

Web scraping

Data Science in a Box
datasciencebox.org



Scraping the web



Scraping the web: what? why?

- Increasing amount of data is available on the web



Scraping the web: what? why?

- Increasing amount of data is available on the web
- These data are provided in an unstructured format: you can always copy&paste, but it's time-consuming and prone to errors



Scraping the web: what? why?

- Increasing amount of data is available on the web
- These data are provided in an unstructured format: you can always copy&paste, but it's time-consuming and prone to errors
- Web scraping is the process of extracting this information automatically and transform it into a structured dataset



Scraping the web: what? why?

- Increasing amount of data is available on the web
- These data are provided in an unstructured format: you can always copy&paste, but it's time-consuming and prone to errors
- Web scraping is the process of extracting this information automatically and transform it into a structured dataset
- Two different scenarios:
 - Screen scraping: extract data from source code of website, with html parser (easy) or regular expression matching (less easy).
 - Web APIs (application programming interface): website offers a set of structured http requests that return JSON or XML files.



Web Scraping with rvest



Hypertext Markup Language

- Most of the data on the web is still largely available as HTML
- It is structured (hierarchical / tree based), but it's often not available in a form useful for analysis (flat / tidy).

```
<html>
  <head>
    <title>This is a title</title>
  </head>
  <body>
    <p align="center">Hello world!</p>
  </body>
</html>
```



rvest

- The **rvest** package makes basic processing and manipulation of HTML data straight forward
- It's designed to work with pipelines built with %>%



Core rvest functions

- `read_html` - Read HTML data from a url or character string
- `html_node` - Select a specified node from HTML document
- `html_nodes` - Select specified nodes from HTML document
- `html_table` - Parse an HTML table into a data frame
- `html_text` - Extract tag pairs' content
- `html_name` - Extract tags' names
- `htmlAttrs` - Extract all of each tag's attributes
- `html_attr` - Extract tags' attribute value by name



SelectorGadget

- Open source tool that eases CSS selector generation and discovery
- Easiest to use with the Chrome Extension
- Find out more on the SelectorGadget vignette

SelectorGadget: point and click CSS selectors



A screenshot of a computer screen showing the Hacker News homepage. The title bar says "Hacker News SelectorGadget Screencast from Andrew Cantino". The main content area shows a list of news items. The 19th item in the list, which is highlighted with a yellow background, is "New theory may explain the notorious cold fusion experiment from two decades ago (discovermagazine.com)". The SelectorGadget extension icon, a blue smiley face, is visible in the top-left corner of the browser window.

1. ▲ AnandTech: Microsoft Surface Review (anandtech.com)
77 points by bartola 2 hours ago | 37 comments

2. ▲ Wired's Review of the Microsoft Surface (wired.com)
42 points by colinplamondon 2 hours ago | 16 comments

3. ▲ Zynga May Have Just Laid Off 100+ Employees From Its Austin Office (techcrunch.com)
386 points by hompawze 10 hours ago | 1 comment

4. ▲ The Hardware Renaissance (techcrunch.com)
366 points by niquresh 11 hours ago | 171 comments

5. ▲ Don't Call The New Microsoft Surface RT A Tablet, This Is A PC (techcrunch.com)
23 points by vyrtek 2 hours ago | 36 comments

6. ▲ Why we buy into ideas: how to convince others of our thoughts (bufferapp.com)
6 points by sunas34 23 minutes ago | discuss

7. ▲ The rise of the "successful" unsustainable company (asmartbear.com)
281 points by yannickmache 12 hours ago | 105 comments

8. ▲ Under the hood of Windows 8, or why desktop users should upgrade from Windows 7 (extremetech.com)
261 points by eve_ 9 12 hours ago | 170 comments

9. ▲ Marc Andreessen's Productivity Trick to Feeling Marvelously Efficient (idonethis.com)
106 points by mikemk 7 hours ago | 34 comments

10. ▲ Show HN: Taurus.io - Create a product tour for your web app in 15 minutes (taurus.io)
31 points by etzio 3 hours ago | 30 comments

11. ▲ The PC isn't dead (dendory.net)
9 points by dendory 1 hour ago | 6 comments

12. ▲ Ceefax Final Broadcast: "Goodbye, cruel world." (h4ck.in)
76 points by learmers 7 hours ago | 24 comments

13. ▲ Show HN: Fact check last night's Presidential debate with Quip (quipvideo.com)
32 points by dmvaldman 4 hours ago | 12 comments

14. ▲ Increasing wireless network speed by 1000%, by replacing packets with algebra (extremetech.com)
98 points by oliver 7 hours ago | 30 comments

15. ▲ Amazon reopen wiped Kindle account (translate.google.com)
2518 points by EwanTee 15 hours ago | 137 comments

16. ▲ Zynga CEO Mark Pincus Confirms Layoffs: 5% of Workforce (techcrunch.com)
47 points by nikunj 6 hours ago | 11 comments

17. ▲ Stanford grad's site nets Southwest 'cease and desist' (paloaltoonline.com)
21 points by cb33 4 hours ago | 18 comments

18. ▲ OrderAhead is hiring a Marketing Associate
2 hours ago

19. ▲ New theory may explain the notorious cold fusion experiment from two decades ago (discovermagazine.com)



Using the SelectorGadget

The screenshot shows a web browser displaying the [IMDb Top 250](https://www.imdb.com/chart/top) chart. The page features a large banner at the top announcing "PREMIERES TONIGHT STARTING 8/7c CBS". Below the banner, the "IMDb Charts" section is visible, specifically the "Top Rated Movies" chart. The chart lists the top 250 movies based on user ratings. The first entry, "The Shawshank Redemption (1994)", is highlighted with a yellow selection box from the SelectorGadget extension. The extension's interface is overlaid on the bottom right of the page, showing options like "SHARE", "IMDb Charts", and "Top Rated Movies".

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)	9.1	☆

IMDb Top 250 - IMDb

imdb.com/chart/top/

IMDb Menu All Search IMDb

SHARE

Sort by: Ranking

IMDb Rating Your Rating

Rank & Title

1. The Shawshank Redemption (1994) ★ 9.2 ★ +

2. The Godfather (1972) ★ 9.1 ★ +

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
Music
Musical
Mystery
Romance

Click on the app logo next to the search bar in your browser

IMDb Top 250 - IMDb

imdb.com/chart/top/

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)	★ 9.1	★	[+]
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)	No valid path found.		Clear Toggle Position XPath ? X

SHARE

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

Box will open in the bottom right of the browser

Click on a page element, and it will turn green

IMDb Charts

Top Rated Movies

Top 250 as rated by IMDb Users

Showing 250 Titles

Sort by: Ranking

	IMDb Rating	Your Rating	
1. The Shawshank Redemption (1994)	9.2	★	[+]
2. The Godfather (1972)	9.1	★	[+]
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)			

SHARE

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
- Music
- Musical
- Mystery

.titleColumn

Clear (250) Toggle Position XPath ? X

selectorbad get will generate a minimal CSS selector for that element, and will highlight everything that is matched by the selector in yellow

IMDb Top 250 - IMDb

imdb.com/chart/top/

IMDb Menu All Search IMDb

SHARE

Sort by: Ranking

IMDb Rating Your Rating

Rank & Title

1. The Shawshank Redemption (1994) ★ 9.2

2. The Godfather (1972) ★ 9.1

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

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

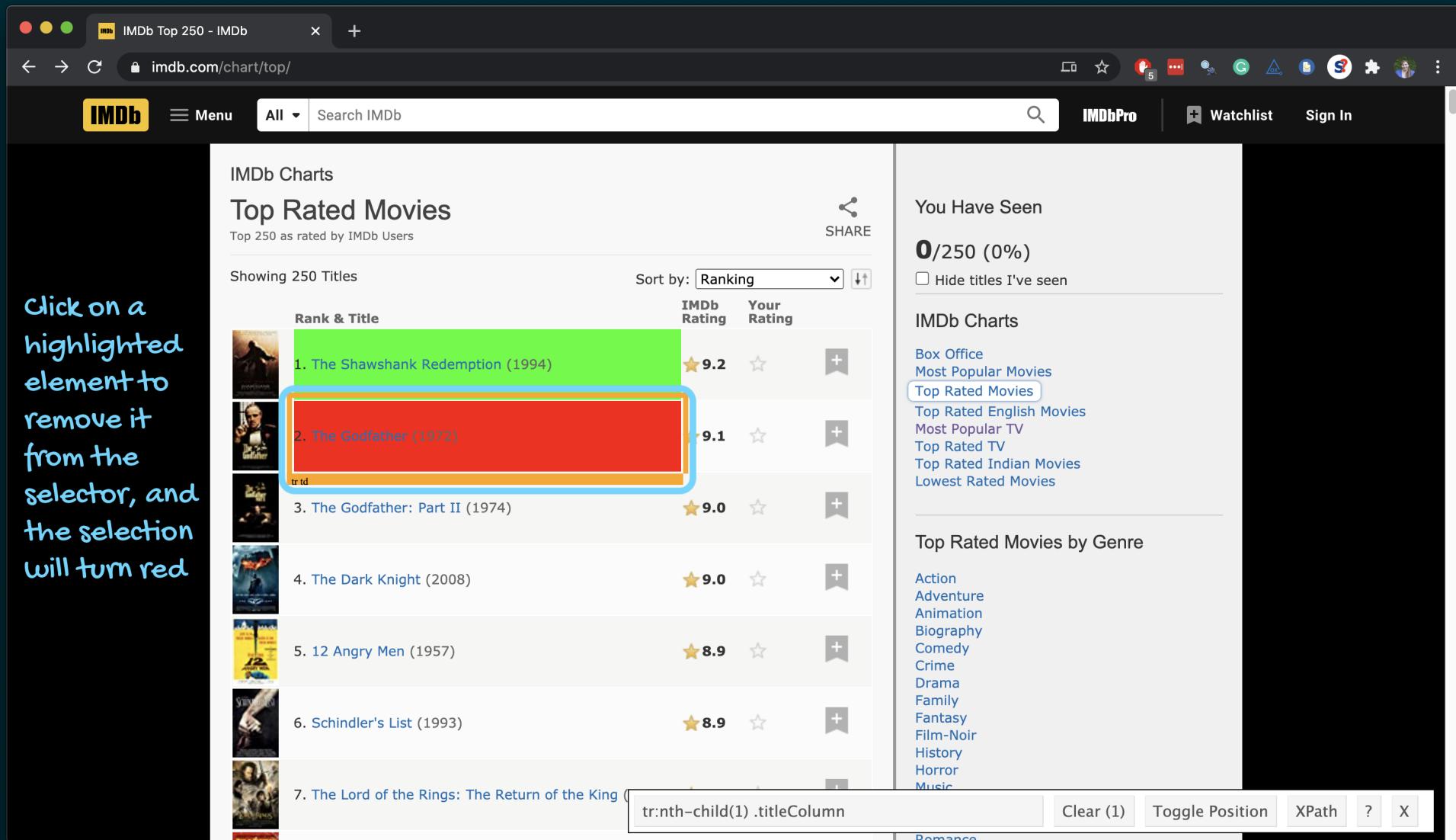
- Action
- Adventure
- Animation
- Biography
- Comedy
- Crime
- Drama
- Family
- Fantasy
- Film-Noir
- History
- Horror
- Music

tr:nth-child(1) .titleColumn

Romance

Clear (1) Toggle Position XPath ? X

Click on a highlighted element to remove it from the selector, and the selection will turn red



IMDb Top 250 - IMDb

imdb.com/chart/top/

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)	9.1	☆
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	☆

SHARE

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
- Music

tr~ tr+ tr .titleColumn , tr:nth-child(1) .titleColumn

Clear (249) Toggle Position XPath ? X

Romance

Click on an unhighlighted element to add it to the selector and it will turn green

Using the SelectorGadget

Through this process of selection and rejection, SelectorGadget helps you come up with the appropriate CSS selector for your needs

