

# Mini Project 01 - IMDB Web Scrapping

```
library(tidyverse) # Prepare DataSet
library(rvest)
```

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'

```
url1 <- "https://www.imdb.com/search/title/?groups=top_100&sort=user_rating%2Cdesc"
```

```
print(url1)
```

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

```
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 .
```

```
imdb %>%  
  html_node("h3.lister-item-header") %>%  
  html_text()
```

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

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

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

```
ratings
```

```
9.3·9.2·9·9·9·9·9·9·8.9·8.8·8.8·8.8·8.8·8.8·8.7·8.7·8.7·8.7·8.6·8.6·8.6·8.6·8.6·8.6·8.6·8.6·8.6·  
8.6·8.6·8.6·8.6·8.5·8.5·8.5·8.5·8.5·8.5·8.5·8.5·8.5·8.5·8.5·8.5·8.5·8.5·8.5·8.5·8.5·8.5
```

```
# Number Of Votes  
imdb %>%  
  html_nodes("p.sort-num_votes-visible") %>%  
  html_text2() -> num_votes
```

```
# 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: 2,658,462   Gross: \$28.34M   Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,842,443   Gross: \$134.97M   Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,631,280   Gross: \$534.86M   Top 250: #3
4	4. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,833,020   Gross: \$377.85M   Top 250: #7
5	5. Schindler's List (1993)	9.0	Votes: 1,346,554   Gross: \$96.90M   Top 250: #6
6	6. The Godfather Part II (1974)	9.0	Votes: 1,262,269   Gross: \$57.30M   Top 250: #4

df

A data.frame: 50 × 3

title	rating	num_vote
<chr>	<dbl>	<chr>
1. The Shawshank Redemption (1994)	9.3	Votes: 2,658,462   Gross: \$28.34M   Top 250: #1
2. The Godfather (1972)	9.2	Votes: 1,842,443   Gross: \$134.97M   Top 250: #2
3. The Dark Knight (2008)	9.0	Votes: 2,631,280   Gross: \$534.86M   Top 250: #3
4. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,833,020   Gross: \$377.85M   Top 250: #7
5. Schindler's List (1993)	9.0	Votes: 1,346,554   Gross: \$96.90M   Top 250: #6
6. The Godfather Part II (1974)	9.0	Votes: 1,262,269   Gross: \$57.30M   Top 250: #4
7. 12 Angry Men (1957)	9.0	Votes: 784,944   Gross: \$4.36M   Top 250: #5
8. Pulp Fiction (1994)	8.9	Votes: 2,034,779   Gross: \$107.93M   Top 250: #8
9. Inception (2010)	8.8	Votes: 2,331,665   Gross: \$292.58M   Top 250: #14
10. The Lord of the Rings: The Two Towers (2002)	8.8	Votes: 1,655,188   Gross: \$342.55M   Top 250: #13
11. Fight Club (1999)	8.8	Votes: 2,103,589   Gross: \$37.03M   Top 250: #12
12. The Lord of the Rings: The Fellowship of the Ring (2001)	8.8	Votes: 1,861,807   Gross: \$315.54M   Top 250: #9
13. Forrest Gump (1994)	8.8	Votes: 2,059,803   Gross: \$330.25M   Top 250: #11
14. Il buono, il brutto, il cattivo (1966)	8.8	Votes: 757,987   Gross: \$6.10M   Top 250: #10
15. The Matrix (1999)	8.7	Votes: 1,900,021   Gross: \$171.48M   Top 250: #16
16. Goodfellas (1990)	8.7	Votes: 1,151,903   Gross: \$46.84M   Top 250: #17
17. The Empire Strikes Back (1980)	8.7	Votes: 1,284,128   Gross: \$290.48M   Top 250: #15
18. One Flew Over the Cuckoo's Nest (1975)	8.7	Votes: 1,003,032   Gross: \$112.00M   Top 250: #18
19. Interstellar (2014)	8.6	Votes: 1,803,818   Gross: \$188.02M   Top 250: #26
20. Cidade de Deus (2002)	8.6	Votes: 753,813   Gross: \$7.56M   Top 250: #23
21. Sen to Chihiro no kamikakushi (2001)	8.6	Votes: 756,948   Gross: \$10.06M   Top 250: #31
22. Saving Private Ryan (1998)	8.6	Votes: 1,381,688   Gross: \$216.54M   Top 250: #24
23. The Green Mile (1999)	8.6	Votes: 1,291,737   Gross: \$136.80M   Top 250: #27
24. La vita è bella (1997)	8.6	Votes: 691,306   Gross: \$57.60M   Top 250: #25
25. Se7en (1995)	8.6	Votes: 1,638,049   Gross: \$100.13M   Top 250: #19
26. Terminator 2: Judgment Day (1991)	8.6	Votes: 1,092,691   Gross: \$204.84M   Top 250: #29
27. The Silence of the Lambs (1991)	8.6	Votes: 1,422,179   Gross: \$130.74M   Top 250: #22
28. Star Wars (1977)	8.6	Votes: 1,356,765   Gross: \$322.74M   Top 250: #28

29. Seppuku (1962)	8.6	Votes: 57,058   Top 250: #45
30. Shichinin no samurai (1954)	8.6	Votes: 345,224   Gross: \$0.27M   Top 250: #20
31. It's a Wonderful Life (1946)	8.6	Votes: 453,665   Top 250: #21
32. Gisaengchung (2019)	8.5	Votes: 788,655   Gross: \$53.37M   Top 250: #34
33. Whiplash (2014)	8.5	Votes: 849,286   Gross: \$13.09M   Top 250: #42
34. The Intouchables (2011)	8.5	Votes: 852,457   Gross: \$13.18M   Top 250: #44
35. The Prestige (2006)	8.5	Votes: 1,324,450   Gross: \$53.09M   Top 250: #41
36. The Departed (2006)	8.5	Votes: 1,316,723   Gross: \$132.38M   Top 250: #39
37. The Pianist (2002)	8.5	Votes: 826,431   Gross: \$32.57M   Top 250: #33
38. Gladiator (2000)	8.5	Votes: 1,490,047   Gross: \$187.71M   Top 250: #37
39. American History X (1998)	8.5	Votes: 1,117,303   Gross: \$6.72M   Top 250: #38
40. The Usual Suspects (1995)	8.5	Votes: 1,080,193   Gross: \$23.34M   Top 250: #40
41. Léon (1994)	8.5	Votes: 1,153,744   Gross: \$19.50M   Top 250: #35
42. The Lion King (1994)	8.5	Votes: 1,051,256   Gross: \$422.78M   Top 250: #36
43. Nuovo Cinema Paradiso (1988)	8.5	Votes: 260,593   Gross: \$11.99M   Top 250: #52
44. Hotaru no haka (1988)	8.5	Votes: 276,031   Top 250: #46
45. Back to the Future (1985)	8.5	Votes: 1,195,328   Gross: \$210.61M   Top 250: #30
46. Apocalypse Now (1979)	8.5	Votes: 664,785   Gross: \$83.47M   Top 250: #53
47. Alien (1979)	8.5	Votes: 877,516   Gross: \$78.90M   Top 250: #50
48. Once Upon a Time in the West (1968)	8.5	Votes: 328,782   Gross: \$5.32M   Top 250: #48
49. Psycho (1960)	8.5	Votes: 669,158   Gross: \$32.00M   Top 250: #32
50. Rear Window (1954)	8.5	Votes: 490,388   Gross: \$36.76M   Top 250: #49

## Mini Project 2: Samsung

```
library(tidyverse) # Prepare DataSet
library(rvest)
```

```
url1 <- "https://specphone.com/Samsung-Galaxy-A04.html"
ss <- read_html(url1)
```

```
ss %>%
  html_nodes("div.topic") %>%
  html_text2() -> attr
```

```
ss %>%
  html_nodes("div.detail") %>%
  html_text2() -> value
```

```
df <- data.frame(
  attribute = attr,
  value = value
)
df
```

A data.frame: 31 × 2

attribute	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

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

```
# Link To All Samsung Smart Phone
ss_url %>%
  html_nodes("li.mobile-brand-item a") %>%
  html_attr("href")-> links
```

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]) {
  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 ...")
}

```

```

[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."
[1] "Progress ..."

```

```

result

```

A data.frame: 319 × 2

attribute	value
<chr>	<chr>
วันเปิดตัว	มิถุนายน 2565
วันวางจำหน่าย	ยังไม่วางจำหน่าย
ขนาด	165.40 x 76.90 x 8.40 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
5G	-
ความเร็ว	HSPA 42.2/5.76 Mbps, LTE-A
ประเภท	PLS LCD
ขนาดหน้าจอ	6.60 นิ้ว
ความละเอียด	1080 x 2408 pixels
ระบบปฏิบัติการ	Android 12
ชิปประมวลผล	Samsung Exynos 850 S5E3830 2 GHz
ชิปกราฟิก	Mali-G52 MP1
หน่วยความจำ	4 GB
ความจุ	64 GB
Memory Card	microSD (1)
กล้องหลัก	ตัวที่ 1: 50 MP, f/1.8, (wide), PDAF ตัวที่ 2: 5 MP, f/2.2, 123° (ultrawide) ตัวที่ 3: 2 MP, f/2.4, (depth)
ความละเอียดวิดีโอ	1080p@30fps
กล้องหน้า	ตัวที่ 1: 8 MP, f/2.2, (wide)
Bluetooth	5.0, A2DP, LE
Wi-Fi	802.11 a/b/g/n/ac, dual-b
USB	Type-C
GPS	A-GPS, GLONASS, GALILEO,
NFC	รองรับ
ความจุ	5,000 mAh
:	:
ขนาด	121.40 x 62.90 x 10.70 มม.
น้ำหนัก	123 กรัม
วัสดุ	Plastic
SIM	รองรับ 2 ซิมการ์ด (micro sim, micro sim)

Technology	EDGE, HSPA
2G	850/900/1800/1900
3G	900/2100
4G	-
5G	-
ความเร็ว	EDGE, HSPA
ประเภท	TFT LCD
ขนาดหน้าจอ	4.00 นิ้ว
ความละเอียด	800 x 480 pixels
ฟีเจอร์เพิ่มเติม	สามารถให้ PC ใช้อินเทอร์เน็ตของสมาร์ทโฟนได้ด้วย Internet Tethering สามารถใช้งาน Mobile Hotspot โดยแชร์อินเทอร์เน็ตบนสมาร์ทโฟนให้กับเครื่องอื่นๆแบบไร้สายได้ สามารถใช้งานเป็น USB Mass Storage ได้เพื่อโอนหรือย้ายข้อมูลบน PC ใช้อินเตอร์เฟซ TouchWiz UX มาพร้อมกับ Android 4.4 รองรับ 2 ซิม
ระบบปฏิบัติการ	Android 4.4
ชิปประมวลผล	Dual Core 0 MHz
ชิปกราฟิก	VideoCore IV
หน่วยความจำ	512 MB
ความจุ	4 GB
Memory Card	microSD (64)
กล้องหลัก	ตัวที่ 1: 3 MP
ความละเอียดวิดีโอ	480
กล้องหน้า	ตัวที่ 1: MP
Bluetooth	4
Wi-Fi	802.11b/g/n
USB	micro USB

```
# Write csv
write_csv(result, "samsung_phone.csv")
```