

Mini Project01 - IMDB Web Scraping

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

```
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"
```

```
# 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" width .
```

```
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

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

```
num_votes
```

'Votes: 2,670,558 | Gross: \$28.34M | Top 250: #1' · 'Votes: 1,850,797 | Gross: \$134.97M | Top 250: #2' ·
 'Votes: 2,643,347 | Gross: \$534.86M | Top 250: #3' · 'Votes: 1,840,604 | Gross: \$377.85M | Top 250: #7' ·
 'Votes: 1,352,020 | Gross: \$96.90M | Top 250: #6' · 'Votes: 1,267,354 | Gross: \$57.30M | Top 250: #4' ·
 'Votes: 788,758 | Gross: \$4.36M | Top 250: #5' · 'Votes: 2,045,475 | Gross: \$107.93M | Top 250: #8' ·
 'Votes: 2,342,888 | Gross: \$292.58M | Top 250: #14' · 'Votes: 1,662,022 | Gross: \$342.55M | Top 250: #13' ·
 'Votes: 2,115,986 | Gross: \$37.03M | Top 250: #12' · 'Votes: 1,869,842 | Gross: \$315.54M | Top 250: #9' ·
 'Votes: 2,070,632 | Gross: \$330.25M | Top 250: #11' · 'Votes: 760,837 | Gross: \$6.10M | Top 250: #10' ·
 'Votes: 1,907,634 | Gross: \$171.48M | Top 250: #16' · 'Votes: 1,157,796 | Gross: \$46.84M | Top 250: #17' ·
 'Votes: 1,289,371 | Gross: \$290.48M | Top 250: #15' · 'Votes: 1,006,539 | Gross: \$112.00M | Top 250: #18' ·
 'Votes: 1,818,973 | Gross: \$188.02M | Top 250: #26' · 'Votes: 756,321 | Gross: \$7.56M | Top 250: #23' ·
 'Votes: 761,574 | Gross: \$10.06M | Top 250: #31' · 'Votes: 1,387,897 | Gross: \$216.54M | Top 250: #24' ·
 'Votes: 1,297,983 | Gross: \$136.80M | Top 250: #27' · 'Votes: 694,178 | Gross: \$57.60M | Top 250: #25' ·
 'Votes: 1,646,557 | Gross: \$100.13M | Top 250: #19' · 'Votes: 1,096,957 | Gross: \$204.84M | Top 250: #29' ·
 'Votes: 1,428,844 | Gross: \$130.74M | Top 250: #22' · 'Votes: 1,361,944 | Gross: \$322.74M | Top 250: #28' ·
 'Votes: 57,600 | Top 250: #44' · 'Votes: 346,343 | Gross: \$0.27M | Top 250: #20' · 'Votes: 456,513 | Top 250: #21' ·
 'Votes: 797,262 | Gross: \$53.37M | Top 250: #34' · 'Votes: 856,807 | Gross: \$13.09M | Top 250: #42' ·
 'Votes: 856,661 | Gross: \$13.18M | Top 250: #46' · 'Votes: 1,329,834 | Gross: \$53.09M | Top 250: #41' ·
 'Votes: 1,322,186 | Gross: \$132.38M | Top 250: #39' · 'Votes: 830,587 | Gross: \$32.57M | Top 250: #33' ·
 'Votes: 1,496,422 | Gross: \$187.71M | Top 250: #37' · 'Votes: 1,120,970 | Gross: \$6.72M | Top 250: #38' ·
 'Votes: 1,083,881 | Gross: \$23.34M | Top 250: #40' · 'Votes: 1,158,563 | Gross: \$19.50M | Top 250: #35' ·
 'Votes: 1,055,794 | Gross: \$422.78M | Top 250: #36' · 'Votes: 261,720 | Gross: \$11.99M | Top 250: #50' ·
 'Votes: 277,760 | Top 250: #45' · 'Votes: 1,201,764 | Gross: \$210.61M | Top 250: #30' ·
 'Votes: 667,291 | Gross: \$83.47M | Top 250: #53' · 'Votes: 881,477 | Gross: \$78.90M | Top 250: #51' ·
 'Votes: 330,060 | Gross: \$5.32M | Top 250: #48' · 'Votes: 671,768 | Gross: \$32.00M | Top 250: #32' ·
 'Votes: 492,076 | Gross: \$36.76M | Top 250: #49'

```
# 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: 2,670,558 Gross: \$28.34M Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,850,797 Gross: \$134.97M Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,643,347 Gross: \$534.86M Top 250: #3
4	4. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,840,604 Gross: \$377.85M Top 250: #7
5	5. Schindler's List (1993)	9.0	Votes: 1,352,020 Gross: \$96.90M Top 250: #6
6	6. The Godfather Part II (1974)	9.0	Votes: 1,267,354 Gross: \$57.30M Top 250: #4

Mini Project02 - SpecPhone 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()
```

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

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

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

```
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)
```

```
[1] "Progress..."
[1] "Progress..."
[1] "Progress..."
[1] "Progress..."
[1] "Progress..."
```

```
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")
```