Mini Project 01 - IMDB Web scraping

```
library(tidyverse)
library(rvest) # scrap 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
read_html(url)
imdb <- read_html(url) ## ฝากค่า
```

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

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

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

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

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

```
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 · 8.5
```

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

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

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

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

```
'Votes: 2,689,604 | Gross: $28.34M | Top 250: #1' · 'Votes: 1,865,515 | Gross: $134.97M | Top 250: #2' ·
'Votes: 2,663,106 | Gross: $534.86M | Top 250: #3' · 'Votes: 1,360,330 | Gross: $96.90M | Top 250: #6' ·
'Votes: 1,276,106 | Gross: $57.30M | Top 250: #4' · 'Votes: 794,577 | Gross: $4.36M | Top 250: #5' ·
'Votes: 1,853,016 | Gross: $377.85M | Top 250: #7' · 'Votes: 2,063,956 | Gross: $107.93M | Top 250: #8' ·
'Votes: 2,362,244 | Gross: $292.58M | Top 250: #14' · 'Votes: 2,135,107 | Gross: $37.03M | Top 250: #12' ·
'Votes: 1,882,593 | Gross: $315.54M | Top 250: #9' · 'Votes: 2,088,146 | Gross: $330.25M | Top 250: #11' ·
'Votes: 765,107 | Gross: $6.10M | Top 250: #10' · 'Votes: 1,673,321 | Gross: $342.55M | Top 250: #13' ·
'Votes: 1,166,961 | Gross: $46.84M | Top 250: #17' · 'Votes: 1,920,205 | Gross: $171.48M | Top 250: #16' ·
'Votes: 1,011,721 | Gross: $112.00M | Top 250: #18' · 'Votes: 1,297,476 | Gross: $290.48M | Top 250: #15' ·
'Votes: 1,843,814 | Gross: $188.02M | Top 250: #26' · 'Votes: 1,660,097 | Gross: $100.13M | Top 250: #19' ·
'Votes: 698,689 | Gross: $57.60M | Top 250: #25' · 'Votes: 1,307,644 | Gross: $136.80M | Top 250: #27' ·
'Votes: 1,438,556 | Gross: $130.74M | Top 250: #22' · 'Votes: 1,370,016 | Gross: $322.74M | Top 250: #28' ·
'Votes: 1,397,147 | Gross: $216.54M | Top 250: #24' · 'Votes: 1,104,032 | Gross: $204.84M | Top 250: #29' ·
'Votes: 768,571 | Gross: $10.06M | Top 250: #31' · 'Votes: 760,147 | Gross: $7.56M | Top 250: #23' ·
'Votes: 465,489 | Top 250: #21' · 'Votes: 348,078 | Gross: $0.27M | Top 250: #20' · 'Votes: 58,392 | Top 250: #44' ·
'Votes: 870,002 | Gross: $13.09M | Top 250: #42' · 'Votes: 812,300 | Gross: $53.37M | Top 250: #34' ·
'Votes: 1,339,186 | Gross: $53.09M | Top 250: #41' · 'Votes: 1,330,970 | Gross: $132.38M | Top 250: #39' ·
'Votes: 1,166,923 | Gross: $19.50M | Top 250: #35' · 'Votes: 1,506,891 | Gross: $187.71M | Top 250: #37' ·
'Votes: 671,359 | Gross: $83.47M | Top 250: #53' · 'Votes: 887,523 | Gross: $78.90M | Top 250: #51' ·
'Votes: 1,211,690 | Gross: $210.61M | Top 250: #30' · 'Votes: 1,089,633 | Gross: $23.34M | Top 250: #40' ·
'Votes: 1,127,137 | Gross: $6.72M | Top 250: #38' · 'Votes: 1,063,345 | Gross: $422.78M | Top 250: #36' ·
'Votes: 836,799 | Gross: $32.57M | Top 250: #32' · 'Votes: 863,560 | Gross: $13.18M | Top 250: #46' ·
'Votes: 332,177 | Gross: $5.32M | Top 250: #48' · 'Votes: 574,922 | Gross: $1.02M | Top 250: #43' ·
'Votes: 675,607 | Gross: $32.00M | Top 250: #33' · 'Votes: 280,229 | Top 250: #45' ·
'Votes: 263,509 | Gross: $11.99M | Top 250: #50'
```

```
# build a dataset
df <- data.frame(
    title = title,
    rating = ratings,
    num_vote = votes
)
head(df)
```

A data.frame: 6 × 3

	title	rating	num_vote
	<chr>	<dbl>	<chr>
1	1. The Shawshank Redemption (1994)	9.3	Votes: 2,689,604 | Gross: $28.34M | Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,865,515 | Gross: $134.97M | Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,663,106 | Gross: $534.86M | Top 250: #3
4	4. Schindler's List (1993)	9.0	Votes: 1,360,330 | Gross: $96.90M | Top 250: #6
5	5. The Godfather Part II (1974)	9.0	Votes: 1,276,106 | Gross: $57.30M | Top 250: #4
6	6. 12 Angry Men (1957)	9.0	Votes: 794,577 | Gross: $4.36M | Top 250: #5

Mini Project 02 - Specphone Phone Database

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

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

```
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: 31 × 2

attribute	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")
```

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

links <- samsung_url %>%
    html_nodes("li.mobile-brand-item a") %>%
    html_attr("href")
```

```
'/Samsung-Galaxy-M13.html' · '/Samsung-Galaxy-A23.html' · '/Samsung-Galaxy-A13.html' ·
'/Samsung-Galaxy-M32-5G.html' · '/Samsung-Galaxy-A12-Nacho.html' · '/Samsung-Galaxy-Pocket-Neo.html' ·
'/Samsung-Galaxy-Young.html' · '/Samsung-Galaxy-J1-Mini.html' · '/Samsung-Galaxy-A01-Core-1-16GB.html' ·
'/Samsung-Galaxy-V-PLUS.html' · '/Samsung-Galaxy-Young-2.html' · '/Samsung-Galaxy-M02.html' ·
'/Samsung-Galaxy-A11.html' · '/Samsung-Galaxy-J2-Pro-2018.html' · '/Samsung-Galaxy-A12-2021.html' ·
'/Samsung-Galaxy-A21s-3-32GB.html' · '/Samsung-Galaxy-J5.html' · '/Samsung-Galaxy-J4.html' ·
'/Samsung-Galaxy-Core-2-Duos.html' · '/Samsung-Galaxy-Ace-Plus.html' · '/Samsung-Galaxy-A20.html' ·
'/Samsung-Galaxy-Chat.html' · '/Samsung-Galaxy-Gio.html' · '/Samsung-Galaxy-Tab-A7-Lite-LTE.html' ·
'/Samsung-Galaxy-Tab-A-10.5WIFI.html' · '/Samsung-Galaxy-Alpha.html' · '/Samsung-Galaxy-S3-Slim.html' ·
'/Samsung-Galaxy-S4-zoom.html' · '/Samsung-Galaxy-Xcover-2.html' · '/Samsung-Galaxy-Tab-8.9-3G-16GB.html' ·
'/Samsung-Galaxy-Tab-A8-LTE-2021.html' · '/Samsung-Galaxy-A8-2018.html' ·
'/Samsung-Galaxy-Tab4-8.0-wifi.html' · '/Samsung-Galaxy-M33-5G.html' · '/Samsung-Galaxy-A50.html' ·
'/Samsung-Galaxy-E7.html' · '/Samsung-Galaxy-S6.html' · '/Samsung-Galaxy-S20-FE.html' ·
'/Samsung-Galaxy-Tab-S4-WIFI.html' · '/Samsung-Galaxy-S7.html' · '/Samsung-Galaxy-Note-5-Exynos.html' ·
'/Samsung-Galaxy-TabPRO-12.2-LTE.html' · '/Samsung-Galaxy-S4-Active.html' ·
'/Samsung-Galaxy-Tab-Active-3.html' · '/Samsung-Galaxy-Tab-S3-9.7.html' · '/Samsung-Galaxy-S6-edge.html' ·
'/Samsung-Galaxy-Note-4-Exynos.html' · '/Samsung-Galaxy-Round.html' ·
'/Samsung-Galaxy-Note-20-Ultra-5G.html' · '/Samsung-ATIV-Q.html' · '/Samsung-ATIV-Smart-PC-PRO.html' ·
'/Samsung-Galaxy-S22-Ultra12-128GB.html' · '/Samsung-Galaxy-Z-Flip-5G.html' · '/Samsung-Galaxy-Z-Flip.html' ·
'/Samsung-Galaxy-Tab-S8-Ultra-5G.html' · '/Samsung-Galaxy-S21-Ultra-16-512GB.html' ·
'/Samsung-Galaxy-S10-Plus-Ram-12GB.html' · '/Samsung-Galaxy-Z-Fold-3.html' · '/Samsung-Galaxy-Z-Fold4.html' ·
'/Samsung-Galaxy-Z-Fold-2-5G.html'
```

```
# full links (full name link)
paste0("https://specphone.com",links)

full_links <- paste0("https://specphone.com",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'

```
# for loop
result <- data.frame()

    for (link in full_links[1:5]) {
        ss_topic <- link %>%
            read_html() %>%  # differance links
            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)
```

```
    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_specphone.csv")
```