# IMDB Web Scraping

***Open library***

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

```
Warning message in system("timedatectl", intern = TRUE):
“running command 'timedatectl' had status 1”
Warning message:
“Failed to locate timezone database”
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

✓ ggplot2 3.3.5     ✓ purrr   0.3.4
✓ tibble  3.1.5     ✓ dplyr   1.0.7
✓ tidyr   1.1.4     ✓ stringr 1.4.0
✓ readr   2.0.2     ✓ forcats 0.5.1

── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter()  masks stats::filter()
✖ purrr::flatten() masks jsonlite::flatten()
✖ dplyr::lag()     masks stats::lag()


Attaching package: ‘rvest’
```

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

***Read html***

```
imdb <- read_html(url)
```

***Gather movie title from website***

```
titles <- imdb %>% html_nodes("h3.lister-item-header") %>%
html_text2()
```

***Gather movie rating from website***

```
rating <- imdb %>% html_nodes("div.inline-block.ratings-imdb-rating") %>%
html_text2() %>% as.numeric()
```

***Gather movie votes from website***

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

**Build a dataset**

```
df <- data.frame(title = titles, rating = rating, vote = votes)
```

```
head(df)
```

A data.frame: 6 × 3

| | title | rating | vote |
|---|---|---|---|
| | <chr> | <dbl> | <chr> |
| 1 | 1. The Shawshank Redemption (1994) | 9.3 | Votes: 2,658,430 \| Gross: $28.34M \| Top 250: #1 |
| 2 | 2. The Godfather (1972) | 9.2 | Votes: 1,842,424 \| Gross: $134.97M \| Top 250: #2 |
| 3 | 3. The Dark Knight (2008) | 9.0 | Votes: 2,631,256 \| Gross: $534.86M \| Top 250: #3 |
| 4 | 4. The Lord of the Rings: The Return of the King (2003) | 9.0 | Votes: 1,832,994 \| Gross: $377.85M \| Top 250: #7 |
| 5 | 5. Schindler's List (1993) | 9.0 | Votes: 1,346,543 \| Gross: $96.90M \| Top 250: #6 |
| 6 | 6. The Godfather Part II (1974) | 9.0 | Votes: 1,262,251 \| Gross: $57.30M \| Top 250: #4 |

# Specphone Web Scraping

**Open library**

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

***Read html***

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

***Gather cellphone attributes and details from website***

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

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

***Create data frame***

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

A data.frame: 31 × 2

| attributes | details |
|---|---|
| <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 |

# Gather data from all Samsung phones

```
sam_url <- read_html("https://specphone.com/brand/samsung")
```

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

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

***Create data frame***

```
result <- data.frame()

for(link in full_link[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(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")
```