Web Scraping from IMDb

```
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"
```

```
# 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's titles
titles <- imdb %>%
    html_nodes("h3.lister-item-header") %>%
    html_text2()
```

```
title[1:10]
```

'1. The Shawshank Redemption (1994)' · '2. The Godfather (1972)' · '3. The Dark Knight (2008)' ·
'4. Schindler\'s List (1993)' · '5. The Lord of the Rings: The Return of the King (2003)' ·
'6. The Godfather Part II (1974)' · '7. 12 Angry Men (1957)' · '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: 2,707,067 | Gross: $28.34M | Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,879,951 | Gross: $134.97M | Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,680,405 | Gross: $534.86M | Top 250: #3
4	4. Schindler's List (1993)	9.0	Votes: 1,368,007 | Gross: $96.90M | Top 250: #6
5	5. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,863,464 | Gross: $377.85M | Top 250: #7
6	6. The Godfather Part II (1974)	9.0	Votes: 1,283,698 | Gross: $57.30M | Top 250: #4

Xiaomi Smartphones (SpecPhone)

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

```
url <- read_html("https://specphone.com/Xiaomi-Redmi-10A.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: 31 × 2

attribute	value
<chr>	<chr>
วันเปิดตัว	**มกราคม 2566**
วันวางจำหน่าย	**ยังไม่วางจำหน่าย**
ขนาด	164.90 x 77.10 x 9.00 มม.
น้ำหนัก	194 กรัม
วัสดุ	**ไม่รองรับ**
SIM	**รองรับ** 2 **ซิมการ์ด** (nano sim, nano sim)
Technology	HSPA, LTE
2G	850/900/1800/1900
3G	850/900/1900/2100
4G	850/900/1900/2100/2600
5G	-
ความเร็ว	HSPA, LTE
ประเภท	IPS LCD
ขนาดหน้าจอ	6.53 นิ้ว
ความละเอียด	720 x 1600 pixels
ระบบปฏิบัติการ	Android 11
ชิปประมวลผล	MediaTek Helio G25 2 GHz
ชิปกราฟิก	PowerVR GE8320
หน่วยความจำ	3 GB
ความจุ	32 GB
Memory Card	**ไม่รองรับ**
กล้องหลัก	**ตัวที่** 1: 13 MP, f/2.2, (wide), 1.0µm, AF
ความละเอียดวีดีโอ	1080p@30fps
กล้องหน้า	**ตัวที่** 1: 5 MP
Bluetooth	5.0, A2DP, LE
Wi-Fi	802.11 a/b/g/n/ac, dual-b
USB	micro USB
GPS	GPS, GLONASS, GALILEO, BD
NFC	**ไม่รองรับ**
ความจุ	5,000 mAh
ประเภท	Non-removable Li-Po Batt

```
#All Xiaomi Smartphones
xiaomi_url <- read_html("https://specphone.com/brand/Xiaomi")
```

```
# Links to all Xiaomi smartphones
links <- xiaomi_url %>%
    html_nodes("li.mobile-brand-item a") %>%
    html_attr("href")
```

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

```
result <- data.frame()

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

    xm_detail <- link %>%
    read_html() %>%
    html_nodes("div.detail") %>%
    html_text2()

    tmp <- data.frame(attribute = xm_topic,
                      value = xm_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
```

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

```
    attribute                        value
1     วันเปิดตัว                   ตุลาคม 2565
2 วันวางจำหน่าย                ยังไม่วางจำหน่าย
3        ขนาด  250.50 x 158.10 x 7.10 มม.
4       น้ำหนัก                      465 กรัม
5         วัสดุ Glass front, aluminum frame
6         SIM
```

```
# write csv
write_csv(result, "result_xm_phones.csv")
```