

# Mini Project 01 - Web Scraping

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

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

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

```
# movie title
# text2 = remove special character
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
```

```
# votes  
num_votes <- imdb %>%  
  html_node("p.sort-num_votes-visible") %>%  
  html_text2()
```

```
num_votes
```

```
'Votes: 2,657,665 | Gross: $28.34M | Top 250: #1'
```

```
# build a dataset  
df <- data.frame(  
  titles = titles,  
  ratings = ratings,  
  num_votes = num_votes  
)
```

```
head(df)
```

A data.frame: 6 × 3

	titles	ratings	num_votes
	<chr>	<dbl>	<chr>
1	1. The Shawshank Redemption (1994)	9.3	Votes: 2,657,665   Gross: \$28.34M   Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 2,657,665   Gross: \$28.34M   Top 250: #1
3	3. The Dark Knight (2008)	9.0	Votes: 2,657,665   Gross: \$28.34M   Top 250: #1
4	4. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 2,657,665   Gross: \$28.34M   Top 250: #1
5	5. Schindler's List (1993)	9.0	Votes: 2,657,665   Gross: \$28.34M   Top 250: #1
6	6. The Godfather Part II (1974)	9.0	Votes: 2,657,665   Gross: \$28.34M   Top 250: #1

# Mini Project 02 - Specphone Phone Database

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

```
# SS galaxy A04
url <- read_html("https://specphone.com/Samsung-Galaxy-A04.html")
```

```
att <- url %>%
  html_nodes("div.topic") %>%
  html_text2()

detail <- url %>%
  html_nodes("div.detail") %>%
  html_text2()
```

```
data.frame(attributes = att, value = detail)
```

A data.frame: 31 × 2

attributes	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 smartphones
```

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

```
# Link to all samsung smartphones
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)
```

```
[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
10     4G
11     5G
12     ความเร็ว
13     ประเภท
```

```
# write csv
write_csv(result, "result_ss_phone.csv")
```