Mini Project 01 : IMDB Web Scraping

```
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

✔ 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_conflic
✖ 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,d
```

```
print(url)
```

```
[1] "https://www.imdb.com/search/title/?groups=top_100&sort=user_rating,de
```

```
# read html
imdb <- read_html(url)
```

```
imdb
```

```
{html_document}
<html xmlns:og="http://ogp.me/ns#" xmlns:fb="http://www.facebook.com/2008/
[1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=Ul
[2] <body id="styleguide-v2" class="fixed">\n            <img height="1" w
```

```
# movie title
titles <- imdb %>%
    html_nodes("h3.lister-item-header") %>% # html_nodes lookup all header
    html_text2() # text2 -> exclude special characters
```

```
titles
```

'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. Schindler\'s List (1993)' · '6. The Godfather Part II (1974)' · '7. 12 Angry Men (1957)' · '8. Pulp Fiction (1994)' · '9. Inception (2010)' · '10. The Lord of the Rings: The Two Towers (2002)' · '11. Fight Club (1999)' · '12. The Lord of the Rings: The Fellowship of the Ring (2001)' · '13. Forrest Gump (1994)' · '14. Il buono, il brutto, il cattivo (1966)' · '15. The Matrix (1999)' · '16. Goodfellas (1990)' · '17. The Empire Strikes Back (1980)' · '18. One Flew Over the Cuckoo\'s Nest (1975)' · '19. Interstellar (2014)' · '20. Cidade de Deus (2002)' · '21. Sen to Chihiro no Kamikakushi (2001)' · '22. Saving Private Ryan (1998)' · '23. The Green Mile (1999)' · '24. La vita è bella (1997)' · '25. Se7en (1995)' · '26. Terminator 2: Judgment Day (1991)' · '27. The Silence of the Lambs (1991)' · '28. Star Wars (1977)' · '29. Seppuku (1962)' · '30. Shichinin no samurai (1954)' · '31. It\'s a Wonderful Life (1946)' · '32. Gisaengchung (2019)' · '33. Whiplash (2014)' · '34. The Intouchables (2011)' · '35. The Prestige (2006)' · '36. The Departed (2006)' · '37. The Pianist (2002)' · '38. Gladiator (2000)' · '39. American History X (1998)' · '40. The Usual Suspects (1995)' · '41. Léon (1994)' · '42. The Lion King (1994)' · '43. Nuovo Cinema Paradiso (1988)' · '44. Hotaru no haka (1988)' · '45. Back to the Future (1985)' · '46. Apocalypse Now (1979)' · '47. Alien (1979)' · '48. Once Upon a Time in the West (1968)' · '49. Psycho (1960)' · '50. Rear Window (1954)'

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

```
ratings
```

9.3 · 9.2 · 9 · 9 · 9 · 9 · 9 · 8.9 · 8.8 · 8.8 · 8.8 · 8.8 · 8.8 · 8.8 · 8.7 · 8.7 · 8.7 · 8.7 · 8.6 · 8.6 · 8.6 · 8.6 · 8.6 · 8.6 · 8.6 · 8.6 · 8.6 · 8.6 · 8.6 · 8.6 · 8.6 · 8.5 · 8.5 · 8.5 · 8.5 · 8.5 · 8.5 · 8.5 · 8.5 · 8.5 · 8.5 · 8.5 · 8.5 · 8.5 · 8.5 · 8.5 · 8.5 · 8.5 · 8.5

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

```
num_votes[1:10]
```

'Votes: 2,660,451 | Gross: $28.34M | Top 250: #1' ·
'Votes: 1,843,774 | Gross: $134.97M | Top 250: #2' ·
'Votes: 2,633,381 | Gross: $534.86M | Top 250: #3' ·
'Votes: 1,834,490 | Gross: $377.85M | Top 250: #7' ·
'Votes: 1,347,489 | Gross: $96.90M | Top 250: #6' ·
'Votes: 1,263,091 | Gross: $57.30M | Top 250: #4' · 'Votes: 785,596 | Gross: $4.36M | Top 250: #5' ·
'Votes: 2,036,409 | Gross: $107.93M | Top 250: #8' ·
'Votes: 2,333,496 | Gross: $292.58M | Top 250: #14' ·
'Votes: 1,656,450 | Gross: $342.55M | Top 250: #13'

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

)
```

```
head(df, 10)
```

A data.frame: 10 × 3

	title	rating	num_vote
	<chr>	<dbl>	<chr>
1	1. The Shawshank Redemption (1994)	9.3	Votes: 2,660,451 | Gross: $28.34M | Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,843,774 | Gross: $134.97M | Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,633,381 | Gross: $534.86M | Top 250: #3
4	4. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,834,490 | Gross: $377.85M | Top 250: #7
5	5. Schindler's List (1993)	9.0	Votes: 1,347,489 | Gross: $96.90M | Top 250: #6
6	6. The Godfather Part II (1974)	9.0	Votes: 1,263,091 | Gross: $57.30M | Top 250: #4
7	7. 12 Angry Men (1957)	9.0	Votes: 785,596 | Gross: $4.36M | Top 250: #5
8	8. Pulp Fiction (1994)	8.9	Votes: 2,036,409 | Gross: $107.93M | Top 250: #8
9	9. Inception (2010)	8.8	Votes: 2,333,496 | Gross: $292.58M | Top 250: #14
10	10. The Lord of the Rings: The Two Towers (2002)	8.8	Votes: 1,656,450 | Gross: $342.55M | Top 250: #13

Mini Project 02 : Specphone Phone Database

```
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

✔ 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_conflic
✖ dplyr::filter()  masks stats::filter()
✖ purrr::flatten() masks jsonlite::flatten()
✖ dplyr::lag()     masks stats::lag()


Attaching package: ‘rvest’
```

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

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

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

```
df <- data.frame(
    attributes = attribute,
    values = value
)
```

```
df
```

A data.frame: 31 × 2

attributes	values
<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

```
# All Sumsung SmartPhones
sumsung_url <- read_html("https://specphone.com/brand/Samsung")
```

```
# links to all sumsung smartphones
links <- sumsung_url %>%
    html_nodes("li.mobile-brand-item a") %>% # space bar and "a" => looking
    html_attr("href") # html_attr => scraping attribute
```

```
# add text to links for complete links
full_links <- paste0("https://specphone.com", links)
```

```
full_links
```

'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' ·
'https://specphone.com/Samsung-Galaxy-Pocket-Neo.html' ·
'https://specphone.com/Samsung-Galaxy-Young.html' ·
'https://specphone.com/Samsung-Galaxy-J1-Mini.html' ·
'https://specphone.com/Samsung-Galaxy-A01-Core-1-16GB.html' ·
'https://specphone.com/Samsung-Galaxy-V-PLUS.html' ·
'https://specphone.com/Samsung-Galaxy-Young-2.html' ·
'https://specphone.com/Samsung-Galaxy-M02.html' ·
'https://specphone.com/Samsung-Galaxy-A11.html' ·
'https://specphone.com/Samsung-Galaxy-J2-Pro-2018.html' ·
'https://specphone.com/Samsung-Galaxy-A12-2021.html' ·
'https://specphone.com/Samsung-Galaxy-A21s-3-32GB.html' ·
'https://specphone.com/Samsung-Galaxy-J5.html' ·
'https://specphone.com/Samsung-Galaxy-J4.html' ·
'https://specphone.com/Samsung-Galaxy-Core-2-Duos.html' ·
'https://specphone.com/Samsung-Galaxy-Ace-Plus.html' ·
'https://specphone.com/Samsung-Galaxy-A20.html' ·
'https://specphone.com/Samsung-Galaxy-Chat.html' ·
'https://specphone.com/Samsung-Galaxy-Gio.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-A7-Lite-LTE.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-A-10.5WIFI.html' ·
'https://specphone.com/Samsung-Galaxy-Alpha.html' ·
'https://specphone.com/Samsung-Galaxy-S3-Slim.html' ·

'https://specphone.com/Samsung-Galaxy-S4-zoom.html' ·
'https://specphone.com/Samsung-Galaxy-Xcover-2.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-8.9-3G-16GB.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-A8-LTE-2021.html' ·
'https://specphone.com/Samsung-Galaxy-A8-2018.html' ·
'https://specphone.com/Samsung-Galaxy-Tab4-8.0-wifi.html' ·
'https://specphone.com/Samsung-Galaxy-M33-5G.html' ·
'https://specphone.com/Samsung-Galaxy-A50.html' ·
'https://specphone.com/Samsung-Galaxy-E7.html' ·
'https://specphone.com/Samsung-Galaxy-S6.html' ·
'https://specphone.com/Samsung-Galaxy-S20-FE.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-S4-WIFI.html' ·
'https://specphone.com/Samsung-Galaxy-S7.html' ·
'https://specphone.com/Samsung-Galaxy-Note-5-Exynos.html' ·
'https://specphone.com/Samsung-Galaxy-TabPRO-12.2-LTE.html' ·
'https://specphone.com/Samsung-Galaxy-S4-Active.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-Active-3.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-S3-9.7.html' ·
'https://specphone.com/Samsung-Galaxy-S6-edge.html' ·
'https://specphone.com/Samsung-Galaxy-Note-4-Exynos.html' ·
'https://specphone.com/Samsung-Galaxy-Round.html' ·
'https://specphone.com/Samsung-Galaxy-Note-20-Ultra-5G.html' ·
'https://specphone.com/Samsung-ATIV-Q.html' ·
'https://specphone.com/Samsung-ATIV-Smart-PC-PRO.html' ·
'https://specphone.com/Samsung-Galaxy-S22-Ultra12-128GB.html' ·
'https://specphone.com/Samsung-Galaxy-Z-Flip-5G.html' ·
'https://specphone.com/Samsung-Galaxy-Z-Flip.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-S8-Ultra-5G.html' ·
'https://specphone.com/Samsung-Galaxy-S21-Ultra-16-512GB.html' ·
'https://specphone.com/Samsung-Galaxy-S10-Plus-Ram-12GB.html' ·
'https://specphone.com/Samsung-Galaxy-Z-Fold-3.html' ·
'https://specphone.com/Samsung-Galaxy-Z-Fold4.html' ·
'https://specphone.com/Samsung-Galaxy-Z-Fold-2-5G.html'

```
result <- data.frame()
# create for loop
for (link in full_links) {
    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...")
}
```

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

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