# Mini Project 01 - IMDb Web Scraping

```
library(tidyverse)  # prep data
library(rvest)      # scrap data from the 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

✓ 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,des
```

```
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/fb
[1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-
[2] <body id="styleguide-v2" class="fixed">\n            <img height="1" wid
```

```
# movie title
titles <- imdb %>%
    html_nodes("h3.lister-item-header") %>%     # all nodes = all h3.liste.. t
    html_text2()                                # text2 = don't include specia
```

```
titles[1:10]        # vector
```

'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)'

```
# rating
ratings <- imdb %>%
    html_nodes("div.ratings-imdb-rating") %>%   # nodes = all div.ratings-imdb
    html_text2() %>%                            # text2 = don't include specia
    as.numeric()                                # convert to numeric
```

```
ratings[1:10]       # vector
```

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") %>%  # nodes = all p.sort-num_votes
    html_text2()                                # text2 = don't include specia
```

```
num_votes[1:10]     # vector
```

'Votes: 2,679,401 | Gross: $28.34M | Top 250: #1' · 'Votes: 1,857,555 | Gross: $134.97M | Top 250: #2' · 'Votes: 2,652,757 | Gross: $534.86M | Top 250: #3' · 'Votes: 1,846,497 | Gross: $377.85M | Top 250: #7' · 'Votes: 1,271,526 | Gross: $57.30M | Top 250: #4' · 'Votes: 1,355,946 | Gross: $96.90M | Top 250: #6' · 'Votes: 791,531 | Gross: $4.36M | Top 250: #5' · 'Votes: 2,054,820 | Gross: $107.93M | Top 250: #8' · 'Votes: 1,875,836 | Gross: $315.54M | Top 250: #9' · 'Votes: 2,124,945 | Gross: $37.03M | Top 250: #12'

```
# 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: 2,679,401 \| Gross: $28.34M \| Top 250: #1 |
| 2 | 2. The Godfather (1972) | 9.2 | Votes: 1,857,555 \| Gross: $134.97M \| Top 250: #2 |
| 3 | 3. The Dark Knight (2008) | 9.0 | Votes: 2,652,757 \| Gross: $534.86M \| Top 250: #3 |
| 4 | 4. The Lord of the Rings: The Return of the King (2003) | 9.0 | Votes: 1,846,497 \| Gross: $377.85M \| Top 250: #7 |
| 5 | 5. The Godfather Part II (1974) | 9.0 | Votes: 1,271,526 \| Gross: $57.30M \| Top 250: #4 |
| 6 | 6. Schindler's List (1993) | 9.0 | Votes: 1,355,946 \| Gross: $96.90M \| Top 250: #6 |

# Mini Project 02 - Specphone Phone Database

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

```
url <- read_html("https://specphone.com/ZTE-nubia-Red-Magic-8-Pro-.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: 32 × 2

| attribute | value |
|---|---|
| <chr> | <chr> |
| วันเปิดตัว | ธันวาคม 2565 |
| วันวางจำหน่าย | ยังไม่วางจำหน่าย |
| ขนาด | 164.00 x 76.40 x 8.90 มม. |
| น้ำหนัก | 230 กรัม |
| วัสดุ | Glass front, glass back, aluminum frame |
| SIM | รองรับ 2 ซิมการ์ด (nano sim, nano sim) |
| Technology | HSPA, LTE-A, 5G |
| 2G | 850/900/1800/1900 |
| 3G | 850/900/1900/2100 |
| 4G | 850/900/1900/2100/2600 |
| 5G | 2100/2600/3500/4700 |
| ความเร็ว | HSPA, LTE-A, 5G |
| ประเภท | AMOLED |
| ขนาดหน้าจอ | 6.80 นิ้ว |
| ความละเอียด | 1116 x 2480 pixels |
| ระบบปฏิบัติการ | Android 13 |
| ชิปประมวลผล | Qualcomm Snapdragon 8 Gen 2 SM8550 3.2 GHz |
| ชิปกราฟิก | Adreno 740 |
| หน่วยความจำ | 12 GB |
| ความจุ | 256 GB |
| Memory Card | ไม่รองรับ |
| กล้องหลัก | ตัวที่ 1: 50 MP, f/1.8, (wide), 1/1.57 ตัวที่ 2: 8 MP, f/2.2, 120˚, 13mm (ultrawide), 1/4.0 ตัวที่ 3: 2 MP, f/2.4, (macro) |
| ความละเอียดวีดีโอ | 8K@30fps, 4K@30/60fps, 1080p@30/60/120/240fps |
| กล้องหน้า | ตัวที่ 1: 16 MP, (wide), under display |
| Bluetooth | 5.3, A2DP, LE |
| Wi-Fi | 802.11 a/b/g/n/ac/6e, dua |
| USB | Type-C |
| GPS | GPS (L1+L5), GLONASS, BDS |
| NFC | รอบรับ |
| ความจุ | 5,000 mAh |
| ประเภท | Non-removable Li-Po Batt |
| Fast Charging | รองรับ (165W) |

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

```
# links to all samsung smartphones
links <- samsung_url %>%
    html_nodes("li.mobile-brand-item a") %>%    # a space = find a in li.mobil
    html_attr("href")                           # get href attribute
```

```
# paste0 = concatenates strings using no space
full_links <- paste0("https://specphone.com", links)
```

```
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(result)
```

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