Mini Project 1 - IMDB Project 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)

imbd <- read_html(url)
imbd
```

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

{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 .
```

```
# read html
imbd <- read_html(url)
imbd
```

```
{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 .
```

```
imdb <- read_html(url)

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

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

```
'9.3' · '9.2' · '9.0' · '9.0' · '9.0' · '9.0' · '9.0' · '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'
```

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

```
'Votes: 2,709,329 | Gross: $28.34M | Top 250: #1' · 'Votes: 1,881,474 | Gross: $134.97M | Top 250: #2' ·
'Votes: 2,682,358 | Gross: $534.86M | Top 250: #3' · 'Votes: 1,368,887 | Gross: $96.90M | Top 250: #6' ·
'Votes: 1,864,698 | Gross: $377.85M | Top 250: #7' · 'Votes: 1,284,571 | Gross: $57.30M | Top 250: #4' ·
'Votes: 800,409 | Gross: $4.36M | Top 250: #5' · 'Votes: 2,080,021 | Gross: $107.93M | Top 250: #8' ·
'Votes: 1,894,210 | Gross: $315.54M | Top 250: #9' · 'Votes: 2,380,317 | Gross: $292.58M | Top 250: #14' ·
'Votes: 2,153,354 | Gross: $37.03M | Top 250: #12' · 'Votes: 2,105,091 | Gross: $330.25M | Top 250: #11' ·
'Votes: 1,683,709 | Gross: $342.55M | Top 250: #13' · 'Votes: 769,005 | Gross: $6.10M | Top 250: #10' ·
'Votes: 1,175,114 | Gross: $46.84M | Top 250: #17' · 'Votes: 1,932,806 | Gross: $171.48M | Top 250: #16' ·
'Votes: 1,016,849 | Gross: $112.00M | Top 250: #18' · 'Votes: 1,305,043 | Gross: $290.48M | Top 250: #15' ·
'Votes: 1,867,991 | Gross: $188.02M | Top 250: #25' · 'Votes: 1,448,430 | Gross: $130.74M | Top 250: #22' ·
'Votes: 1,672,912 | Gross: $100.13M | Top 250: #19' · 'Votes: 1,316,932 | Gross: $136.80M | Top 250: #27' ·
'Votes: 1,405,945 | Gross: $216.54M | Top 250: #23' · 'Votes: 1,377,453 | Gross: $322.74M | Top 250: #28' ·
'Votes: 1,110,378 | Gross: $204.84M | Top 250: #29' · 'Votes: 775,446 | Gross: $10.06M | Top 250: #31' ·
'Votes: 763,454 | Gross: $7.56M | Top 250: #24' · 'Votes: 703,095 | Gross: $57.60M | Top 250: #26' ·
'Votes: 467,521 | Top 250: #21' · 'Votes: 349,333 | Gross: $0.27M | Top 250: #20' · 'Votes: 59,266 | Top 250: #45' ·
'Votes: 882,796 | Gross: $13.09M | Top 250: #42' · 'Votes: 827,323 | Gross: $53.37M | Top 250: #34' ·
'Votes: 1,220,544 | Gross: $210.61M | Top 250: #30' · 'Votes: 1,516,318 | Gross: $187.71M | Top 250: #37' ·
'Votes: 1,339,081 | Gross: $132.38M | Top 250: #39' · 'Votes: 1,347,677 | Gross: $53.09M | Top 250: #41' ·
'Votes: 674,895 | Gross: $83.47M | Top 250: #53' · 'Votes: 1,174,424 | Gross: $19.50M | Top 250: #35' ·
'Votes: 892,504 | Gross: $78.90M | Top 250: #51' · 'Votes: 1,094,587 | Gross: $23.34M | Top 250: #40' ·
'Votes: 1,070,705 | Gross: $422.78M | Top 250: #36' · 'Votes: 1,132,454 | Gross: $6.72M | Top 250: #38' ·
'Votes: 843,147 | Gross: $32.57M | Top 250: #32' · 'Votes: 869,731 | Gross: $13.18M | Top 250: #46' ·
'Votes: 333,639 | Gross: $5.32M | Top 250: #48' · 'Votes: 577,376 | Gross: $1.02M | Top 250: #43' ·
'Votes: 679,083 | Gross: $32.00M | Top 250: #33' · 'Votes: 282,241 | Top 250: #44' ·
'Votes: 264,820 | Gross: $11.99M | Top 250: #50'
```

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

head(df)
```

A data.frame: 6 × 3

	title	rating	num_vote
	<chr>	<chr>	<chr>
1	1. The Shawshank Redemption (1994)	9.3	Votes: 2,709,329 | Gross: $28.34M | Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,881,474 | Gross: $134.97M | Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,682,358 | Gross: $534.86M | Top 250: #3
4	4. Schindler's List (1993)	9.0	Votes: 1,368,887 | Gross: $96.90M | Top 250: #6
5	5. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,864,698 | Gross: $377.85M | Top 250: #7
6	6. The Godfather Part II (1974)	9.0	Votes: 1,284,571 | Gross: $57.30M | Top 250: #4

Mini Project 02 - Specphone phone Database

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

```
url <- read_html("https://specphone.com/Samsung-Galaxy-A30.html")
```

```
url %>%
    html_noted("div.topic") %>%
    html_text2()
```

```
ERROR: Error in html_noted(., "div.topic"): could not find function "html_noted
```

```
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-A14.html' · '/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-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-S23-Ultra-5G8-256GB.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-S23-Ultra-5G.html' ·
'/Samsung-Galaxy-S10-Plus-Ram-12GB.html' · '/Samsung-Galaxy-Z-Fold-3.html'
```

```
full_links <- paste0("https://specphone.com",links) [1:5]
full_links
```

```
'https://specphone.com/Samsung-Galaxy-A14.html' · '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'
```

```
full_links
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_node("div.detail") %>%
        html_text2()

    tmp <- data.frame(attribute = ss_topic
                      value = ss_detail)

    result <- bind_rows(result, tmp)
    print("Progress ...")
}

print(result)
```

```
ERROR: Error in parse(text = x, srcfile = src): <text>:16:23: unexpected symbol
15:     tmp <- data.frame(attribute = ss_topic
16:                           value
                              ^
```

```
write_csv(result, {result_ss_phone.csv})
```