

Mini Project 01 - IMDB web scraping

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

```
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)
```

```
# movie title
# html_text2() delete special
titles <- imdb %>%
  html_nodes("h3.lister-item-header") %>%
  html_text2()
```

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

```
.Primitive("as.double")
```

```
rating
```

```
'9.3'·'9.2'·'9.0'·'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'
```

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

```
#build a 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>	<chr>	<chr>
1	1. The Shawshank Redemption (1994)	9.3	Votes: 2,680,677 Gross: \$28.34M Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,858,473 Gross: \$134.97M Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,653,998 Gross: \$534.86M Top 250: #3
4	4. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,847,287 Gross: \$377.85M Top 250: #7
5	5. The Godfather Part II (1974)	9.0	Votes: 1,272,091 Gross: \$57.30M Top 250: #4
6	6. Schindler's List (1993)	9.0	Votes: 1,356,477 Gross: \$96.90M Top 250: #6

```
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 <- 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()
```

```
data.frame(Atrribuute = att, 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

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

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

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

```
full_links[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'
```

```
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(ss_topic, ss_detail )
  result <- bind_rows(result, tmp)
  print("Progress ...")
}
print(head(result),3)
```

```
[1] "Progress ..."
[1] "Progress ..."
```

```
[1] "Progress ..."
```

```
[1] "Progress ..."
```

```
[1] "Progress ..."
```

	ss_topic	ss_detail
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
write_csv(result, "result_ss_phone.csv")
```