Mini Projects -- Web Scraping

Contents

- IMDB

- eCommerce Website

IMDB

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

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

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

```
# movie title
# <h3 class="lister-item-header">
titles <- imdb %>%
    html_nodes("h3.lister-item-header") %>% #node no "s"...it'll read the first n
    html_text2() # test2 delete some
                 #pull texts in that sec, which are titles
                 #text includes \n \n special char. So we use text2
```

```
titles
```

'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)' · '11. Fight Club (1999)' ·
'12. The Lord of the Rings: The Fellowship of the Ring (2001)' · '13. Forrest Gump (1994)' ·
'14. Il buono, il brutto, il cattivo (1966)' · '15. The Matrix (1999)' · '16. Goodfellas (1990)' ·
'17. The Empire Strikes Back (1980)' · '18. One Flew Over the Cuckoo\'s Nest (1975)' ·
'19. Interstellar (2014)' · '20. Cidade de Deus (2002)' · '21. Sen to Chihiro no kamikakushi (2001)' ·
'22. Saving Private Ryan (1998)' · '23. The Green Mile (1999)' · '24. La vita è bella (1997)' ·
'25. Se7en (1995)' · '26. Terminator 2: Judgment Day (1991)' · '27. The Silence of the Lambs (1991)' ·
'28. Star Wars (1977)' · '29. Seppuku (1962)' · '30. Shichinin no samurai (1954)' ·
'31. It\'s a Wonderful Life (1946)' · '32. Gisaengchung (2019)' · '33. Whiplash (2014)' ·
'34. The Intouchables (2011)' · '35. The Prestige (2006)' · '36. The Departed (2006)' ·
'37. The Pianist (2002)' · '38. Gladiator (2000)' · '39. American History X (1998)' ·
'40. The Usual Suspects (1995)' · '41. Léon (1994)' · '42. The Lion King (1994)' ·
'43. Nuovo Cinema Paradiso (1988)' · '44. Hotaru no haka (1988)' · '45. Back to the Future (1985)' ·
'46. Apocalypse Now (1979)' · '47. Alien (1979)' · '48. Once Upon a Time in the West (1968)' ·
'49. Psycho (1960)' · '50. Rear Window (1954)'

```
# rating
# <div class="inline-block ratings-imdb-rating" name="ir" data-value="9.3">
ratings <- imdb %>%
    html_nodes("div.ratings-imdb-rating") %>%
    html_text2()
```

```
ratings
```

'9.3' · '9.2' · '9.0' · '9.0' · '9.0' · '9.0' · '9.0' · '8.9' · '8.8' · '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' · '8.5' · '8.5'

```
# num of votes
#<p class="sort-num_votes-visible">
#                <span class="text-muted">Votes:</span>
#                <span name="nv" data-value="2678441">2,678,441</span>
#    <span class="ghost">|</span>
num_votes <- imdb %>%
    html_nodes("p.sort-num_votes-visible") %>%
    html_text2()

#span
num_votes_only <- imdb %>%
    #html_nodes("p.sort-num_votes-visible") %>%
    #html_text2()node, "span"
    html_nodes("p.sort-num_votes-visible span:nth-child(2)")%>%
    html_text2() #convert the votes data to text

num_votes_only
```

'2,678,441' · '1,856,865' · '2,651,816' · '1,845,842' · '1,355,557' · '1,271,119' · '791,285' · '2,053,994' · '2,351,325' · '1,666,768' · '2,124,038' · '1,875,189' · '2,078,037' · '762,620' · '1,912,831' · '1,161,816' · '1,292,793' · '1,008,793' · '1,829,410' · '757,928' · '764,488' · '1,391,912' · '1,302,179' · '696,116' · '1,652,282' · '1,100,120' · '1,432,960' · '1,365,364' · '57,963' · '347,108' · '462,825' · '803,408' · '862,414' · '859,623' · '1,333,810' · '1,325,963' · '833,231' · '1,500,848' · '1,123,618' · '1,086,407' · '1,161,745' · '1,059,011' · '262,488' · '278,839' · '1,206,087' · '668,904' · '884,081' · '331,077' · '673,410' · '493,280'

```
num_votes
```

'Votes: 2,678,441 | Gross: $28.34M | Top 250: #1' · 'Votes: 1,856,865 | Gross: $134.97M | Top 250: #2' · 'Votes: 2,651,816 | Gross: $534.86M | Top 250: #3' · 'Votes: 1,845,842 | Gross: $377.85M | Top 250: #7' · 'Votes: 1,355,557 | Gross: $96.90M | Top 250: #6' · 'Votes: 1,271,119 | Gross: $57.30M | Top 250: #4' · 'Votes: 791,285 | Gross: $4.36M | Top 250: #5' · 'Votes: 2,053,994 | Gross: $107.93M | Top 250: #8' · 'Votes: 2,351,325 | Gross: $292.58M | Top 250: #14' · 'Votes: 1,666,768 | Gross: $342.55M | Top 250: #13' · 'Votes: 2,124,038 | Gross: $37.03M | Top 250: #12' · 'Votes: 1,875,189 | Gross: $315.54M | Top 250: #9' · 'Votes: 2,078,037 | Gross: $330.25M | Top 250: #11' · 'Votes: 762,620 | Gross: $6.10M | Top 250: #10' · 'Votes: 1,912,831 | Gross: $171.48M | Top 250: #16' · 'Votes: 1,161,816 | Gross: $46.84M | Top 250: #17' · 'Votes: 1,292,793 | Gross: $290.48M | Top 250: #15' · 'Votes: 1,008,793 | Gross: $112.00M | Top 250: #18' · 'Votes: 1,829,410 | Gross: $188.02M | Top 250: #26' · 'Votes: 757,928 | Gross: $7.56M | Top 250: #23' · 'Votes: 764,488 | Gross: $10.06M | Top 250: #31' · 'Votes: 1,391,912 | Gross: $216.54M | Top 250: #24' · 'Votes: 1,302,179 | Gross: $136.80M | Top 250: #27' · 'Votes: 696,116 | Gross: $57.60M | Top 250: #25' · 'Votes: 1,652,282 | Gross: $100.13M | Top 250: #19' · 'Votes: 1,100,120 | Gross: $204.84M | Top 250: #29' · 'Votes: 1,432,960 | Gross: $130.74M | Top 250: #22' · 'Votes: 1,365,364 | Gross: $322.74M | Top 250: #28' · 'Votes: 57,963 | Top 250: #44' · 'Votes: 347,108 | Gross: $0.27M | Top 250: #20' · 'Votes: 462,825 | Top 250: #21' · 'Votes: 803,408 | Gross: $53.37M | Top 250: #34' · 'Votes: 862,414 | Gross: $13.09M | Top 250: #42' · 'Votes: 859,623 | Gross: $13.18M | Top 250: #46' · 'Votes: 1,333,810 | Gross: $53.09M | Top 250: #41' · 'Votes: 1,325,963 | Gross: $132.38M | Top 250: #39' · 'Votes: 833,231 | Gross: $32.57M | Top 250: #33' · 'Votes: 1,500,848 | Gross: $187.71M | Top 250: #37' · 'Votes: 1,123,618 | Gross: $6.72M | Top 250: #38' · 'Votes: 1,086,407 | Gross: $23.34M | Top 250: #40' · 'Votes: 1,161,745 | Gross: $19.50M | Top 250: #35' · 'Votes: 1,059,011 | Gross: $422.78M | Top 250: #36' · 'Votes: 262,488 | Gross: $11.99M | Top 250: #50' · 'Votes: 278,839 | Top 250: #45' · 'Votes: 1,206,087 | Gross: $210.61M | Top 250: #30' · 'Votes: 668,904 | Gross: $83.47M | Top 250: #53' · 'Votes: 884,081 | Gross: $78.90M | Top 250: #51' · 'Votes: 331,077 | Gross: $5.32M | Top 250: #48' · 'Votes: 673,410 | Gross: $32.00M | Top 250: #32' · 'Votes: 493,280 | Gross: $36.76M | Top 250: #49'

```
# build a dataset
# put in DF
df <- data.frame(
    title = titles,
    rating = ratings,
    num_vote = num_votes_only
)
head(df)
```

A data.frame: 6 × 3

	title	rating	num_vote
	<chr>	<chr>	<chr>
1	1. The Shawshank Redemption (1994)	9.3	2,678,441
2	2. The Godfather (1972)	9.2	1,856,865
3	3. The Dark Knight (2008)	9.0	2,651,816
4	4. The Lord of the Rings: The Return of the King (2003)	9.0	1,845,842
5	5. Schindler's List (1993)	9.0	1,355,557
6	6. The Godfather Part II (1974)	9.0	1,271,119

eCommerce Website

Specphone

```
url2 <- "https://specphone.com/Apple-iPhone-14-Pro-Max.html"
```

```
# read html
spec <- read_html(url2)
```

```
# pull all attributes in ข้อมูลเครื่อง section
# <div class="topic">วันเปิดตัว</div> is att
# <div class="detail">กันยายน 2565</div> is value
att <- spec %>%
    html_nodes("div.topic") %>%
    html_text2()
att
```

'วันเปิดตัว' · 'วันวางจำหน่าย' · 'ขนาด' · 'น้ำหนัก' · 'วัสดุ' · 'SIM' · 'Technology' · '2G' · '3G' · '4G' · '5G' · 'ความเร็ว' · 'ประเภท' · 'ขนาดหน้าจอ' · 'ความละเอียด' · 'ระบบปฏิบัติการ' · 'ชิปประมวลผล' · 'ชิปกราฟิก' · 'หน่วยความจำ' · 'ความจุ' · 'Memory Card' · 'กล้องหลัก' · 'ความละเอียดวีดีโอ' · 'กล้องหน้า' · 'Bluetooth' · 'Wi-Fi' · 'USB' · 'GPS' · 'NFC' · 'ความจุ' · 'ประเภท' · 'Wireless Charging' · 'Fast Charging'

```
# <div class="detail">กันยายน 2565</div> is value
value <- spec %>%
    html_nodes("div.detail") %>%
    html_text2()
```

```
value
```

'กันยายน 2565' · 'ยังไม่วางจำหน่าย' · '160.70 x 77.60 x 7.90 มม.' · '240 กรัม' ·
'Glass front (Gorilla Glass), glass back (Gorilla Glass), stainless steel frame' ·
'รองรับ 1 ซิมการ์ด (esim,nano sim)' · 'HSPA 42.2/5.76 Mbps, LTE-A, 5G, EV-DO Rev.A 3.1 Mbps' ·
'850/900/1800/1900' · '850/900/1900/2100' · '850/900/1900/2100/2600' · '2100/2600/3500/4700' ·
'HSPA 42.2/5.76 Mbps, LTE-A, 5G, EV-DO Rev.A 3.1 Mbps' · 'LTPO Super Retina XDR OLED' · '6.70 นิ้ว' ·
'1290 x 2796 pixels' · 'iOS 16' · 'Apple null A16 Bionic 3.1 GHz' · 'Apple GPU (5-core graphics)' · '6 GB' ·
'128/256/512/1024 GB' · 'ไม่รองรับ' ·
'ตัวที่ 1: 48 MP, f/1.8, 24mm (wide), 1.22µm, dual pixel PDAF, sensor-shift OIS\nตัวที่ 2: 12 MP, f/2.8, 77mm (telephoto), PDAF, OIS, 3x optical zoom\nตัวที่ 3: 12 MP, f/2.2, 13mm, 120˚ (ultrawide), 1.4µm, dual pixel PDAF\nตัวที่ 4: TOF 3D LiDAR scanner (depth)' ·
'4K@24/25/30/60fps, 1080p@25/30/60/120/240fps, HDR, Dolby Vision HDR (up to 60fps), Cinematic mode (4K@30fps), stereo sound rec.' ·
'ตัวที่ 1: 12 MP, f/1.9, 23mm (wide), 1/3.6\nตัวที่ 2: SL 3D, (depth/biometrics sensor)' · '5.3, A2DP, LE' ·
'802.11 a/b/g/n/ac/6, dual' · 'Lightning' · 'dual-band A-GPS, GLONASS,' · 'รอบรับ' · '0 mAh' ·
'Non-removable Li-Ion' · 'รองรับ' · 'รองรับ ()'

```
# build a dataset
df2 <- data.frame(
        attibute = att,
        value = value
)

head(df2)
```

A data.frame: 6 × 2

	attibute	value
	<chr>	<chr>
1	วันเปิดตัว	กันยายน 2565
2	วันวางจำหน่าย	ยังไม่วางจำหน่าย
3	ขนาด	160.70 x 77.60 x 7.90 มม.
4	น้ำหนัก	240 กรัม
5	วัสดุ	Glass front (Gorilla Glass), glass back (Gorilla Glass), stainless steel frame
6	SIM	รองรับ 1 ซิมการ์ด (esim,nano sim)

```
# All Apple Smartphones

apple_url <- read_html("https://specphone.com/brand/Apple")
apple_url
```

```
{html_document}
<html class="no-js" lang="en-US">
[1] <head itemscope itemtype="https://schema.org/WebSite">\n<meta http-equiv= .
[2] <body id="blog" class="wp-custom-logo wp-embed-responsive main" itemscope .
```

```
# link to all Apple smartphones
# we want href (att), <li class="mobile-brand-item col-4 col-sm-4 col-md-3 col-xl
# att -> orange texts
# val -> white
links <- apple_url %>%
    html_nodes("li.mobile-brand-item a") %>% #css selector: find a that is in li.
    html_attr("href")
links
```

'/Apple-iPad-Mini-Wifi.html' · '/Apple-iPhone-3GS-8GB.html' · '/Apple-iPad-Mini-2-Wifi.html' · '/Apple-iPad-10.2-WiFi-2020.html' · '/Apple-iPad-10.2.html' · '/Apple-iPad-10.2-WiFi2021.html' · '/Apple-iPad-9.7-2018-WIFI.html' · '/Apple-iPhone-4-8GB.html' · '/Apple-iPhone-6s-Plus.html' · '/Apple-iPad-4-Wifi.html' · '/Apple-iPhone-6.html' · '/Apple-iPad-Air-2-WiFi.html' · '/Apple-iPad-Wi-Fi-16GB.html' · '/Apple-iPad-Mini-Wifi+Cellular.html' · '/Apple-iPad-Mini-3-WiFi.html' · '/Apple-iPad-Air-5-2022-WiFi.html' · '/Apple-iPhone-5.html' · '/Apple-iPad-2-Wi-Fi-64GB.html' · '/Apple-iPhone-XR.html' · '/Apple-iPhone-4.html' · '/Apple-iPad-Air-2019-Cellular.html' · '/Apple-iPad-9.7-WIFI.html' · '/Apple-iPad-2-3G-32GB.html' · '/Apple-iPad-3G-64GB.html' · '/Apple-iPad-mini-6-Cellular.html' · '/Apple-iPad-Air-2020-Cellular.html' · '/Apple-iPad-Pro-10.5-WIFI.html' · '/Apple-iPhone-14-Pro.html' · '/Apple-iPhone-13-Pro-Max.html' · '/Apple-iPad-Pro-12.9-Cellular-2021.html' · '/Apple-iPhone-14-Pro-Max.html'

```
# to make urls above fully clickable url
#1

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

```
full_links[1:10]
```

```
#2 loop and put the result in DF
#- for each link...
#- create empty DF to put result into
#- put the result in DF

# better to work in R studio
result <- data_frame() #empty
for (link in full_links[1:10]){
    ap_topic <- link %>%
        read_html() %>% #Dont forget to add this
        html_nodes("div.topic") %>%
        html_text2()

    ap_detail <- link %>%
        read_html() %>% #Dont forget to add this
        html_nodes("div.detail") %>%
        html_text2()

    tmp <- data.frame(attributes = ap_topic,
                      value = ap_detail)

    #All all result into a single DF
    result <- bind_rows(result, tmp)

    print("Progressing...")
}
print(result)
```

```
[1] "Progressing..."
[1] "Progressing..."
[1] "Progressing..."
[1] "Progressing..."
[1] "Progressing..."
[1] "Progressing..."
```

```
[1] "Progressing..."
[1] "Progressing..."
[1] "Progressing..."
[1] "Progressing..."
# A tibble: 319 × 2
   attributes  value
   <chr>       <chr>
 1 วันเปิดตัว     พฤศจิกายน 2555
 2 วันวางจำหน่าย พฤษภาคม 2556, สินค้าจำหน่ายหมดแล้ว
 3 ขนาด         200.00 x 134.70 x 7.20 มม.
 4 น้ำหนัก       308 กรัม
 5 วัสดุ         Aluminium, Plastic
 6 SIM          รองรับ 1 ซิมการ์ด
 7 Technology  EDGE, HSPA, HSPA+, LTE
```

```
# write csv
write_csv(result, "result_ap_phone.csv")
```