

Mini Project 01 - IMDB Web Scrapping

```
library(tidyverse)
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'

```
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"
```

```
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 .
```

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

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

```
'1. The Shawshank Redemption (1994)' · '2. The Godfather (1972)' · '3. The Dark Knight (2008)' ·
'4. Schindler's List (1993)' · '5. The Lord of the Rings: The Return of the King (2003)' ·
'6. The Godfather Part II (1974)' · '7. 12 Angry Men (1957)' · '8. Pulp Fiction (1994)' · '9. Fight Club (1999)' ·
'10. Inception (2010)'
```

```
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
```

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

```
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: 2,688,960 Gross: \$28.34M Top 250: #1
2	2. The Godfather (1972)	9.2	Votes: 1,864,973 Gross: \$134.97M Top 250: #2
3	3. The Dark Knight (2008)	9.0	Votes: 2,662,422 Gross: \$534.86M Top 250: #3
4	4. Schindler's List (1993)	9.0	Votes: 1,360,010 Gross: \$96.90M Top 250: #6
5	5. The Lord of the Rings: The Return of the King (2003)	9.0	Votes: 1,852,593 Gross: \$377.85M Top 250: #7
6	6. The Godfather Part II (1974)	9.0	Votes: 1,275,810 Gross: \$57.30M Top 250: #4