

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"
```

```
# Read HTML
```

```
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()
```

```
titles[1:10]
```

```
'1. The Shawshank Redemption (1994)' · '2. The Godfather (1972)' · '3. The Dark Knight (2008)' ·
'4. The Godfather Part II (1974)' · '5. Schindler's List (1993)' · '6. 12 Angry Men (1957)' ·
'7. The Lord of the Rings: The Return of the King (2003)' · '8. Pulp Fiction (1994)' ·
'9. The Lord of the Rings: The Fellowship of the Ring (2001)' · '10. Inception (2010)'
```

```
# Rating
```

```
ratings <- imdb %>%
  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 votes
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: 2,706,278 Gross: \$28.34M Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,879,311 Gross: \$134.97M Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,679,766 Gross: \$534.86M Top 250: #3
4	4. The Godfather Part II (1974)	9.0	Votes: 1,283,331 Gross: \$57.30M Top 250: #4
5	5. Schindler's List (1993)	9.0	Votes: 1,367,635 Gross: \$96.90M Top 250: #6
6	6. 12 Angry Men (1957)	9.0	Votes: 799,482 Gross: \$4.36M Top 250: #5

SpecPhone Phone Database

```
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/Xiaomi-Poco-X5-5G.html")
```

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

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

A data.frame: 32 × 2

attribute	value
<chr>	<chr>
วันเปิดตัว	กุมภาพันธ์ 2566
วันวางจำหน่าย	วางจำหน่ายแล้ว
ขนาด	165.90 x 76.20 x 8.00 มม.
น้ำหนัก	189 กรัม
วัสดุ	Glass front (Gorilla Glass 5), glass back, plastic frame
SIM	รองรับ 2 ซิมการ์ด (nano sim, nano sim)
Technology	HSPA, 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, LTE-A (CA), 5G
ประเภท	AMOLED
ขนาดหน้าจอ	6.67 นิ้ว
ความละเอียด	1080 x 2400 pixels
ระบบปฏิบัติการ	Android 12
ชิปประมวลผล	Qualcomm Snapdragon 695 5G SM6375 2.2 GHz
ชิปกราฟิก	Adreno 619
หน่วยความจำ	6 GB
ความจุ	128 GB
Memory Card	ไม่รองรับ
กล้องหลัก	ตัวที่ 1: 48 MP, f/1.8, (wide), PDAF ตัวที่ 2: 8 MP, f/2.2, 118° (ultrawide), 1/4 ตัวที่ 3: 2 MP, f/2.4, (macro)
ความละเอียดวิดีโอ	1080p@30fps
กล้องหน้า	ตัวที่ 1: 13 MP, f/2.5, (wide)
Bluetooth	5.1, A2DP, LE
Wi-Fi	802.11 a/b/g/n/ac, dual-b
USB	Type-C
GPS	GPS
NFC	รองรับ
ความจุ	5,000 mAh
ประเภท	Non-removable Li-Po Batt
Fast Charging	รองรับ (33W)

```
# All Samsung's smartphones

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

```
# Links to all Samsung smart phones

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 (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(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..."

```

```
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)

```

# Write CSV

write_csv(result, "result_ss_phone.csv")

```