

Mini Project 01 - IMDB web scraping

```
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 1.3.1

✓ 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"
```

```
# 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 .
```

```
# movie title
imbd %>%
  html_node("h3.lister-item-header") %>%
  html_text2()
```

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

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

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

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

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

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

```
num_votes[1:10]
```

```
'Votes: 2,680,710 | Gross: $28.34M | Top 250: #1' · 'Votes: 1,858,497 | Gross: $134.97M | Top 250: #2' ·
'Votes: 2,654,029 | Gross: $534.86M | Top 250: #3' · 'Votes: 1,847,302 | Gross: $377.85M | Top 250: #7' ·
'Votes: 1,272,110 | Gross: $57.30M | Top 250: #4' · 'Votes: 1,356,500 | Gross: $96.90M | Top 250: #6' ·
'Votes: 791,905 | Gross: $4.36M | Top 250: #5' · 'Votes: 2,056,001 | Gross: $107.93M | Top 250: #8' ·
'Votes: 1,876,685 | Gross: $315.54M | Top 250: #9' · 'Votes: 2,126,247 | Gross: $37.03M | Top 250: #12'
```

```
#build a dataset
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: 2,680,710 Gross: \$28.34M Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,858,497 Gross: \$134.97M Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,654,029 Gross: \$534.86M Top 250: #3
4	4. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,847,302 Gross: \$377.85M Top 250: #7
5	5. The Godfather Part II (1974)	9.0	Votes: 1,272,110 Gross: \$57.30M Top 250: #4
6	6. Schindler's List (1993)	9.0	Votes: 1,356,500 Gross: \$96.90M Top 250: #6

Mini Project 02 - Specphone Phone Database

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

```
url1 <- read_html("https://specphone.com/Samsung-Galaxy-A04.html")
```

```
att <- url1 %>%
  html_nodes("div.topic") %>%
  html_text2()

value <- url1 %>%
  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)
ความละเอียดวิดีโอ	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
links <- samsung_url %>%
  html_nodes("li.mobile-brand-item a") %>%
  html_attr("href")
```

```
full_links <- paste0("https://specphone.com",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_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 ..."
[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

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

```
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)