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,des
```

```
print(url)
```

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

```
imdb <- read_html(url)
```

```
imdb
```

```
{html_document}
<html xmlns:og="http://ogp.me/ns#" xmlns:fb="http://www.facebook.com/2008/fb
[1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-
[2] <body id="styleguide-v2" class="fixed">\n            <img height="1" wid
```

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

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

```
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: 2,704,069 | Gross: $28.34M | Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,877,558 | Gross: $134.97M | Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,677,877 | Gross: $534.86M | Top 250: #3
4	4. The Godfather Part II (1974)	9.0	Votes: 1,282,451 | Gross: $57.30M | Top 250: #4
5	5. Schindler's List (1993)	9.0	Votes: 1,366,741 | Gross: $96.90M | Top 250: #6
6	6. 12 Angry Men (1957)	9.0	Votes: 798,852 | Gross: $4.36M | Top 250: #5

Mini Project 02 - Specphone Phone Database

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

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

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

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

```
df <- data.frame(attributee = att, value = value)
```

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

```
links <- samsung_url %>%
    html_nodes("li.mobile-brand-item a") %>%
    html_attr("href")
```

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

```
result <- data.frame()

for (link in fullLinks[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(result)
```

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