Mini Project01 - IMDB web scraping

```
library(tidyverse)
library(rvest)
```

```
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"
```

```
print(url)
```

```
[1] "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 .
```

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

```
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 · 8.8 · 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

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

```
num_votes
```

'Votes: 2,681,809 | Gross: $28.34M | Top 250: #1' · 'Votes: 1,859,304 | Gross: $134.97M | Top 250: #2' · 'Votes: 2,655,144 | Gross: $534.86M | Top 250: #3' · 'Votes: 1,848,011 | Gross: $377.85M | Top 250: #7' · 'Votes: 1,272,627 | Gross: $57.30M | Top 250: #4' · 'Votes: 1,356,988 | Gross: $96.90M | Top 250: #6' · 'Votes: 792,253 | Gross: $4.36M | Top 250: #5' · 'Votes: 2,057,037 | Gross: $107.93M | Top 250: #8' · 'Votes: 1,877,445 | Gross: $315.54M | Top 250: #9' · 'Votes: 2,127,330 | Gross: $37.03M | Top 250: #12' · 'Votes: 2,354,657 | Gross: $292.58M | Top 250: #14' · 'Votes: 2,081,013 | Gross: $330.25M | Top 250: #11' · 'Votes: 1,668,752 | Gross: $342.55M | Top 250: #13' · 'Votes: 763,333 | Gross: $6.10M | Top 250: #10' · 'Votes: 1,163,378 | Gross: $46.84M | Top 250: #17' · 'Votes: 1,915,073 | Gross: $171.48M | Top 250: #16' · 'Votes: 1,009,658 | Gross: $112.00M | Top 250: #18' · 'Votes: 1,294,158 | Gross: $290.48M | Top 250: #15' · 'Votes: 464,091 | Top 250: #21' · 'Votes: 1,833,702 | Gross: $188.02M | Top 250: #26' · 'Votes: 1,654,568 | Gross: $100.13M | Top 250: #19' · 'Votes: 1,303,777 | Gross: $136.80M | Top 250: #27' · 'Votes: 1,366,777 | Gross: $322.74M | Top 250: #28' · 'Votes: 1,434,577 | Gross: $130.74M | Top 250: #22' · 'Votes: 1,101,269 | Gross: $204.84M | Top 250: #29' · 'Votes: 1,393,433 | Gross: $216.54M | Top 250: #24' · 'Votes: 758,582 | Gross: $7.56M | Top 250: #23' · 'Votes: 765,701 | Gross: $10.06M | Top 250: #31' · 'Votes: 696,889 | Gross: $57.60M | Top 250: #25' · 'Votes: 347,387 | Gross: $0.27M | Top 250: #20' · 'Votes: 58,081 | Top 250: #44' · 'Votes: 864,619 | Gross: $13.09M | Top 250: #42' · 'Votes: 1,502,623 | Gross: $187.71M | Top 250: #37' · 'Votes: 806,038 | Gross: $53.37M | Top 250: #34' · 'Votes: 1,207,797 | Gross: $210.61M | Top 250: #30' · 'Votes: 1,163,239 | Gross: $19.50M | Top 250: #35' · 'Votes: 885,120 | Gross: $78.90M | Top 250: #51' · 'Votes: 1,327,484 | Gross: $132.38M | Top 250: #39' · 'Votes: 1,335,426 | Gross: $53.09M | Top 250: #41' · 'Votes: 1,124,684 | Gross: $6.72M | Top 250: #38' · 'Votes: 669,617 | Gross: $83.47M | Top 250: #53' · 'Votes: 493,721 | Gross: $36.76M | Top 250: #49' · 'Votes: 1,087,394 | Gross: $23.34M | Top 250: #40' · 'Votes: 1,060,296 | Gross: $422.78M | Top 250: #36' · 'Votes: 331,401 | Gross: $5.32M | Top 250: #48' · 'Votes: 860,860 | Gross: $13.18M | Top 250: #46' · 'Votes: 834,275 | Gross: $32.57M | Top 250: #33' · 'Votes: 573,839 | Gross: $1.02M | Top 250: #43' · 'Votes: 674,062 | Gross: $32.00M | Top 250: #32' · 'Votes: 279,257 | Top 250: #45'

```
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: 2,681,809 | Gross: $28.34M | Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,859,304 | Gross: $134.97M | Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,655,144 | Gross: $534.86M | Top 250: #3
4	4. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,848,011 | Gross: $377.85M | Top 250: #7
5	5. The Godfather Part II (1974)	9.0	Votes: 1,272,627 | Gross: $57.30M | Top 250: #4
6	6. Schindler's List (1993)	9.0	Votes: 1,356,988 | Gross: $96.90M | Top 250: #6

Project 02 - Specphone Phone Database

```
library(tidyverse)
library(rvest)
```

```
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()
```

```
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
```

'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 Phone

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

```
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_links <- paste0("http://specphone.com", links)
```

```
full_links
```

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

```
result <- data.frame()

for (link in full_links[1:10]) {
    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(attrebutes = 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..."
```

```
head(result)
```

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