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

Mini Project 01 - IMDB web scraping

```
url <- "https://www.imdb.com/search/title/?groups=top_100&sort=user_rating,desc"
```

```
print(url)
```

```
function (description, open = "", blocking = TRUE, encoding = getOption("encodi
    method = getOption("url.method", "default"), headers = NULL)
{
    method <- match.arg(method, c("default", "internal", "libcurl",
        "wininet"))
    if (!is.null(headers)) {
        nh <- names(headers)
        if (length(nh) != length(headers) || any(nh == "") ||
            anyNA(headers) || anyNA(nh))
            stop("'headers' must have names and must not be NA")
        headers <- paste0(nh, ": ", headers)
        headers <- list(headers, paste0(headers, "\r\n", collapse = ""))
    }
    .Internal(url(description, open, blocking, encoding, method,
        headers))
}
<bytecode: 0x5568e2074aa8>
<environment: namespace:base>
```

```
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
titles <- imdb %>%
    html_nodes("h3.lister-item-header")  %>% # node, nodes choose to use!
    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()
```

```
rating[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 = rating,
    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: 2,704,786 | Gross: $28.34M | Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,878,119 | Gross: $134.97M | Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,678,499 | Gross: $534.86M | Top 250: #3
4	4. The Godfather Part II (1974)	9.0	Votes: 1,282,718 | Gross: $57.30M | Top 250: #4
5	5. Schindler's List (1993)	9.0	Votes: 1,367,009 | Gross: $96.90M | Top 250: #6
6	6. 12 Angry Men (1957)	9.0	Votes: 799,026 | Gross: $4.36M | Top 250: #5

Mini Project 02 - Specphone Phone Database

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

```
url <- read_html("https://specphone.com/Samsung-Galaxy-A04s.html")
```

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

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

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

A data.frame: 31 × 2

attributes	value
<chr>	<chr>
วันเปิดตัว	ตุลาคม 2565
วันวางจำหน่าย	ยังไม่วางจำหน่าย
ขนาด	164.70 x 76.70 x 9.10 มม.
น้ำหนัก	196 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	Samsung Exynos 850 S5E3830 2 GHz
ชิปกราฟิก	Mali-G52 MP1
หน่วยความจำ	4 GB
ความจุ	64 GB
Memory Card	microSD (1)
กล้องหลัก	ตัวที่ 1: 50 MP, f/1.8, (wide), PDAF ตัวที่ 2: 2 MP, f/2.4, (macro) ตัวที่ 3: 2 MP, f/2.4, (depth)
ความละเอียดวีดีโอ	1080p@30fps
กล้องหน้า	ตัวที่ 1: 5 MP, f/2.2, (wide)
Bluetooth	5.0, A2DP, LE
Wi-Fi	802.11 a/b/g/n/ac, dual-b
USB	Type-C
GPS	A-GPS, GLONASS, BDS, GALI
NFC	รอบรับ
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")
```

```
links[1]
```

'/Samsung-Galaxy-M13.html'

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

```
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))
```

```
    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_phone1.csv")
```