Mini Project 01 - IMDB web scraping

```
library(tidyverse) #prep data
library(rvest) #scrape data
```

```
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%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
#html_node top1, html_nodes all
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()
```

```
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 from 3 vectors
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: 2,658,071 | Gross: $28.34M | Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,842,229 | Gross: $134.97M | Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,630,904 | Gross: $534.86M | Top 250: #3
4	4. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,832,726 | Gross: $377.85M | Top 250: #7
5	5. Schindler's List (1993)	9.0	Votes: 1,346,381 | Gross: $96.90M | Top 250: #6
6	6. The Godfather Part II (1974)	9.0	Votes: 1,262,113 | Gross: $57.30M | Top 250: #4

```
df
```

A data.frame: 50 × 3

<table>
<tr><th>title</th><th>rating</th><th>num_vote</th></tr>
<tr><td>&lt;chr&gt;</td><td>&lt;dbl&gt;</td><td>&lt;chr&gt;</td></tr>
<tr><td>1. The Shawshank Redemption (1994)</td><td>9.3</td><td>Votes: 2,658,071 | Gross: $28.34M | Top 250: #1</td></tr>
<tr><td>2. The Godfather (1972)</td><td>9.2</td><td>Votes: 1,842,229 | Gross: $134.97M | Top 250: #2</td></tr>
<tr><td>3. The Dark Knight (2008)</td><td>9.0</td><td>Votes: 2,630,904 | Gross: $534.86M | Top 250: #3</td></tr>
<tr><td>4. The Lord of the Rings: The Return of the King (2003)</td><td>9.0</td><td>Votes: 1,832,726 | Gross: $377.85M | Top 250: #7</td></tr>
<tr><td>5. Schindler's List (1993)</td><td>9.0</td><td>Votes: 1,346,381 | Gross: $96.90M | Top 250: #6</td></tr>
<tr><td>6. The Godfather Part II (1974)</td><td>9.0</td><td>Votes: 1,262,113 | Gross: $57.30M | Top 250: #4</td></tr>
</table>

Mini Project 02 - Specphone Phone Database

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

```
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()
```

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

```
#Links to all samsung smartphone
links <- samsung_url %>%
            html_nodes("li.mobile-brand-item a") %>% #spacebar_a find_child a ir
            html_attr("href") #hyperlink reference
```

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

```
#Samsung 10 smartphone
result <- data.frame()
for (links in full_links[1:10]){
    ss_topic <- links %>%
        read_html() %>%
        html_nodes("div.topic") %>%
        html_text2()

    ss_detail <- links %>%
        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 ..."
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
         attribute
1          วันเปิดตัว
2      วันวางจำหน่าย
3             ขนาด
```

```
4            น้ำหนัก
5              วัสดุ
6               SIM
7        Technology
8                2G
```

```
result
```

A data.frame: 319 × 2

attribute	value
<chr>	<chr>
วันเปิดตัว	มิถุนายน 2565
วันวางจำหน่าย	ยังไม่วางจำหน่าย
ขนาด	165.40 x 76.90 x 8.40 มม.
น้ำหนัก	192 กรัม
วัสดุ	Glass front, plastic back, plastic frame
SIM	รองรับ 2 ซิมการ์ด (nano sim, nano sim)
Technology	HSPA 42.2/5.76 Mbps, LTE-A
2G	850/900/1800/1900
3G	850/900/1900/2100
4G	850/900/1900/2100
5G	-
ความเร็ว	HSPA 42.2/5.76 Mbps, LTE-A
ประเภท	PLS LCD
ขนาดหน้าจอ	6.60 นิ้ว
ความละเอียด	1080 x 2408 pixels
ระบบปฏิบัติการ	Android 12
ชิปประมวลผล	Samsung Exynos 850 S5E3830 2 GHz
ชิปกราฟิก	Mali-G52 MP1
หน่วยความจำ	4 GB
ความจุ	64 GB
Memory Card	microSD (1)
กล้องหลัก	ตัวที่ 1: 50 MP, f/1.8, (wide), PDAF ตัวที่ 2: 5 MP, f/2.2, 123˚ (ultrawide) ตัวที่ 3: 2 MP, f/2.4, (depth)
ความละเอียดวีดีโอ	1080p@30fps
กล้องหน้า	ตัวที่ 1: 8 MP, f/2.2, (wide)
Bluetooth	5.0, A2DP, LE

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