Mini Project 01 - IMDB web scraping

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

```
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

title	rating	num_vote
<chr>	<dbl>	<chr>
1. The Shawshank Redemption (1994)	9.3	Votes: 2,658,071 | Gross: $28.34M | Top 250: #1
2. The Godfather (1972)	9.2	Votes: 1,842,229 | Gross: $134.97M | Top 250: #2
3. The Dark Knight (2008)	9.0	Votes: 2,630,904 | Gross: $534.86M | Top 250: #3
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. Schindler's List (1993)	9.0	Votes: 1,346,381 | Gross: $96.90M | Top 250: #6
6. The Godfather Part II (1974)	9.0	Votes: 1,262,113 | Gross: $57.30M | Top 250: #4
7. 12 Angry Men (1957)	9.0	Votes: 784,825 | Gross: $4.36M | Top 250: #5
8. Pulp Fiction (1994)	8.9	Votes: 2,034,486 | Gross: $107.93M | Top 250: #8
9. Inception (2010)	8.8	Votes: 2,331,297 | Gross: $292.58M | Top 250: #14
10. The Lord of the Rings: The Two Towers (2002)	8.8	Votes: 1,654,928 | Gross: $342.55M | Top 250: #13
11. Fight Club (1999)	8.8	Votes: 2,103,207 | Gross: $37.03M | Top 250: #12
12. The Lord of the Rings: The Fellowship of the Ring (2001)	8.8	Votes: 1,861,481 | Gross: $315.54M | Top 250: #9
13. Forrest Gump (1994)	8.8	Votes: 2,059,470 | Gross: $330.25M | Top 250: #11
14. Il buono, il brutto, il cattivo (1966)	8.8	Votes: 757,886 | Gross: $6.10M | Top 250: #10
15. The Matrix (1999)	8.7	Votes: 1,899,775 | Gross: $171.48M | Top 250: #16
16. Goodfellas (1990)	8.7	Votes: 1,151,713 | Gross: $46.84M | Top 250: #17
17. The Empire Strikes Back (1980)	8.7	Votes: 1,283,982 | Gross: $290.48M | Top 250: #15
18. One Flew Over the Cuckoo's Nest (1975)	8.7	Votes: 1,002,912 | Gross: $112.00M | Top 250: #18
19. Interstellar (2014)	8.6	Votes: 1,803,351 | Gross: $188.02M | Top 250: #26

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 in
            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
9               3G
```

```
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
Wi-Fi	802.11 a/b/g/n/ac, dual-b
USB	Type-C
GPS	A-GPS, GLONASS, GALILEO,
NFC	รอบรับ
ความจุ	5,000 mAh
⋮	⋮
ขนาด	121.40 x 62.90 x 10.70 มม.
น้ำหนัก	123 กรัม
วัสดุ	Plastic
SIM	รองรับ 2 ซิมการ์ด (micro sim, micro sim)

Technology	EDGE, HSPA
2G	850/900/1800/1900
3G	900/2100
4G	-
5G	-
ความเร็ว	EDGE, HSPA
ประเภท	TFT LCD
ขนาดหน้าจอ	4.00 นิ้ว

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