

Mini Project 01 - IMDB web sraping

Ctrl + m ใช้เปลี่ยนประเภทของ cell ว่าจะให้เป็น code หรือ note

```
library(tidyverse) #
library(rvest) #scraape data from internet ยังไม่เหมาะกับเว็บที่ต้องมีการล็อกอิน

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%2Cdesc"
```

```
print(url)
```

```
[1] "https://www.imdb.com/search/title/?groups=top_100&sort=user_rating%2Cdesc"
```

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

```
{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 .
```

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

# ถ้า node ไม่มี s มันจะวิ่งไปหา box แรกที่มันเจอ
# ถ้าใช้ text() มันจะโชว์อักขระพิเศษ ถ้า text2() มันจะไม่โชว์อักขระพิเศษ
```

```
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() %>%      #convert to numeric
  as.numeric()
```

```
# number of votes
num_votes <- imdb %>%
  html_nodes("p.sort-num_votes-visible") %>%
  html_text2()
```

```
#build a dataset
# สุดท้ายแล้วเราต้องการนำ data แต่ละส่วนที่เราดึงมารวมกัน เพื่อนำไปวิเคราะห์ต่อ
df <- data.frame(
  title=titles,
  rating = ratings,
  numvote = num_votes
)
head(df)
```

A data.frame: 6 × 3

	title	rating	numvote
	<chr>	<dbl>	<chr>
1	1. The Shawshank Redemption (1994)	9.3	Votes: 2,657,665 Gross: \$28.34M Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,841,944 Gross: \$134.97M Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,630,432 Gross: \$534.86M Top 250: #3
4	4. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,832,422 Gross: \$377.85M Top 250: #7
5	5. Schindler's List (1993)	9.0	Votes: 1,346,191 Gross: \$96.90M Top 250: #6
6	6. The Godfather Part II (1974)	9.0	Votes: 1,261,942 Gross: \$57.30M Top 250: #4

วิธีการดึงข้อมูลแบบนี้ ไม่เหมาะกับเว็บที่ต้องมีการ log in และเหมาะกับเว็บที่ไม่ได้มีลูกเล่นในการเขียนเว็บมาก

Mini Project 02- Specphone phone

```
library(tidyverse) #
library(rvest) #scraape data from internet ยังไม่เหมาะกับเว็บที่ต้องมีการล็อกอิน
```

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

```
# function ตัวแรกต้องเป็นฟังก์ชันที่เราใช้ในการดึงข้อมูลขึ้นมาก่อน : read_html
# ไปดูว่าแต่ละกล่องนะ มันใช้โค้ดอะไร และมีแพทเทินอะไรใช้ซ้ำ ๆ กันทุกกล่องไหม

att <- url %>%
  #read_html() %>% # ซึ่งถ้าเราเอาไปฝากไว้ด้านบนแล้ว เราก็ลบด้านล่างนี้ทิ้งไปได้
  html_nodes("div.topic") %>% # ถ้า node ไม่มี s มันจะดึงมาแค่ตัวแรกตัวเดียว
  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	164.40 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)
ความละเอียดวิดีโอ	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 Smart phone
```

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

```
# link to all samsung smartphone
links <- samsung_url %>%
  html_nodes("li.mobile-brand-item a") %>%      # จงหาตัว a ที่อยู่ภายใน li.mobile-br
  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-S23-Ultra-5G8-256GB.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-S23-Ultra-5G.html' ·
'/Samsung-Galaxy-S10-Plus-Ram-12GB.html' · '/Samsung-Galaxy-Z-Fold-3.html'
```

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

```

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(attribute = ss_topic,
                    value = ss_detail)

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

print(result)

# เขียนด้วย R studio น่าจะง่ายกว่า

```

```

print(head(result),3) # print results แด่3แถวแรก

```

	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 in CSV น่าจะดูง่ายกว่า
write_csv(result, "result_ss_phone.csv")

```