chr>	<chr>
วันเปิดตัว	**วันเปิดตัว**
วันวางจำหน่าย	**วันวางจำหน่าย**
ขนาด	**ขนาด**
น้ำหนัก	**น้ำหนัก**
วัสดุ	**วัสดุ**
SIM	SIM
Technology	Technology
2G	2G
3G	3G
4G	4G
5G	5G
ความเร็ว	**ความเร็ว**
ประเภท	**ประเภท**
ขนาดหน้าจอ	**ขนาดหน้าจอ**
ความละเอียด	**ความละเอียด**
ระบบปฏิบัติการ	**ระบบปฏิบัติการ**
ชิปประมวลผล	**ชิปประมวลผล**
ชิปกราฟิก	**ชิปกราฟิก**
หน่วยความจำ	**หน่วยความจำ**
ความจุ	**ความจุ**
Memory Card	Memory Card
กล้องหลัก	**กล้องหลัก**
ความละเอียดวีดีโอ	**ความละเอียดวีดีโอ**
กล้องหน้า	**กล้องหน้า**
Bluetooth	Bluetooth
Wi-Fi	Wi-Fi
USB	USB
GPS	GPS
NFC	NFC
ความจุ	**ความจุ**
ประเภท	**ประเภท**

```
# 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("Progess ...")
}

print(result)
```

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

```
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
# After run, Click Attached data (Left icon) >> Export data >> Excel >> data >>
write_csv(result, "result_ss_phone.csv")
```