

# Mini Project 01 - IMDB web scraping

```
##prepare data, build dataset, data vis
library(tidyverse)
##scrape data from internet
library(rvest)
```

```
#get url from website
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"
```

```
#access website and read Doc 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
##node
imdb %>%
  html_node("h3.lister-item-header") %>% #first node
  html_text2() #get text in node and remove special char
```

```
'1. The Shawshank Redemption (1994)'
```

```
#movie title
##nodes
titles <- imdb %>%
  html_nodes("h3.lister-item-header") %>% #every nodes
  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. The Godfather Part II (1974)' · '6. Schindler's List (1993)' ·
'7. 12 Angry Men (1957)' · '8. Pulp Fiction (1994)' · '9. The Lord of the Rings: The Fellowship of the Ring (2001)' ·
'10. Fight Club (1999)'
```

```
#ratings
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
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: 2,685,355   Gross: \$28.34M   Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,862,147   Gross: \$134.97M   Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,658,698   Gross: \$534.86M   Top 250: #3
4	4. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,850,227   Gross: \$377.85M   Top 250: #7
5	5. The Godfather Part II (1974)	9.0	Votes: 1,274,223   Gross: \$57.30M   Top 250: #4
6	6. Schindler's List (1993)	9.0	Votes: 1,358,441   Gross: \$96.90M   Top 250: #6

## Mini Project 02 - Specphone Phone Database

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

```
#get url from website
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 Smartphones
```

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

```
links <- samsung_url %>%
  html_nodes("li.mobile-brand-item a") %>% #find "a" within box
  html_attr("href") #get attribute 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	การเชื่อมต่อ

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

```
result2 <- data.frame()

for (link in full_links[1:3]) {
  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)

  result2 <- bind_rows(result2, tmp)

  print("progress...")
}

print(result2)
```

```
[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
14	ขนาดหน้าจอ
15	ความละเอียด

```
print(head(result2),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)