

Mini project IMDB We scrabing

```
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" width .
```

```
#Movie title
titles <- imdb %>%
  html_nodes("h3.lister-item-header") %>%
  html_text2() #2 is to remove special character
```

```
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 · 8.5 · 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()
```

```
#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: 2,696,944 Gross: \$28.34M Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,871,486 Gross: \$134.97M Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,670,648 Gross: \$534.86M Top 250: #3
4	4. Schindler's List (1993)	9.0	Votes: 1,363,536 Gross: \$96.90M Top 250: #6
5	5. The Godfather Part II (1974)	9.0	Votes: 1,279,218 Gross: \$57.30M Top 250: #4
6	6. 12 Angry Men (1957)	9.0	Votes: 796,425 Gross: \$4.36M Top 250: #5

Mini Project 02

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

```
url <- read_html("https://specphone.com/Samsung-Galaxy-S23-Ultra-5G.html")
```

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

detail <- url %>%
  html_nodes("div.detail") %>%
  html_text2()
```

```
data.frame(attribute = att, value = detail)
```

A data.frame: 33 x 2

attribute	value
<chr>	<chr>
วันเปิดตัว	มกราคม 2566
วันวางจำหน่าย	ยังไม่วางจำหน่าย
ขนาด	163.40 x 78.10 x 8.90 มม.
น้ำหนัก	233 กรัม
วัสดุ	Glass front (Gorilla Glass Victus 2), glass back (Gorilla Glass Victus 2), aluminum frame
SIM	รองรับ 2 ซิมการ์ด (nano sim, nano sim)
Technology	HSPA 42.2/5.76 Mbps, LTE-A (CA), 5G
2G	850/900/1800/1900
3G	850/900/1900/2100
4G	850/900/1900/2100/2600
5G	2100/2600/3500/4700
ความเร็ว	HSPA 42.2/5.76 Mbps, LTE-A (CA), 5G
ประเภท	Dynamic AMOLED 2X
ขนาดหน้าจอ	6.80 นิ้ว
ความละเอียด	1440 x 3088 pixels
ระบบปฏิบัติการ	Android 12
ชิปประมวลผล	Qualcomm Snapdragon 8 Gen 2 SM8550 3.2 GHz
ชิปกราฟิก	Adreno 740
หน่วยความจำ	12 GB
ความจุ	512/1024 GB
Memory Card	ไม่รองรับ
กล้องหลัก	ตัวที่ 1: 200 MP, f/1.7, 23mm (wide), 1/1.3 ตัวที่ 2: 10 MP, f/4.9, 230mm (periscope telephoto), 1/3.52 ตัวที่ 3: 10 MP, f/2.4, 70mm (telephoto), 1/3.52 ตัวที่ 4: 12 MP, f/2.2, 13mm, 120° (ultrawide), 1/2.55
ความละเอียดวิดีโอ	8K@24/30fps, 4K@30/60fps, 1080p@30/60/240fps, 720p@960fps, HDR10+, stereo sound rec., gyro-EIS
กล้องหน้า	ตัวที่ 1: 12 MP, f/2.2, 25mm (wide), PDAF
Bluetooth	5.3, A2DP, LE
Wi-Fi	802.11 a/b/g/n/ac/6e, tri
USB	Type-C
GPS	GPS, GLONASS, BDS, GALILE
NFC	รองรับ
ความจุ	5,000 mAh
ประเภท	Non-removable Li-Po Batt
Wireless Charging	รองรับ
Fast Charging	รองรับ (45W)

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

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

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

```
full_links
```

'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' ·
'https://specphone.com/Samsung-Galaxy-Young-2.html' ·
'https://specphone.com/Samsung-Galaxy-M02.html' · 'https://specphone.com/Samsung-Galaxy-A11.html' ·
'https://specphone.com/Samsung-Galaxy-J2-Pro-2018.html' ·
'https://specphone.com/Samsung-Galaxy-A12-2021.html' ·
'https://specphone.com/Samsung-Galaxy-A21s-3-32GB.html' ·
'https://specphone.com/Samsung-Galaxy-J5.html' · 'https://specphone.com/Samsung-Galaxy-J4.html' ·
'https://specphone.com/Samsung-Galaxy-Core-2-Duos.html' ·
'https://specphone.com/Samsung-Galaxy-Ace-Plus.html' ·
'https://specphone.com/Samsung-Galaxy-A20.html' · 'https://specphone.com/Samsung-Galaxy-Chat.html' ·
'https://specphone.com/Samsung-Galaxy-Gio.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-A7-Lite-LTE.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-A-10.5WIFI.html' ·
'https://specphone.com/Samsung-Galaxy-Alpha.html' ·
'https://specphone.com/Samsung-Galaxy-S3-Slim.html' ·
'https://specphone.com/Samsung-Galaxy-S4-zoom.html' ·
'https://specphone.com/Samsung-Galaxy-Xcover-2.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-8.9-3G-16GB.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-A8-LTE-2021.html' ·
'https://specphone.com/Samsung-Galaxy-A8-2018.html' ·
'https://specphone.com/Samsung-Galaxy-Tab4-8.0-wifi.html' ·
'https://specphone.com/Samsung-Galaxy-M33-5G.html' ·
'https://specphone.com/Samsung-Galaxy-A50.html' · 'https://specphone.com/Samsung-Galaxy-E7.html' ·
'https://specphone.com/Samsung-Galaxy-S6.html' ·
'https://specphone.com/Samsung-Galaxy-S20-FE.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-S4-WIFI.html' ·
'https://specphone.com/Samsung-Galaxy-S7.html' ·
'https://specphone.com/Samsung-Galaxy-Note-5-Exynos.html' ·
'https://specphone.com/Samsung-Galaxy-TabPRO-12.2-LTE.html' ·
'https://specphone.com/Samsung-Galaxy-S4-Active.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-Active-3.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-S3-9.7.html' ·
'https://specphone.com/Samsung-Galaxy-S6-edge.html' ·
'https://specphone.com/Samsung-Galaxy-Note-4-Exynos.html' ·
'https://specphone.com/Samsung-Galaxy-Round.html' ·
'https://specphone.com/Samsung-Galaxy-Note-20-Ultra-5G.html' ·
'https://specphone.com/Samsung-ATIV-Q.html' ·
'https://specphone.com/Samsung-ATIV-Smart-PC-PRO.html' ·
'https://specphone.com/Samsung-Galaxy-S23-Ultra-5G8-256GB.html' ·
'https://specphone.com/Samsung-Galaxy-S22-Ultra12-128GB.html' ·
'https://specphone.com/Samsung-Galaxy-Z-Flip-5G.html' ·
'https://specphone.com/Samsung-Galaxy-Z-Flip.html' ·
'https://specphone.com/Samsung-Galaxy-Tab-S8-Ultra-5G.html' ·
'https://specphone.com/Samsung-Galaxy-S21-Ultra-16-512GB.html' ·
'https://specphone.com/Samsung-Galaxy-S23-Ultra-5G.html' ·
'https://specphone.com/Samsung-Galaxy-S10-Plus-Ram-12GB.html' ·
'https://specphone.com/Samsung-Galaxy-Z-Fold-3.html'

```

result <- data.frame()

for (links in full_links[1:10]) {
  ss_topic <- links %>%
    read_html %>%
      html_nodes("div.topic") %>%
      html_text2()

  ss_detail <- links %>%
    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
9   3G

```

```

#write CSV
write_csv(result, "Result_ss.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)