MINI PROJECT 01 - IMDB WEB SCRAPING

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

```
Warning message in system("timedatectl", intern = TRUE):
“running command 'timedatectl' had status 1”
Warning message:
“Failed to locate timezone database”
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✓ dplyr     1.1.1     ✓ readr     2.1.4
✓ forcats   1.0.0     ✓ stringr   1.5.0
✓ ggplot2   3.4.2     ✓ tibble    3.2.1
✓ lubridate 1.9.2     ✓ tidyr     1.3.0
✓ purrr     1.0.1
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter()  masks stats::filter()
✖ purrr::flatten() masks jsonlite::flatten()
✖ dplyr::lag()     masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all confl

Attaching package: ‘rvest’


The following object is masked from ‘package:readr’:
```

```
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" widt ...
```

```
# 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. Schindler\'s List (1993)' · '5. The Godfather Part II (1974)' · '6. The Lord of the Rings: The Return of the King (2003)' ·
'7. 12 Angry Men (1957)' · '8. Pulp Fiction (1994)' · '9. Spider-Man: Across the Spider-Verse (2023)' ·
'10. Inception (2010)'

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

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

head(df)
```

A data.frame: 6 × 3

	title	rating	numvote
	<chr>	<dbl>	<chr>
1	1. The Shawshank Redemption (1994)	9.3	Votes: 2,789,245 | Gross: $28.34M | Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,942,563 | Gross: $134.97M | Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,769,147 | Gross: $534.86M | Top 250: #3
4	4. Schindler's List (1993)	9.0	Votes: 1,402,946 | Gross: $96.90M | Top 250: #6
5	5. The Godfather Part II (1974)	9.0	Votes: 1,319,868 | Gross: $57.30M | Top 250: #4
6	6. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,910,736 | Gross: $377.85M | Top 250: #7

Mini Project 02 - SPECPHONE

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

```
url <- read_html("https://specphone.com/Samsung-Galaxy-Z-Fold5-12-1024GB.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: 33 × 2

attribute	value
<chr>	<chr>
วันเปิดตัว	**กรกฎาคม** 2566
วันวางจำหน่าย	**สิงหาคม** 2566, **ยังไม่วางจำหน่าย**
ขนาด	154.90 x 129.90 x 6.10 **มม.**
น้ำหนัก	253 **กรัม**
วัสดุ	Glass , Plastic , Aluminum frame
SIM	**รองรับ** 2 **ซิมการ์ด** (Nano-SIM, eSIM)
Technology	HSPA, LTE-A, 5G
2G	GSM 850 / 900 / 1800 / 1900
3G	UMTS 850 / 900 / 1900 / 2100 MHz
4G	1/ 2/ 3/ 4/ 5/ 7/ 8/ 13/ 18/ 20/ 25/ 26/ 28/ 31/ 34/ 38/ 39/ 40/ 41
5G	1/ 3/ 5/ 7/ 8/ 20/ 28/ 38/ 41/ 77/ 78
ความเร็ว	HSPA, LTE-A, 5G
ประเภท	AMOLED
ขนาดหน้าจอ	7.60 **นิ้ว**
ความละเอียด	1812 x 2176 pixels
ระบบปฏิบัติการ	Android 13
ชิปประมวลผล	Qualcomm Snapdragon 8 Gen 2 SM8550-AB 3.2 GHz
ชิปกราฟิก	Adreno 740
หน่วยความจำ	12 GB
ความจุ	1024 GB
Memory Card	Not Support
กล้องหลัก	**ตัวที่** 1: 50 MP, f/1.8, 23mm (wide), 1.0µm, Dual Pixel PDAF, OIS **ตัวที่** 2: 10 MP, f/2.4, (telephoto), PDAF, OIS, 3x optical zoom **ตัวที่** 3: 12 MP, f/2.2, 123˚, 12mm (ultrawide), 1.12µm
ความละเอียดวีดีโอ	8K@24fps, 4K@60fps, 1080p@60/240fps (gyro-EIS), 720p@960fps (gyro-EIS), HDR10+
กล้องหน้า	**ตัวที่** 1: 4 MP, f/1.8, 26mm (wide), 2.0µm, under display **ตัวที่** 2: Cover camera :10 MP, f/2.2, 24mm (wide), 1/3
Bluetooth	5.3, A2DP, LE, aptX HD
Wi-Fi	Wi-Fi 802.11 a/b/g/n/ac/6e/7
USB	Type-C 3.2
GPS	GPS, GLONASS, GALILEO, BDS
NFC	**รองรับ**
ความจุ	4,400 mAh
ประเภท	Li-Polymer
Wireless Charging	**รองรับ**
Fast Charging	**รองรับ** (25W / 15W wireless)

```
#All Samsung Smartphones

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

```
# links to all samsung smartphone
links <- samsung_url %>%
    html_nodes("li.mobile-brand-item a") %>%
    html_attr("href")
```

```
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),3)
```

	attribute	value
1	วันเปิดตัว	เมษายน 2566
2	วันวางจำหน่าย	มิถุนายน 2566, ยังไม่วางจำหน่าย
3	ขนาด	162.10 x 77.60 x 8.30 มม.
4	น้ำหนัก	195 กรัม
5	วัสดุ	Glass , Plastic
6	SIM	รองรับ 2 ซิมการ์ด (Nano-SIM, Nano-SIM)

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