Mini project1

```
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 1.3

✓ 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 <- "https://www.imdb.com/search/title/?groups=top_100&sort=user_rating%2Cd
```

```
print(url)
```

```
[1] "https://www.imdb.com/search/title/?groups=top_100&sort=user_rating%2Cde
```

```
# read html
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 title

```
# movie title
imdb %>%
    html_node("h3.lister-item-header") %>%
    html_text() #select text in h3.lister-item-header node
```

'\n   1.\n   \n   The Shawshank Redemption\n   (1994)\n'

```
# movie title
imdb %>%
    html_node("h3.lister-item-header") %>%
    html_text2() #select text but remove special character
```

'1. The Shawshank Redemption (1994)'

add "s" after html_node -> html_nodes to select all nodes

```
# movie title
imdb %>%
    html_nodes("h3.lister-item-header") %>% #add "s"
    html_text2()
```

'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)'

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

Rating

```
# rating
imdb %>%
    html_node("div.ratings-imdb-rating") %>%
    html_text2()
```

'9.3'

all ratings and convert to be `numeric` type

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

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

```
head(num_votes)
```

'Votes: 2,657,665 | Gross: $28.34M | Top 250: #1'

Build data set

```
df <- data.frame(
    title = titles,
    rating = ratings,
    num_vote = num_votes
)

head(df)
```

A data.frame: 6 × 3

	title	rating	num_vote
	<chr>	<dbl>	<chr>
1	1. The Shawshank Redemption (1994)	9.3	Votes: 2,657,665 | Gross: $28.34M | Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 2,657,665 | Gross: $28.34M | Top 250: #1
3	3. The Dark Knight (2008)	9.0	Votes: 2,657,665 | Gross: $28.34M | Top 250: #1
4	4. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 2,657,665 | Gross: $28.34M | Top 250: #1
5	5. Schindler's List (1993)	9.0	Votes: 2,657,665 | Gross: $28.34M | Top 250: #1
6	6. The Godfather Part II (1974)	9.0	Votes: 2,657,665 | Gross: $28.34M | Top 250: #1

Mini Project 02 - Specphone Phone Database

load library

```
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 1.3

✓ 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/Samsung-Galaxy-A04.html")
```

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

value
```

'ตุลาคม 2565' · 'ยังไม่วางจำหน่าย' · '164.40 x 76.30 x 9.10 มม.' · '192 กรัม' · 'Glass front, plastic back, plastic frame' · 'รองรับ 2 ซิมการ์ด (nano sim, nano sim)' · 'HSPA 42.2/5.76 Mbps, LTE-A' · '850/900/1800/1900' · '850/900/1900/2100' · '850/900/1900/2100/2600' · '-' · 'HSPA 42.2/5.76 Mbps, LTE-A' · 'PLS LCD' · '6.50 นิ้ว' · '720 x 1600 pixels' · 'Android 12' · 'Spreadtrum Unisoc SC9863A 1.6 GHz' · 'PowerVR GE8322' · '3 GB' · '32 GB' · 'microSD (1)' · 'ตัวที่ 1: 50 MP, f/1.8, (wide), AF\nตัวที่ 2: 2 MP, f/2.4, (depth)' · '1080p@30fps' · 'ตัวที่ 1: 5 MP, f/2.2' · '5.0, A2DP, LE' · '802.11 a/b/g/n/ac, dual-b' · 'Type-C' · 'GLONASS, GALILEO, BDS' · 'ไม่รองรับ' · '5,000 mAh' · 'Non-removable Li-Po Batt'

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

เคาะ space ตามด้วยชื่อ tag หมายถึงเข้าไปหา child ที่ชื่อ tag นั้น

```
# find links samsung smartphone
samsung_url %>%
    html_node("li.mobile-brand-item a") %>% #find a in li.mobile-brand-item
    html_attr("href") # find href
```

'/Samsung-Galaxy-M13.html'

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

```
links
```

'/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' ·

```
paste0("https://specphone.com",links[5:10])
```

'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'

```
full_links <- paste0("https://specphone.com", links)
```

```
result <- data.frame()

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...")
}

print(result)
```

```
601
602
603
604
605
606
607
```

```
608
609
610
611
612
613
614
615
616
617
618
619
```

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

```
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..."
         attribute
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
14        ขนาดหน้าจอ
```

```
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)
```