

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 Scraping?



Example 1

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


IMDb Find Movies, TV shows, Celebrities and more... All  **IMDbPro** | Help   















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

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

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

- Sponsor Sessions
- Virtual Posters
- Workshops
- Code of Conduct
- RS

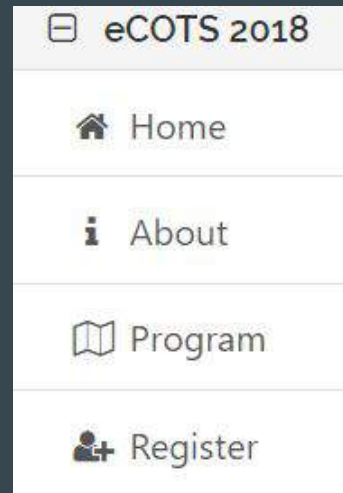

```

<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>

```



bit.ly/SelectorGadget

IMDb Top 250 | IMDb

Find Movies, TV shows, Celebrities, and more.

IMDb Pro | Help | Facebook | Twitter | YouTube

Movies, TV & Showtime | Celebs, Events & Photos | News & Community | Watchlist | Sign In with Facebook | What's Hot is behind

IMDb Charts

Top Rated Movies

Top 250 as rated by IMDb users

Showing 250 Titles

Sort by: **Ranking**

Rank	IMDb Rating	Year	Rating	Watchlist
1	9.2			
2	9.2			
3	9.0			
4	9.0			
5	8.9			
6	8.9			

SHAILENE WOODLEY TOM CLANCY
ADRIFT
SHE'S ON THE BRINK OF DEATH
IN THEATERS JUNE 1 | WATCH TRAILER

You Have Seen

0/250 (0%)

IMDb: Watch The user

IMDb Charts

IMDb Column | Gear (150) | Toggle Position | X

IMDb.com

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

Home, TV & Musicals | Credits, Events & Photos | News & Community | Watchlist | [Sign in with Facebook](#) | Other Sign In Options

Most Popular Feature 2017-12-31

(1 to 50 of 11,034 items) | [View All](#)

Sort by: [Relevance](#) | [Alphabetical](#) | [Runtime](#) | [Year](#) | [Release Date](#)

<p>1. The Shape of Water</p> <p>PG-13 128 min Drama, Fantasy, Romance</p> <p>8.5 Add to Watchlist</p> <p>New in theaters and on DVD. Watch on Amazon Watch on iTunes Watch on YouTube</p> <p>Director: Guillermo del Toro Stars: Sally Hawkins, Michael Shannon, Richard Jenkins, Doug Jones, Michael Ballhaug</p> <p>Box Office: \$113,227,169 Gross: \$133,244,000</p>	<p>2. The Greatest Showman (2017)</p> <p>PG 105 min Biography, Drama, Musical</p> <p>7.4 Add to Watchlist Release Date</p> <p>Suburban the birth of show business, and tells of a visionary who saw that nothing is more beautiful than people in harmony.</p> <p>Director: Michael Gracey Stars: Hugh Jackman, Michelle Williams, Zac Efron, Keaton</p> <p>Box Office: \$106,871,111 Gross: \$477,989,000</p>	<p>3. Revenge (III) (2017)</p> <p>R 109 min Action, Thriller</p> <p>6.5 Add to Watchlist</p>	<p>4. Star Wars: The Last Jedi</p> <p>PG-13 132 min Adventure, Fantasy, Sci-Fi</p> <p>8.3 Add to Watchlist</p>	<p>5. The Star Wars Holiday Special</p> <p>TV-14 45 min Comedy, Family, Fantasy, Sci-Fi</p> <p>6.5 Add to Watchlist</p>	<p>6. The Star Wars Christmas Special</p> <p>TV-14 45 min Comedy, Family, Fantasy, Sci-Fi</p> <p>6.5 Add to Watchlist</p>	<p>7. Star Wars: The Force Awakens</p> <p>PG 151 min Adventure, Fantasy, Sci-Fi</p> <p>7.8 Add to Watchlist</p>	<p>8. Star Wars: The Force Awakens</p> <p>PG 151 min Adventure, Fantasy, Sci-Fi</p> <p>7.8 Add to Watchlist</p>	<p>9. Star Wars: The Force Awakens</p> <p>PG 151 min Adventure, Fantasy, Sci-Fi</p> <p>7.8 Add to Watchlist</p>	<p>10. Star Wars: The Force Awakens</p> <p>PG 151 min Adventure, Fantasy, Sci-Fi</p> <p>7.8 Add to Watchlist</p>
--	--	--	--	---	---	---	---	---	--



Tom Cruise
 \$188 on IMDb.com

IMDb.com

Search: | [Clear](#) | [Cancel](#) | [OK](#)

11:01 PM 1/1/2018

Things to Consider



I'm not a robot



reCAPTCHA

[Privacy](#) - [Terms](#)

```
library(robotstxt)
```

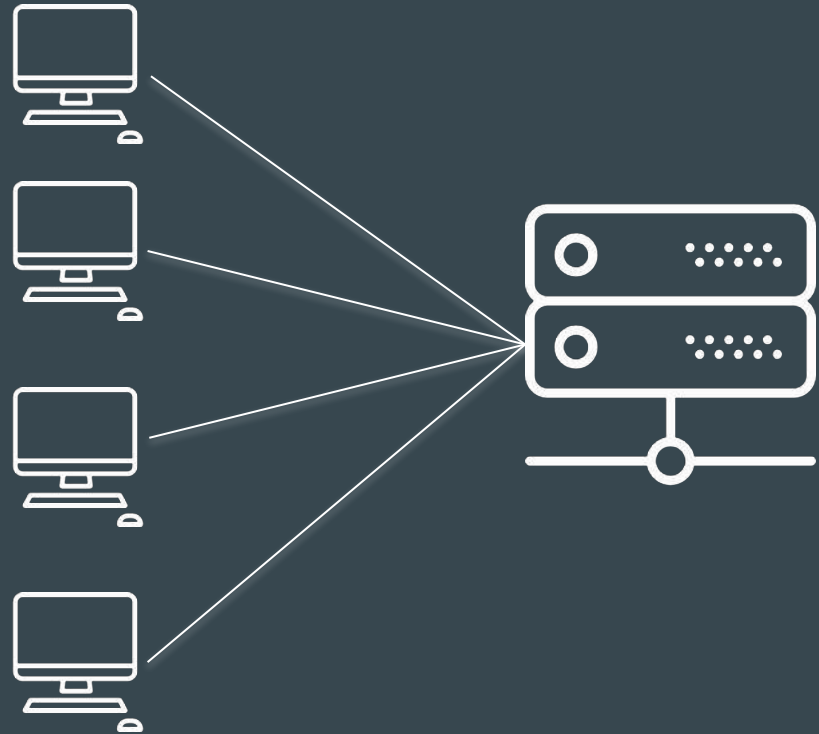
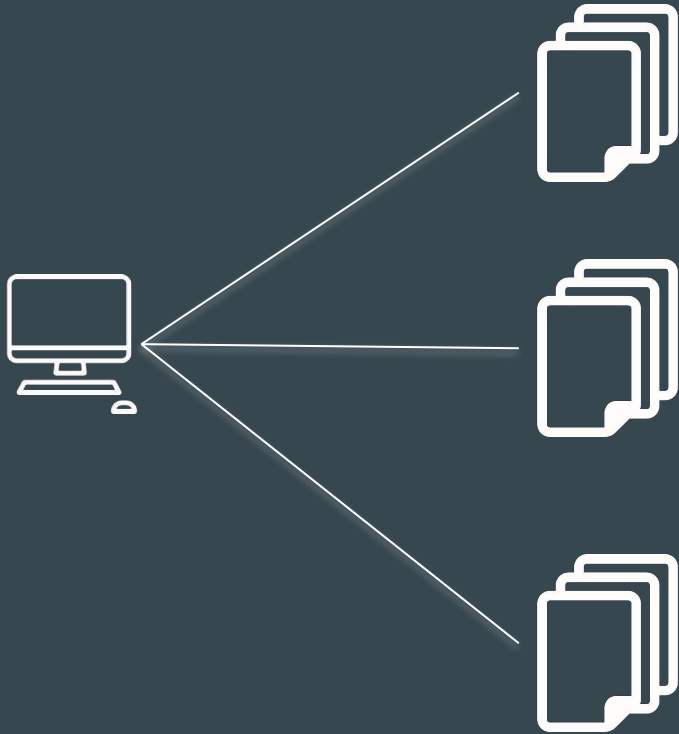
```
paths_allowed("http://www.imdb.com")
```

```
#>
```

```
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-webscraping

<https://github.com/mdogucu/eCOTS2018>