Web scraping by R

- IMDB Website

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

Get data from html

- Use read_html from `rvest` to read a document from url link
- Find and get data from all node from header and class name and then by using `html_nodes`

```
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
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" widt .
```

```
# movie title node
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 Lord of the Rings: The Return of the King (2003)' · '5. Schindler\'s List (1993)' ·
'6. The Godfather Part II (1974)' · '7. 12 Angry Men (1957)' · '8. Pulp Fiction (1994)' · '9. Inception (2010)' ·
'10. The Lord of the Rings: The Two Towers (2002)'

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

num_votes[1:10]
```

'Votes: 2,665,497 | Gross: $28.34M | Top 250: #1' · 'Votes: 1,847,169 | Gross: $134.97M | Top 250: #2' ·
'Votes: 2,638,482 | Gross: $534.86M | Top 250: #3' · 'Votes: 1,837,604 | Gross: $377.85M | Top 250: #7' ·
'Votes: 1,349,699 | Gross: $96.90M | Top 250: #6' · 'Votes: 1,265,096 | Gross: $57.30M | Top 250: #4' ·
'Votes: 787,189 | Gross: $4.36M | Top 250: #5' · 'Votes: 2,040,310 | Gross: $107.93M | Top 250: #8' ·
'Votes: 2,338,167 | Gross: $292.58M | Top 250: #14' · 'Votes: 1,659,271 | Gross: $342.55M | Top 250: #13'

```
# 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: 2,665,497 | Gross: $28.34M | Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,847,169 | Gross: $134.97M | Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,638,482 | Gross: $534.86M | Top 250: #3
4	4. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,837,604 | Gross: $377.85M | Top 250: #7
5	5. Schindler's List (1993)	9.0	Votes: 1,349,699 | Gross: $96.90M | Top 250: #6
6	6. The Godfather Part II (1974)	9.0	Votes: 1,265,096 | Gross: $57.30M | Top 250: #4

Web scraping

- Specphone website
- Create a phone database

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

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

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

topic
```

'วันเปิดตัว' · 'วันวางจำหน่าย' · 'ขนาด' · 'น้ำหนัก' · 'วัสดุ' · 'SIM' · 'Technology' · '2G' · '3G' · '4G' · '5G' · 'ความเร็ว' · 'ประเภท' · 'ขนาดหน้าจอ' · 'ความละเอียด' · 'ระบบปฏิบัติการ' · 'ชิปประมวลผล' · 'ชิปกราฟิก' · 'หน่วยความจำ' · 'ความจุ' · 'Memory Card' · 'กล้องหลัก' · 'ความละเอียดวีดีโอ' · 'กล้องหน้า' · 'Bluetooth' · 'Wi-Fi' · 'USB' · 'GPS' · 'NFC' · 'ความจุ' · 'ประเภท'

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

value
```

'ตุลาคม 2565' · 'ยังไม่วางจำหน่าย' · '164.40 x 76.30 x 9.10 มม.' · '192 กรัม' · 'Glass front, plastic back, plastic frame' · 'รองรับ 2 ซิมการ์ด (nano sim, nano sim)' · 'HSPA 42.2/5.76 Mbps, LTE-A' · '850/900/1800/1900' · '850/900/1900/2100' · '850/900/1900/2100/2600' · '-' · 'HSPA 42.2/5.76 Mbps, LTE-A' · 'PLS LCD' · '6.50 นิ้ว' · '720 x 1600 pixels' · 'Android 12' · 'Spreadtrum Unisoc SC9863A 1.6 GHz' · 'PowerVR GE8322' · '3 GB' · '32 GB' · 'microSD (1)' · 'ตัวที่ 1: 50 MP, f/1.8, (wide), AF\nตัวที่ 2: 2 MP, f/2.4, (depth)' · '1080p@30fps' · 'ตัวที่ 1: 5 MP, f/2.2' · '5.0, A2DP, LE' · '802.11 a/b/g/n/ac, dual-b' · 'Type-C' · 'GLONASS, GALILEO, BDS' · 'ไม่รองรับ' · '5,000 mAh' · 'Non-removable Li-Po Batt'

```
df = data.frame(attribute = topic, value = value)

df
```

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)
ความละเอียดวีดีโอ	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")
```

```
# link to all samsung smartphone
links <- samsung_url %>%
    html_nodes("li.mobile-brand-item a") %>% # link is in a li box class mobile-b
    html_attr("href") # get link href in sub box a

links
```

'/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-M02.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-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-S10-Plus-Ram-12GB.html' · '/Samsung-Galaxy-Z-Fold-3.html' · '/Samsung-Galaxy-Z-Fold4.html' · '/Samsung-Galaxy-Z-Fold-2-5G.html'

```
# create a full link of all samsung smartphone website
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-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-S10-Plus-Ram-12GB.html' · 'https://specphone.com/Samsung-Galaxy-Z-Fold-3.html' · 'https://specphone.com/Samsung-Galaxy-Z-Fold4.html' · 'https://specphone.com/Samsung-Galaxy-Z-Fold-2-5G.html'

```
# Create a dataframe of all samsung smart phone
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(attributes = ss_topic,
                      value = ss_detail)

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

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

```
print(head(result), 3)
```

```
   attributes                                        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")
```