Mini Project 01 - IMDB web scraping

```
# Install the rvest package (run this only once)
#install.packages("rvest")

# Load the rvest package
library(rvest)
library(tidyverse)
```

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

```
print(url)
```

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

```
imbd <- read_html(url)
```

```
imbd
```

```
{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 <- imbd %>%
    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. Schindler\'s List (1993)' · '5. The Lord of the Rings: The Return of the King (2003)' ·
'6. The Godfather Part II (1974)' · '7. 12 Angry Men (1957)' · '8. Pulp Fiction (1994)' ·
'9. The Lord of the Rings: The Fellowship of the Ring (2001)' · '10. Inception (2010)'

```
# rating
imbd %>%
    html_node("div.ratings-imdb-rating") %>%
    html_text2()
```

'9.3'

```
# hyml_nodes & convert character > numeric
ratings <- imbd %>%
    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 <- imbd %>%
    html_nodes("p.sort-num_votes-visible") %>%
    html_text2()
```

```
num_votes
```

'Votes: 2,709,540 | Gross: $28.34M | Top 250: #1' · 'Votes: 1,881,633 | Gross: $134.97M | Top 250: #2' ·
'Votes: 2,682,563 | Gross: $534.86M | Top 250: #3' · 'Votes: 1,368,963 | Gross: $96.90M | Top 250: #6' ·
'Votes: 1,864,829 | Gross: $377.85M | Top 250: #7' · 'Votes: 1,284,657 | Gross: $57.30M | Top 250: #4' ·
'Votes: 800,472 | Gross: $4.36M | Top 250: #5' · 'Votes: 2,080,186 | Gross: $107.93M | Top 250: #8' ·
'Votes: 1,894,349 | Gross: $315.54M | Top 250: #9' · 'Votes: 2,380,514 | Gross: $292.58M | Top 250: #14' ·
'Votes: 2,153,568 | Gross: $37.03M | Top 250: #12' · 'Votes: 2,105,279 | Gross: $330.25M | Top 250: #11' ·
'Votes: 1,683,816 | Gross: $342.55M | Top 250: #13' · 'Votes: 769,051 | Gross: $6.10M | Top 250: #10' ·
'Votes: 1,175,203 | Gross: $46.84M | Top 250: #17' · 'Votes: 1,932,933 | Gross: $171.48M | Top 250: #16' ·
'Votes: 1,016,904 | Gross: $112.00M | Top 250: #18' · 'Votes: 1,305,142 | Gross: $290.48M | Top 250: #15' ·
'Votes: 1,868,222 | Gross: $188.02M | Top 250: #25' · 'Votes: 1,448,533 | Gross: $130.74M | Top 250: #22' ·
'Votes: 1,673,048 | Gross: $100.13M | Top 250: #19' · 'Votes: 1,317,028 | Gross: $136.80M | Top 250: #27' ·
'Votes: 1,406,076 | Gross: $216.54M | Top 250: #23' · 'Votes: 1,377,544 | Gross: $322.74M | Top 250: #28' ·
'Votes: 1,110,442 | Gross: $204.84M | Top 250: #29' · 'Votes: 775,519 | Gross: $10.06M | Top 250: #31' ·
'Votes: 763,482 | Gross: $7.56M | Top 250: #24' · 'Votes: 703,125 | Gross: $57.60M | Top 250: #26' ·
'Votes: 467,548 | Top 250: #21' · 'Votes: 349,344 | Gross: $0.27M | Top 250: #20' · 'Votes: 59,270 | Top 250: #45' ·
'Votes: 882,901 | Gross: $13.09M | Top 250: #42' · 'Votes: 827,451 | Gross: $53.37M | Top 250: #34' ·
'Votes: 1,220,643 | Gross: $210.61M | Top 250: #30' · 'Votes: 1,516,400 | Gross: $187.71M | Top 250: #37' ·
'Votes: 1,339,150 | Gross: $132.38M | Top 250: #39' · 'Votes: 1,347,770 | Gross: $53.09M | Top 250: #41' ·
'Votes: 674,935 | Gross: $83.47M | Top 250: #53' · 'Votes: 1,174,477 | Gross: $19.50M | Top 250: #35' ·
'Votes: 892,556 | Gross: $78.90M | Top 250: #51' · 'Votes: 1,094,638 | Gross: $23.34M | Top 250: #40' ·
'Votes: 1,070,785 | Gross: $422.78M | Top 250: #36' · 'Votes: 1,132,495 | Gross: $6.72M | Top 250: #38' ·
'Votes: 843,212 | Gross: $32.57M | Top 250: #32' · 'Votes: 869,791 | Gross: $13.18M | Top 250: #46' ·
'Votes: 333,652 | Gross: $5.32M | Top 250: #48' · 'Votes: 577,398 | Gross: $1.02M | Top 250: #43' ·
'Votes: 679,129 | Gross: $32.00M | Top 250: #33' · 'Votes: 282,260 | Top 250: #44' ·
'Votes: 264,830 | Gross: $11.99M | Top 250: #50'

```
# 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: 2,709,540 | Gross: $28.34M | Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,881,633 | Gross: $134.97M | Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,682,563 | Gross: $534.86M | Top 250: #3
4	4. Schindler's List (1993)	9.0	Votes: 1,368,963 | Gross: $96.90M | Top 250: #6
5	5. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,864,829 | Gross: $377.85M | Top 250: #7
6	6. The Godfather Part II (1974)	9.0	Votes: 1,284,657 | Gross: $57.30M | Top 250: #4

Mini Project 02 - Specphone Phone Database

```
library(rvest)
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

attribute	value
<chr>	<chr>
วันเปิดตัว	**ตุลาคม** 2565
วันวางจำหน่าย	**ยังไม่วางจำหน่าย**
ขนาด	164.40 x 76.30 x 9.10 **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	850/900/1900/2100/2600
5G	-
ความเร็ว	HSPA 42.2/5.76 Mbps, LTE-A
ประเภท	PLS LCD
ขนาดหน้าจอ	6.50 **นิ้ว**
ความละเอียด	720 x 1600 pixels
ระบบปฏิบัติการ	Android 12
ชิปประมวลผล	Spreadtrum Unisoc SC9863A 1.6 GHz
ชิปกราฟิก	PowerVR GE8322
หน่วยความจำ	3 GB
ความจุ	32 GB
Memory Card	microSD (1)
กล้องหลัก	**ตัวที่** 1: 50 MP, f/1.8, (wide), AF **ตัวที่** 2: 2 MP, f/2.4, (depth)
ความละเอียดวีดีโอ	1080p@30fps
กล้องหน้า	**ตัวที่** 1: 5 MP, f/2.2
Bluetooth	5.0, A2DP, LE
Wi-Fi	802.11 a/b/g/n/ac, dual-b
USB	Type-C
GPS	GLONASS, GALILEO, BDS
NFC	**ไม่รองรับ**
ความจุ	5,000 mAh
ประเภท	Non-removable Li-Po Batt

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

```
# links to all sumsung smartphone
links <- sumsung_url %>%
    html_nodes("li.mobile-brand-item a") %>% # find li and then find a
    html_attr("href")
```

```
# str_c() or paste0()
full_links <- paste0("http://specphone.com", links)
```

```
result <- data.frame()

for (link in full_links[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("Progress ...")
}

print(result)
```

```
[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
```

```
# write csv
write_csv(result, "result_ss_phone_csv")
```