

*Mini Project 01 - TOP IMDB

```
#install.packages("tidyverse") -> for program r
#install.packages("rvest") -> for program r
library(tidyverse)
library(rvest) # scrape data from internet
```

```
url <- "https://www.imdb.com/search/title/?groups=top_100&sort=user_rating%2Cd
```

```
print(url)
```

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

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

```
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
```

```
#movie title
titles <- imdb %>%
  html_nodes("h3,list-item-header")%>% #node คือ ดึงแค่ชื่อแรก nodes คือ ดึงทุกชื่อ
  html_text2()%>% #text ดึงอีกขระแปลกๆมาด้วย text2 ดึงแค่ชื่อ
  head
```

```
titles[1:10]
```

'1. The Shawshank Redemption (1994)' · '2. The Godfather (1972)' · '3. The Dark Knight (2008)' ·
'4. Schindler's List (1993)' · '5. The Godfather Part II (1974)' · '6. 12 Angry Men (1957)' · NA · NA · NA · NA

```
#rating
ratings <- imdb%>%
  html_nodes("div.ratings-imdb-rating") %>%
  html_text2()%>%
  as.numeric()%>%
  head
```

```
ratings[1:10]
```

```
9.3 · 9.2 · 9 · 9 · 9 · 9 · <NA> · <NA> · <NA> · <NA>
```

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

```
num_votes
```

```
'Votes: 2,699,167 | Gross: $28.34M | Top 250: #1' · 'Votes: 1,873,368 | Gross: $134.97M | Top 250: #2' ·  
'Votes: 2,672,895 | Gross: $534.86M | Top 250: #3' · 'Votes: 1,364,525 | Gross: $96.90M | Top 250: #6' ·  
'Votes: 1,280,233 | Gross: $57.30M | Top 250: #4' · 'Votes: 797,150 | Gross: $4.36M | Top 250: #5'
```

```
#build a database
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: 2,699,167 Gross: \$28.34M Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,873,368 Gross: \$134.97M Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,672,895 Gross: \$534.86M Top 250: #3
4	4. Schindler's List (1993)	9.0	Votes: 1,364,525 Gross: \$96.90M Top 250: #6
5	5. The Godfather Part II (1974)	9.0	Votes: 1,280,233 Gross: \$57.30M Top 250: #4
6	6. 12 Angry Men (1957)	9.0	Votes: 797,150 Gross: \$4.36M Top 250: #5

df

A data.frame: 51 × 1

title
<chr>
1. The Shawshank Redemption (1994)
2. The Godfather (1972)
3. The Dark Knight (2008)
4. Schindler's List (1993)
5. The Godfather Part II (1974)
6. 12 Angry Men (1957)
7. The Lord of the Rings: The Return of the King (2003)
8. Pulp Fiction (1994)
9. Inception (2010)
10. Fight Club (1999)
11. The Lord of the Rings: The Fellowship of the Ring (2001)
12. Forrest Gump (1994)
13. Il buono, il brutto, il cattivo (1966)
14. The Lord of the Rings: The Two Towers (2002)
15. GoodFellas (1990)
16. The Matrix (1999)
17. One Flew Over the Cuckoo's Nest (1975)
18. The Empire Strikes Back (1980)
19. Interstellar (2014)
20. Se7en (1995)
21. The Silence of the Lambs (1991)
22. Saving Private Ryan (1998)
23. The Green Mile (1999)
24. Star Wars (1977)
25. La vita è bella (1997)
26. Terminator 2: Judgment Day (1991)
27. Sen to Chihiro no kamikakushi (2001)
28. Cidade de Deus (2002)
29. It's a Wonderful Life (1946)
30. Shichinin no samurai (1954)
31. Seppuku (1962)
32. Whiplash (2014)
33. Gisaengchung (2019)
34. Gladiator (2000)
35. The Departed (2006)
36. Léon (1994)
37. The Prestige (2006)
38. Apocalypse Now (1979)
39. Back to the Future (1985)
40. Alien (1979)
41. The Lion King (1994)
42. The Usual Suspects (1995)

43. American History X (1998)
44. The Pianist (2002)
45. Casablanca (1942)
46. The French Lieutenant's Woman (2011)
47. Psycho (1960)
48. Once Upon a Time in the West (1968)
49. Hotaru no haka (1988)
50. Rear Window (1954)

****Mini Project 02 - Specphone Phone project**

```
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(attitude= att,value=value)
```

A tibble: 31 × 2

attitude	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

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

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

```
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(attitude = ss_topic,
                    value=ss_detail)

  result <- bind_rows(result,tmp)
  print("Progress ...")
}
print(head(result),3)
```

```
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
      attitude                                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_phone.csv")
```