Mini Project 01 - IMDB web scraping

```
library(tidyverse)
library(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
```

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

```
movie[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. 12 Angry Men (1957)' · '7. The Lord of the Rings: The Return of the King (2003)' · '8. Pulp Fiction (1994)' · '9. Inception (2010)' · '10. Fight Club (1999)'

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

```
rating[1:10]
```

9.3 · 9.2 · 9 · 9 · 9 · 9 · 9 · 8.9 · 8.8 · 8.8

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

```
df <- data_frame(movie,
                 rating ,
                 num_vote )
```

```
head(df)
```

A tibble: 6 × 3

movie	rating	num_vote
<chr>	<dbl>	<chr>
1. The Shawshank Redemption (1994)	9.3	Votes: 2,700,197 | Gross: $28.34M | Top 250: #1
2. The Godfather (1972)	9.2	Votes: 1,874,281 | Gross: $134.97M | Top 250: #2
3. The Dark Knight (2008)	9.0	Votes: 2,673,959 | Gross: $534.86M | Top 250: #3
4. Schindler's List (1993)	9.0	Votes: 1,365,018 | Gross: $96.90M | Top 250: #6
5. The Godfather Part II (1974)	9.0	Votes: 1,280,718 | Gross: $57.30M | Top 250: #4
6. 12 Angry Men (1957)	9.0	Votes: 797,527 | Gross: $4.36M | Top 250: #5

Mini Project 02 - Specphone Phone Database

```
library(tidyverse)
library(rvest)
```

```
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 <- read_html("https://specphone.com/Realme-GT-3.html")
```

```
url
```

```
{html_document}
<html class="no-js" lang="en-US">
[1] <head itemscope itemtype="https://schema.org/WebSite">\n<meta http-equiv
[2] <body id="blog" class="wp-custom-logo wp-embed-responsive main" itemscop
```

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

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

```
att
```

'วันเปิดตัว' · 'วันวางจำหน่าย' · 'ขนาด' · 'น้ำหนัก' · 'วัสดุ' · 'SIM' · 'Technology' · '2G' · '3G' · '4G' · '5G' · 'ความเร็ว' · 'ประเภท' · 'ขนาดหน้าจอ' · 'ความละเอียด' · 'ระบบปฏิบัติการ' · 'ชิปประมวลผล' · 'ชิปกราฟิก' · 'หน่วยความจำ' · 'ความจุ' · 'Memory Card' · 'กล้องหลัก' · 'ความละเอียดวีดีโอ' · 'กล้องหน้า' · 'Bluetooth' · 'Wi-Fi' · 'USB' · 'GPS' · 'NFC' · 'ความจุ' · 'ประเภท'

```
value
```

'กุมภาพันธ์ 2566' · 'ยังไม่วางจำหน่าย' · '163.90 x 75.80 x 8.90 มม.' · '199 กรัม' · 'ไม่รองรับ' · 'รองรับ 2 ซิมการ์ด (nano sim, nano sim)' · 'HSPA, LTE-A, 5G' · '850/900/1800/1900' · '850/900/1900/2100' · '850/900/1900/2100/2600' · '2100/2600/3500/4700' · 'HSPA, LTE-A, 5G' · 'AMOLED' · '6.74 นิ้ว' · '1240 x 2772 pixels' · 'Android 13' · 'Qualcomm Snapdragon 8+ Gen 1 SM8475 3.19 GHz' · 'Adreno 730' · '16 GB' · '256 GB' · 'ไม่รองรับ' · 'ตัวที่ 1: 50 MP, f/1.9, 24mm (wide), 1/1.56\nตัวที่ 2: 8 MP, f/2.2, 16mm, 112˚ (ultrawide), 1/4.0\nตัวที่ 3: 2 MP, f/3.3, 20mm (microscope)' · '4K@30/60fps, 1080p@30/60fps, gyro-EIS' · 'ตัวที่ 1: 16 MP, f/2.5, 25mm (wide), 1/3.09' · '5.3, A2DP, LE, aptX HD' · '802.11 a/b/g/n/ac, dual-b' · 'Type-C' · 'A-GPS, GLONASS, BDS, GALI' · 'รอบรับ' · '4,600 mAh' · 'Non-removable Li-Po Batt'

```
data_frame(attribute = att,
           value = value)
```

A tibble: 31 × 2

attribute	value
<chr>	<chr>
วันเปิดตัว	กุมภาพันธ์ 2566
วันวางจำหน่าย	ยังไม่วางจำหน่าย
ขนาด	163.90 x 75.80 x 8.90 มม.
น้ำหนัก	199 กรัม
วัสดุ	ไม่รองรับ
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	6.74 นิ้ว
ความละเอียด	1240 x 2772 pixels
ระบบปฏิบัติการ	Android 13
ชิปประมวลผล	Qualcomm Snapdragon 8+ Gen 1 SM8475 3.19 GHz
ชิปกราฟิก	Adreno 730
หน่วยความจำ	16 GB
ความจุ	256 GB
Memory Card	ไม่รองรับ
กล้องหลัก	ตัวที่ 1: 50 MP, f/1.9, 24mm (wide), 1/1.56 ตัวที่ 2: 8 MP, f/2.2, 16mm, 112˚ (ultrawide), 1/4.0 ตัวที่ 3: 2 MP, f/3.3, 20mm (microscope)
ความละเอียดวีดีโอ	4K@30/60fps, 1080p@30/60fps, gyro-EIS
กล้องหน้า	ตัวที่ 1: 16 MP, f/2.5, 25mm (wide), 1/3.09
Bluetooth	5.3, A2DP, LE, aptX HD
Wi-Fi	802.11 a/b/g/n/ac, dual-b
USB	Type-C
GPS	A-GPS, GLONASS, BDS, GALI
NFC	รอบรับ
ความจุ	4,600 mAh
ประเภท	Non-removable Li-Po Batt

```
Warning message:
“`data_frame()` was deprecated in tibble 1.1.0.
Please use `tibble()` instead.
This warning is displayed once every 8 hours.
Call `lifecycle::last_lifecycle_warnings()` to see where this warning was ge
```

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

```
url_sumsung
```

```
{html_document}
<html class="no-js" lang="en-US">
[1] <head itemscope itemtype="https://schema.org/WebSite">\n<meta http-equiv
[2] <body id="blog" class="wp-custom-logo wp-embed-responsive main" itemscop
```

```
links <- url_sumsung %>%
    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[1:5]) {
    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(attributes = ss_topic,
                      value = ss_detail)
    result <- bind_rows(result, tmp)
    print("Progress")

}

print(result)
```

```
[1] "Progress"
[1] "Progress"
[1] "Progress"
[1] "Progress"
[1] "Progress"
```

	attributes
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

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

```
print(head(result),3)
```

```
   attributes                                        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)
```