Mini Project 01 - IMDB web scraping

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

```
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. The Shawshank Redemption (1994)' · '2. The Godfather (1972)' · '3. The Dark Knight (2008)' ·
'4. Schindler\'s List (1993)' · '5. The Lord of the Rings: The Return of the King (2003)' ·
'6. The Godfather Part II (1974)' · '7. 12 Angry Men (1957)' · '8. Pulp Fiction (1994)' ·
'9. The Lord of the Rings: The Fellowship of the Ring (2001)' · '10. Inception (2010)' · '11. Fight Club (1999)' ·
'12. Forrest Gump (1994)' · '13. The Lord of the Rings: The Two Towers (2002)' ·
'14. Il buono, il brutto, il cattivo (1966)' · '15. The Matrix (1999)' · '16. One Flew Over the Cuckoo\'s Nest (1975)' ·
'17. GoodFellas (1990)' · '18. The Empire Strikes Back (1980)' · '19. Interstellar (2014)' · '20. Se7en (1995)' ·
'21. The Silence of the Lambs (1991)' · '22. The Green Mile (1999)' · '23. Star Wars (1977)' ·
'24. Saving Private Ryan (1998)' · '25. Terminator 2: Judgment Day (1991)' ·
'26. Sen to Chihiro no kamikakushi (2001)' · '27. La vita è bella (1997)' · '28. Cidade de Deus (2002)' ·
'29. It\'s a Wonderful Life (1946)' · '30. Shichinin no samurai (1954)' · '31. Seppuku (1962)' · '32. Whiplash (2014)' ·
'33. Gladiator (2000)' · '34. Gisaengchung (2019)' · '35. The Departed (2006)' · '36. Back to the Future (1985)' ·
'37. The Prestige (2006)' · '38. Apocalypse Now (1979)' · '39. Léon (1994)' · '40. Alien (1979)' ·
'41. The Usual Suspects (1995)' · '42. The Lion King (1994)' · '43. American History X (1998)' ·
'44. Once Upon a Time in the West (1968)' · '45. The Pianist (2002)' · '46. The Intouchables (2011)' ·
'47. Casablanca (1942)' · '48. Psycho (1960)' · '49. Hotaru no haka (1988)' · '50. Rear Window (1954)'
```

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

```r
ratings[1:10]
```

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

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

```r
#build a data set
df <- data_frame(
    title = titles,
    rating = ratings,
    num_vote = num_votes
)

head(df)
```

A tibble: 6 × 3

title	rating	num_vote
<chr>	<dbl>	<chr>
1. The Shawshank Redemption (1994)	9.3	Votes: 2,703,010 | Gross: $28.34M | Top 250: #1
2. The Godfather (1972)	9.2	Votes: 1,876,695 | Gross: $134.97M | Top 250: #2
3. The Dark Knight (2008)	9.0	Votes: 2,676,843 | Gross: $534.86M | Top 250: #3
4. Schindler's List (1993)	9.0	Votes: 1,366,281 | Gross: $96.90M | Top 250: #6
5. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,861,157 | Gross: $377.85M | Top 250: #7
6. The Godfather Part II (1974)	9.0	Votes: 1,282,012 | Gross: $57.30M | Top 250: #4

```
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 gener
```

Mini Project 02 - Specphone Database

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

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

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

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

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

A tibble: 31 × 2

attributes	value
<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 Samsung smartphone

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

```
# link to all samsumg 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[1:5]) {
    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("Progress ...")
}

#print(result)
```

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

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

```
# A
#   tibble:
#   6
#   ×
#   2
# … with 2 more variables: attribute <chr>, value <chr>
```

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