Mini project 01 web scraping

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

```
Warning message in system("timedatectl", intern = TRUE):
“running command 'timedatectl' had status 1”
Warning message:
“Failed to locate timezone database”
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 

✓ ggplot2 3.3.5     ✓ purrr   0.3.4
✓ tibble  3.1.5     ✓ dplyr   1.0.7
✓ tidyr   1.1.4     ✓ stringr 1.4.0
✓ readr   2.0.2     ✓ forcats 0.5.1

── Conflicts ────────────────────────────────────────── tidyverse_conflicts() 
✖ dplyr::filter()  masks stats::filter()
✖ purrr::flatten() masks jsonlite::flatten()
✖ dplyr::lag()     masks stats::lag()


Attaching package: ‘rvest’
```

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

```
print(url)
```

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

```
#read html
imdb <- read_html(url)
```

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

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

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

```
rating[1:10]
```

9.3 · 9.2 · 9 · 9 · 9 · 9 · 9 · 8.9 · 8.8 · 8.8

```

```

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

head(df)
```

A data.frame: 6 × 3

	titles	ratings	num_votes
	<chr>	<dbl>	<chr>
1	1. The Shawshank Redemption (1994)	9.3	Votes: 2,658,802 | Gross: $28.34M | Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,842,619 | Gross: $134.97M | Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,631,601 | Gross: $534.86M | Top 250: #3
4	4. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,833,269 | Gross: $377.85M | Top 250: #7
5	5. Schindler's List (1993)	9.0	Votes: 1,346,705 | Gross: $96.90M | Top 250: #6
6	6. The Godfather Part II (1974)	9.0	Votes: 1,262,382 | Gross: $57.30M | Top 250: #4

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()
```

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

head(df)
```

A data.frame: 6 × 2

	atts	vaues
	<chr>	<chr>
1	วันเปิดตัว	ตุลาคม 2565
2	วันวางจำหน่าย	ยังไม่วางจำหน่าย
3	ขนาด	164.40 x 76.30 x 9.10 มม.
4	น้ำหนัก	192 กรัม
5	วัสดุ	Glass front, plastic back, plastic frame
6	SIM	รองรับ 2 ซิมการ์ด (nano sim, nano sim)

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

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

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

```
full_links[1:5]
```

'https://specphone.com/Samsung-Galaxy-M13.html' · 'https://specphone.com/Samsung-Galaxy-A23.html' · 'https://specphone.com/Samsung-Galaxy-A13.html' · 'https://specphone.com/Samsung-Galaxy-M32-5G.html' · 'https://specphone.com/Samsung-Galaxy-A12-Nacho.html'

```
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..")
}

#print (result)
```

```
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")
```