Author : NanT.

Mini Project 01 - 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

✓ 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
```

```r
# read html
imdb <- read_html(url)
```

```r
imdb
```

```
{html_document}
<html xmlns:og="http://ogp.me/ns#" xmlns:fb="http://www.facebook.com/2008/fb
[1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-
[2] <body id="styleguide-v2" class="fixed">\n              <img height="1" wid
```

```r
# movie title
imdb %>%
    html_node("h3.lister-item-header") %>%
    html_text()  #html_text2() จะลบอักขระแปลกๆทิ้งไป
```

'\n   1.\n \n   The Shawshank Redemption\n   (1994)\n'

```r
# all movie title at this page (nodes)
titles <- imdb %>%
    html_nodes("h3.lister-item-header") %>%
    html_text2()
```

```r
# rating box
imdb %>%
    html_node("div.ratings-bar") %>%
    html_text2()
```

'9.3\nRate this\n 1 2 3 4 5 6 7 8 9 10 9.3/10 X \n81 Metascore'

```r
# rating only
imdb %>%
    html_node("div.ratings-imdb-rating") %>%
    html_text2()
```

'9.3'

```
# rating 1-10
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()
```

```
# 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: 2,677,998 | Gross: $28.34M | Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,856,459 | Gross: $134.97M | Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,651,336 | Gross: $534.86M | Top 250: #3
4	4. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,845,465 | Gross: $377.85M | Top 250: #7
5	5. Schindler's List (1993)	9.0	Votes: 1,355,383 | Gross: $96.90M | Top 250: #6
6	6. The Godfather Part II (1974)	9.0	Votes: 1,270,903 | Gross: $57.30M | Top 250: #4

Mini Project 02 - Specphone Phone Database

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

```
url <- read_html("https://specphone.com/Samsung-Galaxy-A30.html")
```

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

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

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

A data.frame: 32 × 2

attribute	value
<chr>	<chr>
วันเปิดตัว	เมษายน 2562
วันวางจำหน่าย	เมษายน 2562, วางจำหน่ายแล้ว
ขนาด	157.50 x 74.70 x 7.70 มม.
น้ำหนัก	ไม่รองรับ
วัสดุ	Front glass, plastic body
SIM	รองรับ 2 ซิมการ์ด (nano sim, nano sim)
Technology	EDGE,, HSPA+, LTE Cat.9
2G	850/900/1800/1900
3G	850/900/1900/2100
4G	850/900/1900/2100
5G	-
ความเร็ว	EDGE,, HSPA+, LTE Cat.9
ประเภท	
ขนาดหน้าจอ	6.40 นิ้ว
ความละเอียด	2340 x 1080 pixels
ระบบปฏิบัติการ	Android 9
ชิปประมวลผล	Samsung Exynos 7 7904 1.6 GHz
ชิปกราฟิก	Mali-G71 MP2
หน่วยความจำ	4 GB
ความจุ	64 GB
Memory Card	microSD (512)
กล้องหลัก	ตัวที่ 1: 16 MP ตัวที่ 2: 5 MP
ความละเอียดวีดีโอ	1080p
กล้องหน้า	ตัวที่ 1: 16 MP
Bluetooth	5.0, A2DP, LE
Wi-Fi	802.11 a/b/g/n/ac
USB	Type-C
GPS	GPS,A-GPS, GLONASS, BDS
NFC	ไม่รองรับ
ความจุ	4,000 mAh
ประเภท	Li - Polymer
Fast Charging	รองรับ (Fast battery charging 15W)

All SS smartphones

```
# all samsung smartphones
ss_url <- read_html("https://specphone.com/brand/Samsung")
```

```
links <- ss_url %>%
    html_nodes("li.mobile-brand-item a") %>% # หา a ที่อยู่ในกล่อง li
    html_attr("href") #ดึงลิ้งค์ใน a
```

```
links #link ยังกดไม่ได้ ต้องมี http
```

'/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'

```
# เพิ่ม web
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'

```r
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 ...")
}

print(result)
```

```
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
        attributes
1        วันเปิดตัว
2      วันวางจำหน่าย
3            ขนาด
4          น้ำหนัก
5            วัสดุ
6             SIM
7      Technology
8              2G
9              3G
```

```r
# write csv
write_csv(result, "ss_phone.csv")
```

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

```
  attributes                    value
1    วันเปิดตัว              มิถุนายน 2565
2 วันวางจำหน่าย          ยังไม่วางจำหน่าย
3       ขนาด 165.40 x 76.90 x 8.40 มม.
```