

Mini Project01 - 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"
```

```
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" width="1" alt="Facebook logo" data-bbox="125 225 145 245" />
```

```
#movie title
titles <-imdb %>%
  html_nodes("h3.lister-item-header") %>%          #html_node("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
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 data frame

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: 2,664,096 Gross: \$28.34M Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,846,148 Gross: \$134.97M Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,637,049 Gross: \$534.86M Top 250: #3
4	4. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,836,721 Gross: \$377.85M Top 250: #7
5	5. Schindler's List (1993)	9.0	Votes: 1,349,035 Gross: \$96.90M Top 250: #6
6	6. The Godfather Part II (1974)	9.0	Votes: 1,264,515 Gross: \$57.30M Top 250: #4

Mini Project02 - Specphone on database

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

```
url <- read_html("https://specphone.com/Samsung-Galaxy-A04.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 x 2

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

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

```
#links to all samsung smartphone
links <- samsung_url %>%
  html_nodes("li.mobile-brand-item a")%>% #เข้าไปหา a ใน li
  html_attr("href")
```

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'
```

```
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'

```

result <- data.frame()

for (link in full_links[1:10]) {    ##link ที่อยู่ใน Full link
  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..."
[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
9      3G

```

#NYC flights 2013 Analysis


```
library(dplyr)
```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

filter, lag

The following objects are masked from 'package:base':

intersect, setdiff, setequal, union

1.## Q1 : How many flights in May 2013

สายการบินไหน Popular

สายการบินไหนบินเลขที่สุด

ระยะเวลาในการบิน

Point คือต้องการใช้ Function ที่เราเรียน