

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
- 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
- `html_attrs` - 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



The image shows a screenshot of a web browser window displaying the Hacker News homepage. A blue status bar at the top of the browser indicates "SelectorGadget Screencast" and "from Andrew Cantino". The main content area of the browser shows a list of news items from Hacker News, each with a title, author, points, and timestamp. The first few items are highlighted with yellow boxes, indicating they were selected using the SelectorGadget extension. The browser interface includes a navigation bar with back, forward, and search buttons, a toolbar with various icons, and a menu bar with "File", "Edit", "View", "Search", "Bookmarks", "Help", and "Other Bookmarks". The status bar also shows "login".

Rank	Title	Author	Points	Comments	Link
1.	AnandTech: Microsoft Surface Review	(anandtech.com)	77	37	link
2.	Wired's Review of the Microsoft Surface	(wired.com)	42	16	link
3.	Zynga May Have Just Laid Off 100+ Employees From Its Austin Office	(techcrunch.com)	386	10	link
4.	The Hardware Renaissance	(com.com)	366	1171	link
5.	Don't Call The New Microsoft Surface RT A Tablet, This Is A PC	(techcrunch.com)	23	36	link
6.	Why we buy into ideas: how to convince others of our thoughts	(bufferapp.com)	6	discuss	link
7.	The rise of the "successful" unsustainable company	(asmartbear.com)	281	105	link
8.	Under the hood of Windows 8, or why desktop users should upgrade from Windows 7	(extremetech.com)	261	170	link
9.	Marc Andreessen's Productivity Trick to Feeling Marvelously Efficient	(idonethis.com)	106	34	link
10.	Show HN: Taurus.io - Create a product tour for your web app in 15 minutes	(taurus.io)	31	30	link
11.	The PC isn't dead	(dendory.net)	9	6	link
12.	Ceefax Final Broadcast: "Goodbye, cruel world."	(h4ck.in)	76	24	link
13.	Show HN: Fact check last night's Presidential debate with Quip	(quipvideo.com)	32	12	link
14.	Increasing wireless network speed by 1000%, by replacing packets with algebra	(extremetech.com)	98	30	link
15.	Amazon reopen wiped Kindle account	(translate.google.com)	2518	137	link
16.	Zynga CEO Mark Pincus Confirms Layoffs: 5% of Workforce	(techcrunch.com)	47	11	link
17.	Stanford grad's site nets Southwest 'cease and desist'	(paloaltoonline.com)	21	18	link
18.	OrderAhead is hiring a Marketing Associate		2	18	link
19.	New theory may explain the notorious cold fusion experiment from two decades ago	(discovermagazine.com)			link



Using the SelectorGadget

The screenshot shows the IMDb Top 250 chart page. A SelectorGadget overlay is active, highlighting the first item in the list: "1. The Shawshank Redemption (1994)". The overlay includes a "SHARE" button and a "Sort by:" dropdown set to "Ranking". The "IMDb Rating" column shows a value of "9.2" for both the highlighted row and the second row. The "Your Rating" column shows an empty star icon for both rows.

Rank & Title	IMDb Rating	Your Rating
1. The Shawshank Redemption (1994)	9.2	[empty star]
2. The Godfather (1972)	9.2	[empty star]
3. The Godfather: Part II (1974)	9.1	[empty star]

No valid path found.

Clear Toggle Position XPath ? X

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

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

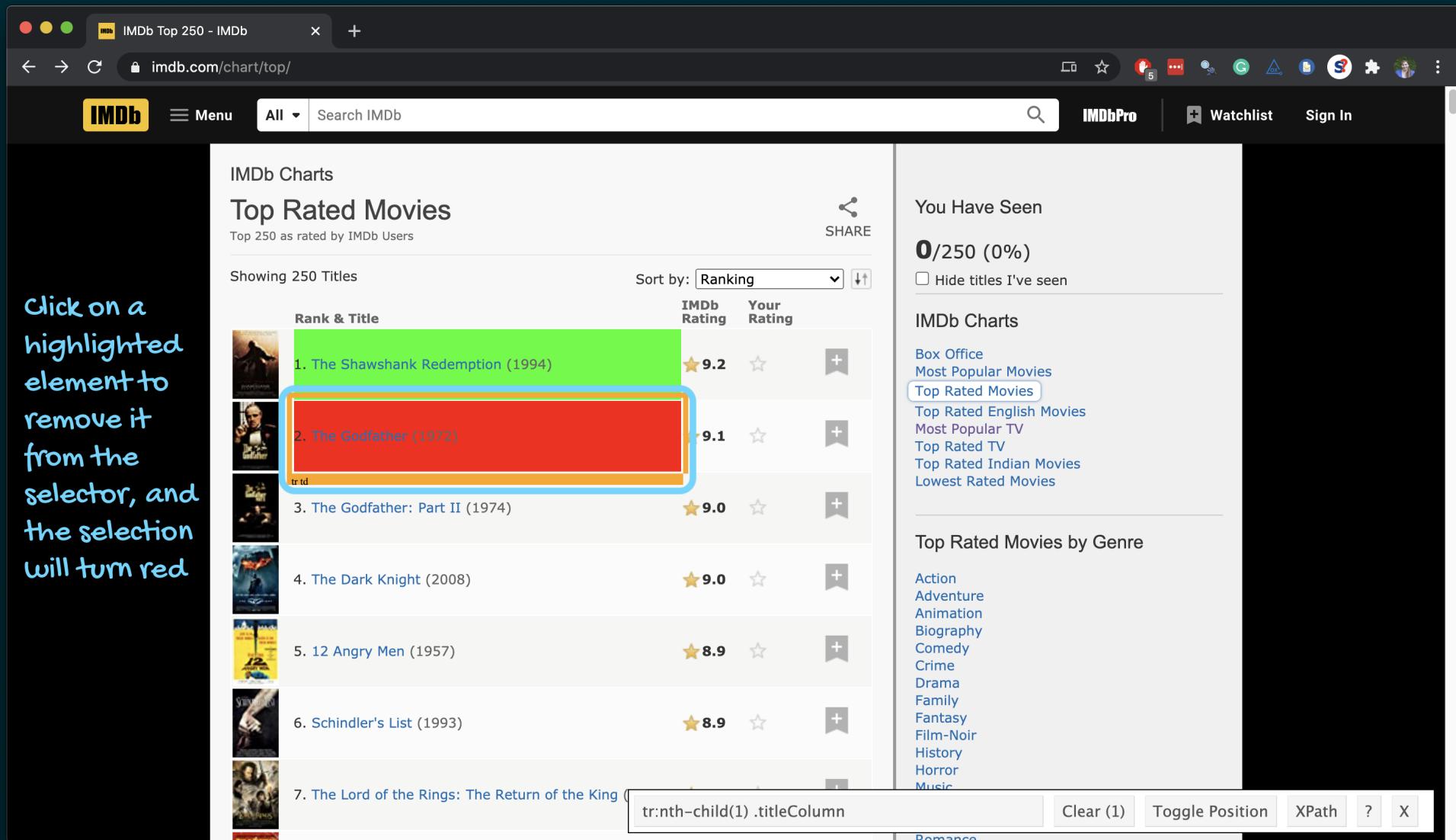
- Action
- Adventure
- Animation
- Biography
- Comedy
- Crime
- Drama
- Family
- Fantasy
- Film-Noir
- History
- Horror
- Music

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

