

# Building Data Portfolio

- Intro to Web Scraping
- Date : 12 November 2022
- Time : 13.00 - 15.00

## Mini Project 01 - 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 .
```

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

```
num_votes[1:10]
```

'Votes: 2,660,041 | Gross: \$28.34M | Top 250: #1' · 'Votes: 1,843,472 | Gross: \$134.97M | Top 250: #2' ·  
 'Votes: 2,632,937 | Gross: \$534.86M | Top 250: #3' · 'Votes: 1,834,138 | Gross: \$377.85M | Top 250: #7' ·  
 'Votes: 1,347,286 | Gross: \$96.90M | Top 250: #6' · 'Votes: 1,262,901 | Gross: \$57.30M | Top 250: #4' ·  
 'Votes: 785,454 | Gross: \$4.36M | Top 250: #5' · 'Votes: 2,036,037 | Gross: \$107.93M | Top 250: #8' ·  
 'Votes: 2,333,090 | Gross: \$292.58M | Top 250: #14' · 'Votes: 1,656,151 | Gross: \$342.55M | Top 250: #13'

```
# build a dataest
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: 2,660,041   Gross: \$28.34M   Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,843,472   Gross: \$134.97M   Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,632,937   Gross: \$534.86M   Top 250: #3
4	4. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,834,138   Gross: \$377.85M   Top 250: #7
5	5. Schindler's List (1993)	9.0	Votes: 1,347,286   Gross: \$96.90M   Top 250: #6
6	6. The Godfather Part II (1974)	9.0	Votes: 1,262,901   Gross: \$57.30M   Top 250: #4

## Mini Project 02 - Specphone Phone Database

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

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

```
links <- samsung_url %>%
```

```
html_nodes("li.mobile-brand-item a") %>%  
html_attr("href")
```

```
links[1:10]
```

```
 '/Samsung-Galaxy-M13.html' · '/Samsung-Galaxy-A23.html' · '/Samsung-Galaxy-A13.html' ·  
 '/Samsung-Galaxy-M32-5G.html' · '/Samsung-Galaxy-A12-Nacho.html' · '/Samsung-Galaxy-Pocket-Neo.html' ·  
 '/Samsung-Galaxy-Young.html' · '/Samsung-Galaxy-J1-Mini.html' · '/Samsung-Galaxy-A01-Core-1-16GB.html' ·  
 '/Samsung-Galaxy-V-PLUS.html'
```

```
full_links <- paste0("https://specphone.com",links[1:10])
```

```
full_links
```

```
'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' ·  
'https://specphone.com/Samsung-Galaxy-Pocket-Neo.html' ·  
'https://specphone.com/Samsung-Galaxy-Young.html' · 'https://specphone.com/Samsung-Galaxy-J1-Mini.html' ·  
'https://specphone.com/Samsung-Galaxy-A01-Core-1-16GB.html' ·  
'https://specphone.com/Samsung-Galaxy-V-PLUS.html'
```

```
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(head(result))
```

	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")
```