Mini Project 01 - IMDB web scraping

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

```
url <- "https://www.imdb.com/search/title/?groups=top_100&sort=user_rating%2Cdesc
```

```
print(url)
```

```
[1] "https://www.imdb.com/search/title/?groups=top_100&sort=user_rating%2Cdesc"
```

```
## 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() # text2 : drop special characters
```

```
titles[1:10]
```

'1. The Shawshank Redemption (1994)' · '2. The Godfather (1972)' · '3. The Dark Knight (2008)' ·

```
## 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()
```

```
num_votes[1:10]
```

'Votes: 2,657,665 | Gross: $28.34M | Top 250: #1' · 'Votes: 1,841,944 | Gross: $134.97M | Top 250: #2' ·

```
## build a dataset
df <- data.frame(
    title = titles,
    rating = ratings,
    num_vote = num_votes
)
```

```
head(df)
```

A data.frame: 6 × 3

Mini Project 02 - SpecPhone Phone Database

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

```
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

```
## All sumsung smartphones
sumsung_url <- read_html("https://specphone.com/brand/Samsung")
```

```
## Links to all sumsung smartphones
links <- sumsung_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[1:5]){
    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("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")
```