Mini Project - IMDB web scraping

```
library(tidyverse) #for prep data visualization
library(rvest) #for web scraping
```

```
Warning message in system("timedatectl", intern = TRUE):
“running command 'timedatectl' had status 1”
Warning message:
“Failed to locate timezone database”
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1

✓ 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_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,desc"
```

```
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 .
```

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

title
```

```
'1. The Shawshank Redemption (1994)'
```

```
#scrape titles
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. 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)'
```

```
#rating
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 · 8.5
```

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

votes
```

```
'Votes: 2,670,850 | Gross: $28.34M | Top 250: #1' · 'Votes: 1,851,020 | Gross: $134.97M | Top 250: #2' ·
'Votes: 2,643,649 | Gross: $534.86M | Top 250: #3' · 'Votes: 1,840,776 | Gross: $377.85M | Top 250: #7' ·
'Votes: 1,352,135 | Gross: $96.90M | Top 250: #6' · 'Votes: 1,267,484 | Gross: $57.30M | Top 250: #4' ·
'Votes: 788,834 | Gross: $4.36M | Top 250: #5' · 'Votes: 2,045,770 | Gross: $107.93M | Top 250: #8' ·
'Votes: 2,343,139 | Gross: $292.58M | Top 250: #14' · 'Votes: 1,662,167 | Gross: $342.55M | Top 250: #13' ·
'Votes: 2,116,266 | Gross: $37.03M | Top 250: #12' · 'Votes: 1,870,016 | Gross: $315.54M | Top 250: #9' ·
'Votes: 2,070,900 | Gross: $330.25M | Top 250: #11' · 'Votes: 760,881 | Gross: $6.10M | Top 250: #10' ·
'Votes: 1,907,801 | Gross: $171.48M | Top 250: #16' · 'Votes: 1,157,915 | Gross: $46.84M | Top 250: #17' ·
'Votes: 1,289,500 | Gross: $290.48M | Top 250: #15' · 'Votes: 1,006,615 | Gross: $112.00M | Top 250: #18' ·
'Votes: 1,819,314 | Gross: $188.02M | Top 250: #26' · 'Votes: 756,373 | Gross: $7.56M | Top 250: #23' ·
'Votes: 761,667 | Gross: $10.06M | Top 250: #31' · 'Votes: 1,388,015 | Gross: $216.54M | Top 250: #24' ·
'Votes: 1,298,112 | Gross: $136.80M | Top 250: #27' · 'Votes: 694,258 | Gross: $57.60M | Top 250: #25' ·
'Votes: 1,646,764 | Gross: $100.13M | Top 250: #19' · 'Votes: 1,097,073 | Gross: $204.84M | Top 250: #29' ·
'Votes: 1,429,006 | Gross: $130.74M | Top 250: #22' · 'Votes: 1,362,069 | Gross: $322.74M | Top 250: #28' ·
'Votes: 57,612 | Top 250: #44' · 'Votes: 346,370 | Gross: $0.27M | Top 250: #20' · 'Votes: 456,608 | Top 250: #21' ·
'Votes: 797,442 | Gross: $53.37M | Top 250: #34' · 'Votes: 856,978 | Gross: $13.09M | Top 250: #42' ·
'Votes: 856,731 | Gross: $13.18M | Top 250: #46' · 'Votes: 1,329,957 | Gross: $53.09M | Top 250: #41' ·
'Votes: 1,322,300 | Gross: $132.38M | Top 250: #39' · 'Votes: 830,668 | Gross: $32.57M | Top 250: #33' ·
'Votes: 1,496,559 | Gross: $187.71M | Top 250: #37' · 'Votes: 1,121,056 | Gross: $6.72M | Top 250: #38' ·
'Votes: 1,083,974 | Gross: $23.34M | Top 250: #40' · 'Votes: 1,158,670 | Gross: $19.50M | Top 250: #35' ·
'Votes: 1,055,876 | Gross: $422.78M | Top 250: #36' · 'Votes: 261,732 | Gross: $11.99M | Top 250: #50' ·
'Votes: 277,799 | Top 250: #45' · 'Votes: 1,201,911 | Gross: $210.61M | Top 250: #30' ·
'Votes: 667,339 | Gross: $83.47M | Top 250: #53' · 'Votes: 881,557 | Gross: $78.90M | Top 250: #51' ·
'Votes: 330,094 | Gross: $5.32M | Top 250: #48' · 'Votes: 671,810 | Gross: $32.00M | Top 250: #32' ·
'Votes: 492,104 | Gross: $36.76M | Top 250: #49'
```

```
#Build Dataset
df <- data.frame(
    title = titles,
    vote = votes,
    rating = ratings
)
```

```
head(df)
```

A data.frame: 6 × 3

	title	vote	rating
	<chr>	<chr>	<dbl>
1	1. The Shawshank Redemption (1994)	Votes: 2,670,850 | Gross: $28.34M | Top 250: #1	9.3
2	2. The Godfather (1972)	Votes: 1,851,020 | Gross: $134.97M | Top 250: #2	9.2
3	3. The Dark Knight (2008)	Votes: 2,643,649 | Gross: $534.86M | Top 250: #3	9.0
4	4. The Lord of the Rings: The Return of the King (2003)	Votes: 1,840,776 | Gross: $377.85M | Top 250: #7	9.0
5	5. Schindler's List (1993)	Votes: 1,352,135 | Gross: $96.90M | Top 250: #6	9.0
6	6. The Godfather Part II (1974)	Votes: 1,267,484 | Gross: $57.30M | Top 250: #4	9.0

Mini Project - Spec phone scraping

```
library(tidyverse) #for prep data visualization
library(rvest) #for web scraping
```

```
Warning message in system("timedatectl", intern = TRUE):
“running command 'timedatectl' had status 1”
Warning message:
“Failed to locate timezone database”
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1

✓ 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_conflicts()
✗ dplyr::filter()  masks stats::filter()
✗ purrr::flatten() masks jsonlite::flatten()
✗ dplyr::lag()     masks stats::lag()


Attaching package: ‘rvest’
```

```
url <- "https://specphone.com/Honor-80.html"
```

```
#scraping attribute
att <- url %>%
    read_html()%>%
    html_nodes("div.topic")%>%
    html_text2()
```

```
#scraping value
val <- url %>%
    read_html()%>%
    html_nodes("div.detail")%>%
    html_text2()
```

```
#create dataframe
df <- data.frame(
    attribute = att,
    value = val
)
```

```
df
```

A data.frame: 32 × 2

attribute	value
<chr>	<chr>
วันเปิดตัว	พฤศจิกายน 2565
วันวางจำหน่าย	ยังไม่วางจำหน่าย
ขนาด	161.10 x 73.90 x 7.70 มม.
น้ำหนัก	180 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	OLED
ขนาดหน้าจอ	6.67 นิ้ว
ความละเอียด	1080 x 2400 pixels
ระบบปฏิบัติการ	Android 12
ชิปประมวลผล	Qualcomm Snapdragon 782G null 2.7 GHz
ชิปกราฟิก	Adreno 642L
หน่วยความจำ	8 GB
ความจุ	256 GB
Memory Card	ไม่รองรับ
กล้องหลัก	ตัวที่ 1: 160 MP, f/1.8, (wide), 1/1.56 ตัวที่ 2: 8 MP, f/2.2, (ultrawide), AF ตัวที่ 3: 2 MP, f/2.4, (macro)
ความละเอียดวีดีโอ	4K@30fps, 1080p@30/60fps, gyro-EIS
กล้องหน้า	ตัวที่ 1: 32 MP, f/2.4, (wide)
Bluetooth	5.2, A2DP, LE, aptX HD
Wi-Fi	802.11 a/b/g/n/ac, dual-b
USB	Type-C
GPS	GPS, GALILEO, GLONASS, BD
NFC	รอบรับ
ความจุ	4,800 mAh
ประเภท	Non-removable Li-Po Batt
Fast Charging	รองรับ (66W)

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

```
#get a link
links <- samsung_url %>%
    html_nodes("li.mobile-brand-item a")%>%
    html_attr("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' ·
'/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_link <- paste0("https://specphone.com",links)
full_link[1:10]
```

```
'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'
```

```
#get a full link in every samsung phone
result <- data.frame()

for (link in full_link[1:10]){

    ss_att <- link %>%
    read_html()%>%
    html_nodes("div.topic")%>%
    html_text2()

    ss_val <- link %>%
    read_html()%>%
    html_nodes("div.detail")%>%
    html_text2()

    tmp <- data.frame(
        attribute = ss_att,
        value = ss_val
    )

    result <- bind_rows(result,tmp)

}
```

```
print(result)
```

```
          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
15       ความละเอียด
16     ระบบปฏิบัติการ
17       ชิปประมวลผล
18          ชิปกราฟิก
19        หน่วยความจำ
```

```
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)
```

```
#Export CSV file
write_csv(result,"result.csv")
```