

Teaching web scraping: Integrating data science into statistics



Mine Dogucu
New College of Florida
@MineDogucu

Mine Çetinkaya-Rundel
Duke University
@minebocek
2018-05-21

What is Web Scrapping?



Example 1

Secure | <https://www.imdb.com/chart/top>

IMDb

Find Movies, TV shows, Celebrities and more...

All

Q

IMDbPro

Help

f

t

i

f Sign in with Facebook

Other Sign in options

IMDb Charts

Top Rated Movies

Top 250 as rated by IMDb Users

Showing 250 Titles

Sort by: Ranking

↓↑

Rank & Title	IMDb Rating	Your Rating	
1. The Shawshank Redemption (1994)	★ 9.2	☆	+
2. The Godfather (1972)	★ 9.2	☆	+
3. The Godfather: Part II (1974)	★ 9.0	☆	+
4. The Dark Knight (2008)	★ 9.0	☆	+
5. 12 Angry Men (1957)	★ 8.9	☆	+
6. Schindler's List (1993)	★ 8.9	☆	+
7. The Lord of the Rings: The Return of the King (2003)	★ 8.9	☆	+

You Have Seen

0/250 (0%)

☐ Hide titles I've seen

IMDb Charts

[Box Office](#)

[Most Popular Movies](#)

[Top Rated Movies](#)

[Top Rated English Movies](#)

[Most Popular TV](#)

[Top Rated TV](#)

[Top Rated Indian Movies](#)

[Lowest Rated Movies](#)

Top Rated Movies by Genre

[Action](#)

[Adventure](#)

[Animation](#)

[Biography](#)

[Comedy](#)

[Crime](#)

[Drama](#)

[Family](#)

[Fantasy](#)

[Film-Noir](#)

[History](#)

[Horror](#)

[Mystery](#)

[Romance](#)

[Sci-Fi](#)

[Thriller](#)

[War](#)

[Western](#)

Hand Scraping

	A	B	C	D
1		Rank & Title	IMDb Rating	Your Rating
2		1. The Shawshank Redemption (1994)	9.2	
3				
4				
5		2. The Godfather (1972)	9.2	
6				
7				
8		3. The Godfather: Part II (1974)	9	
9				
10				
11		4. The Dark Knight (2008)	9	
12				
13				

.xls

	A	B	C	D	
1		Rank & Title	IMDb Rating	Your Rating	
2		1. The Shawshank Redemption (1994)	9.2		
3					
4					
5		2. The Godfather (1972)	9.2		
6					
7					
8		3. The Godfather: Part II (1974)	9		
9					
10					
11		4. The Dark Knight (2008)	9		
12					
13					

.CSV

Web scraping

	title	year	rating
1	The Shawshank Redemption	1994	9.2
2	The Godfather	1972	9.2
3	The Godfather: Part II	1974	9.0
4	The Dark Knight	2008	9.0
5	12 Angry Men	1957	8.9
6	Schindler's List	1993	8.9
7	The Lord of the Rings: The Return of the King	2003	8.9
8	Pulp Fiction	1994	8.9
9	The Good, the Bad and the Ugly	1966	8.8
10	Fight Club	1999	8.8
11	The Lord of the Rings: The Fellowship of the Ring	2001	8.8

Example 2

https://www.imdb.com/search/title?year=2017&title_type=feature&page=1&ref_=adv_nxt


IMDb Find Movies, TV shows, Celebrities and more... All

Movies, TV & Showtimes Celebs, Events & Photos News & Community Watchlist

Most Popular Feature Films Released 2017-01-01 to 2017-12-31

1 to 50 of 11,634 titles | [Next »](#) View Mode: [Compact](#) | [Detailed](#)

Sort by: [Popularity ▲](#) | [Alphabetical](#) | [IMDb Rating](#) | [Number of Votes](#) | [US Box Office](#) | [Runtime](#) | [Year](#) | [Release Date](#)



1. Thor: Ragnarok (2017) [+](#)


PG-13 | 130 min | Action, Adventure, Comedy

★ **7.9** ☆ [Rate this](#) **74** Metascore

Thor is imprisoned on the planet Sakaar, and must race against time to return to Asgard and stop Ragnarök, the destruction of his world, at the hands of the powerful and ruthless villain Hela.

Director: [Taika Waititi](#) | Stars: [Chris Hemsworth](#), [Tom Hiddleston](#), [Cate Blanchett](#), [Mark Ruffalo](#)

Votes: 326,235 | Gross: \$315.06M



2. The Greatest Showman (2017) [+](#)

PG | 105 min | Biography, Drama, Musical

★ **7.8** ☆ [Rate this](#) **48** Metascore

Celebrates the birth of show business, and tells of a visionary who rose from nothing to create a spectacle that became a worldwide sensation.

Director: [Michael Gracey](#) | Stars: [Hugh Jackman](#), [Michelle Williams](#), [Zac Efron](#), [Zendaya](#)

Votes: 120,071 | Gross: \$173.88M

Hypertext Markup Language (HTML) Nodes

```
<html>
<head>
<title>Page Title</title>
</head>
<body>

<h1>My First Heading</h1>
<p>My first paragraph.</p>

</body>
</html>
```

My First Heading

My first paragraph.

ECOTS 2018



Home

About

Program

Register

Birds of a Feather

Breakout Sessions

Keynotes

Regional Conferences

Sponsor Sessions

Virtual Posters

Workshops

Code of Conduct

RS

- | | |
|----------------------|-----------------|
| Back | Alt+Left Arrow |
| Forward | Alt+Right Arrow |
| Reload | Ctrl+R |
| Save as... | Ctrl+S |
| Print... | Ctrl+P |
| Cast... | |
| Translate to English | |
| AdBlock | |
| View page source | Ctrl+U |
| Inspect | Ctrl+Shift+I |


```

<div class="panel panel-default">
  <div class="panel-heading">
    <h4 class="panel-title">
      <a data-toggle="collapse" data-parent="#accordion" href="#ecots18">
        <i id="ecots18_toggle" class="fa fa-minus-square-o fa-fw"></i>
        </span>&nbsp;<b>eCOTS 2018</b></a>
      </h4>
    </div>
    <div id="ecots18" class="panel-collapse collapse in">
      <div class="list-group">
        <a href="/cause/ecots/ecots18/" class="list-group-item">
          <i class="fa fa-fw fa-home">&nbsp;</i> <b>Home</b></a>
        <a href="/cause/ecots/ecots18/about" class="list-group-item">
          <i class="fa fa-fw fa-info">&nbsp;</i> <b>About</b></a>
        <!--<a href="/cause/ecots/ecots18/proposals/submit" class="list-group-item">
          <i class="fa fa-fw fa-file-text">&nbsp;</i> <b>Call for Proposals</b></a-->
        <a href="/cause/ecots/ecots18/program" class="list-group-item">
          <i class="fa fa-fw fa-map-o">&nbsp;</i> <b>Program</b></a>
        <a href="/cause/ecots/ecots18/register" class="list-group-item">
          <i class="fa fa-fw fa-user-plus">&nbsp;</i> <b>Register</b></a>

        <a href="/cause/ecots/ecots18/program/birds-of-a-feather" class="list-group-item">
          <i class="fa fa-fw fa-leaf">&nbsp;</i> Birds of a Feather</a>
        <a href="/cause/ecots/ecots18/program/breakouts" class="list-group-item">
          <i class="fa fa-fw fa-comments">&nbsp;</i> Breakout Sessions</a>

        <a href="/cause/ecots/ecots18/keynotes" class="list-group-item">
          <i class="fa fa-fw fa-microphone">&nbsp;</i> Keynotes</a>

```

☰ eCOTS 2018

🏠 Home

ℹ About

📖 Program

👤+ Register

bit.ly/SelectorGadget

IMDb Top 250 | IMDb

Find Reviews, TV shows, Celebrities and more...

IMDb Charts

Top Rated Movies

Top 250 as rated by IMDb users

Showing 250 Titles

Sort by: **Ranking**

	Rank	Title	Rating	Your Rating
1	1	The Godfather Part II (1973)	9.2	
2	2	The Godfather (1972)	9.2	
3	3	The Godfather: Part II (1973)	9.2	
4	4	The Dark Knight (2008)	9.0	
5	5	12 Angry Men (1957)	9.0	
6	6	The Godfather: Part I (1972)	9.0	

IMDb Charts

Gear (150) Toggle Position Xhtml

The screenshot shows the IMDb website's 'Most Popular Feature' section for the week of 2017-12-31. The page lists the top 25 movies, with 'The Greatest Showman' (2017) at the top. The movie's details, including its cast and crew, are visible on the right side of the page.

Most Popular Feature 2017-12-31
 (1 to 25 of 11,514 titles) [View All](#)

Next up: [Roundup: 4 Signature Movies](#) [New Movies](#)

1. The Greatest Showman (2017)
 PG-13
 7.6
 128 min
 Biography, Drama, Musical
 Celebrates the birth of show business, and tells of a visionary who rose from nothing to create a spectacle that became a worldwide sensation.
 Director: [Michael Gracey](#) Stars: [Hugh Jackman](#), [Michelle Williams](#), [Zac Efron](#), [Rebecca Ferguson](#)
 Votes: 128,871 Gross: \$170.8M

2. Revenge (II) (2017)
 R
 108 min
 Action, Thriller
 6.4
 120 min
 Action, Thriller
 6.4

3. The Shape of Water (2017)
 PG-13
 129 min
 Drama, Fantasy, Romance
 8.3
 129 min
 Drama, Fantasy, Romance
 8.3

4. The Star (2017)
 PG
 94 min
 Animation, Comedy, Family
 6.5
 94 min
 Animation, Comedy, Family
 6.5

5. The Christmas Chronicles (2017)
 PG
 96 min
 Comedy, Family, Fantasy
 6.5
 96 min
 Comedy, Family, Fantasy
 6.5

6. The Christmas Chronicles 2 (2017)
 PG
 96 min
 Comedy, Family, Fantasy
 6.5
 96 min
 Comedy, Family, Fantasy
 6.5

7. The Christmas Chronicles (2017)
 PG
 96 min
 Comedy, Family, Fantasy
 6.5
 96 min
 Comedy, Family, Fantasy
 6.5

8. The Christmas Chronicles (2017)
 PG
 96 min
 Comedy, Family, Fantasy
 6.5
 96 min
 Comedy, Family, Fantasy
 6.5

9. The Christmas Chronicles (2017)
 PG
 96 min
 Comedy, Family, Fantasy
 6.5
 96 min
 Comedy, Family, Fantasy
 6.5

10. The Christmas Chronicles (2017)
 PG
 96 min
 Comedy, Family, Fantasy
 6.5
 96 min
 Comedy, Family, Fantasy
 6.5

11. The Christmas Chronicles (2017)
 PG
 96 min
 Comedy, Family, Fantasy
 6.5
 96 min
 Comedy, Family, Fantasy
 6.5

12. The Christmas Chronicles (2017)
 PG
 96 min
 Comedy, Family, Fantasy
 6.5
 96 min
 Comedy, Family, Fantasy
 6.5

13. The Christmas Chronicles (2017)
 PG
 96 min
 Comedy, Family, Fantasy
 6.5
 96 min
 Comedy, Family, Fantasy
 6.5

14. The Christmas Chronicles (2017)
 PG
 96 min
 Comedy, Family, Fantasy
 6.5
 96 min
 Comedy, Family, Fantasy
 6.5

15. The Christmas Chronicles (2017)
 PG
 96 min
 Comedy, Family, Fantasy
 6.5
 96 min
 Comedy, Family, Fantasy
 6.5

16. The Christmas Chronicles (2017)
 PG
 96 min
 Comedy, Family, Fantasy
 6.5
 96 min
 Comedy, Family, Fantasy
 6.5

17. The Christmas Chronicles (2017)
 PG
 96 min
 Comedy, Family, Fantasy
 6.5
 96 min
 Comedy, Family, Fantasy
 6.5

18. The Christmas Chronicles (2017)
 PG
 96 min
 Comedy, Family, Fantasy
 6.5
 96 min
 Comedy, Family, Fantasy
 6.5

19. The Christmas Chronicles (2017)
 PG
 96 min
 Comedy, Family, Fantasy
 6.5
 96 min
 Comedy, Family, Fantasy
 6.5

20. The Christmas Chronicles (2017)
 PG
 96 min
 Comedy, Family, Fantasy
 6.5
 96 min
 Comedy, Family, Fantasy
 6.5

21. The Christmas Chronicles (2017)
 PG
 96 min
 Comedy, Family, Fantasy
 6.5
 96 min
 Comedy, Family, Fantasy
 6.5

22. The Christmas Chronicles (2017)
 PG
 96 min
 Comedy, Family, Fantasy
 6.5
 96 min
 Comedy, Family, Fantasy
 6.5


23. The Christmas Chronicles (2017)
 PG
 96 min
 Comedy, Family, Fantasy
 6.5
 96 min
 Comedy, Family, Fantasy
 6.5

24. The Christmas Chronicles (2017)
 PG
 96 min
 Comedy, Family, Fantasy
 6.5
 96 min
 Comedy, Family, Fantasy
 6.5

25. The Christmas Chronicles (2017)
 PG
 96 min
 Comedy, Family, Fantasy
 6.5
 96 min
 Comedy, Family, Fantasy
 6.5

Things to Consider

☐ I'm not a robot


reCAPTCHA
[Privacy - Terms](#)

```
library(robotstxt)
```

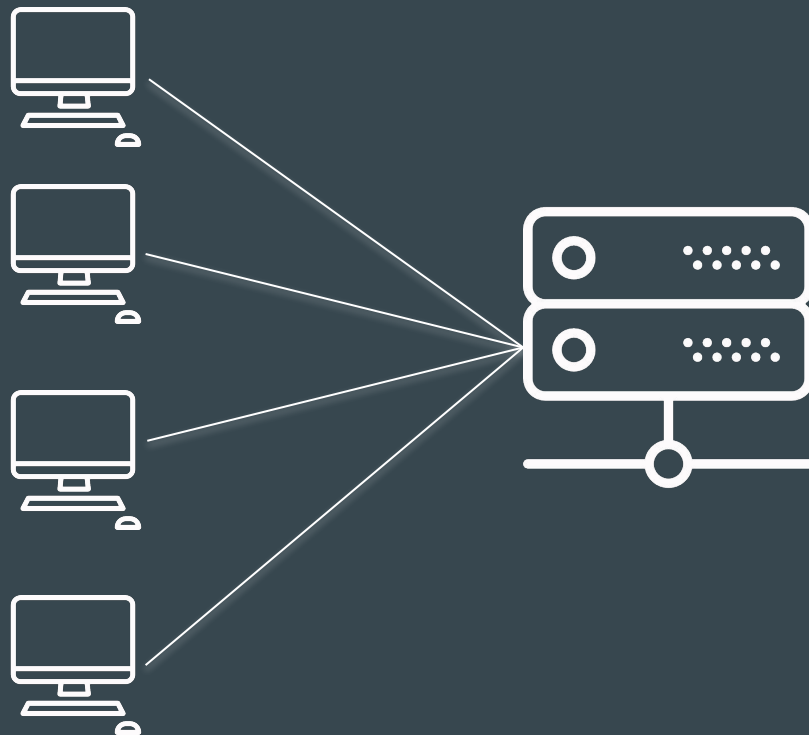
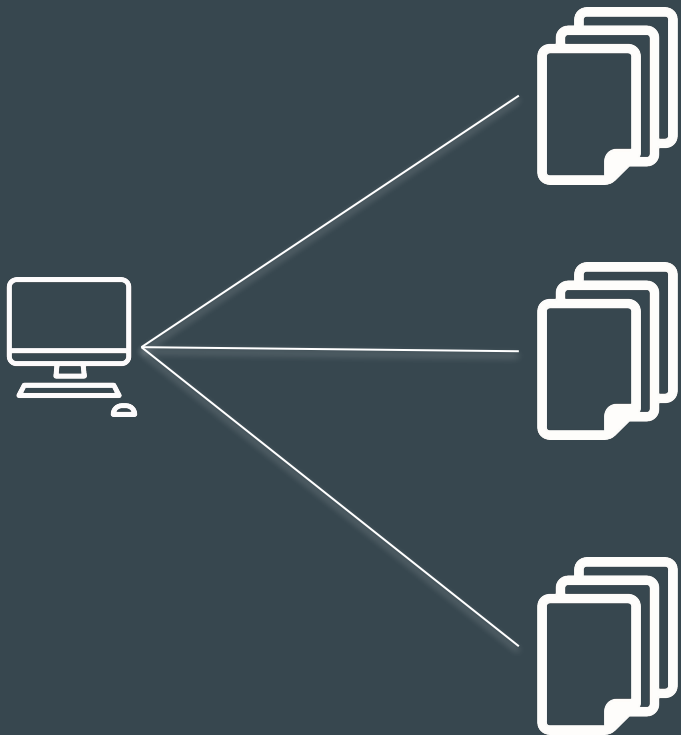
```
paths_allowed("http://www.imdb.com")
```

```
#>
```

```
www.imdb.com
```

```
#> [1] TRUE
```

Things to Consider



```
library(rvest)
```

```
library(tidyverse)
```

```
page <- read_html("http://www.imdb.com/chart/top")
```

Reads an HTML or XML object

```
titles <- page %>%
```

```
  html_nodes("titleColumn a") %>%
```

Selector

```
  html_text()
```

Text value

```
head(titles)
```

```
#> [1] "The Shawshank Redemption" "The Godfather"
```

```
#> [3] "The Godfather: Part II"    "The Dark Knight"
```

```
#> [5] "12 Angry Men"              "Schindler's List"
```


bit.ly/ecots2018

Click on

Projects

Make your own copy of the project called Examples

Examples



Mine Dogucu

 Copy

Benefits

- Students get exposed to non-standard (non-rectangular) data format.
- Students get large amounts of data in a short span of time and in a tidy format.
- Students get exposure to working with strings.
- Students can have more diverse sources of data for statistics projects.
- Web scraping can bring computing topics (e.g. HTML, functions, loops) into the statistics classroom.
- Instructors can use web scraping to curate datasets for classroom use.

Potential Problems

- Website can be down
- NA values

Notes

- More complex scraping is possible
- Timing in the semester
- Web APIs
- Terms of Use

QUESTIONS?

Mine Dogucu

✉ mdogucu@ncf.edu

🐦 @MineDogucu

Mine Çetinkaya-Rundel

✉ mine@stat.duke.edu

🐦 @minebocek

bit.ly/ecots2018-web scraping