

# Web scraping

Filippo Chiarello, Ph.D.

# 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 screenshot shows a Hacker News homepage with the title "SelectorGadget Screencast" and author "from Andrew Cantino". The page lists 19 news items. The 4th item, "The Hardware Renaissance", has a blue rectangular highlight around its link and title, indicating it was selected by the SelectorGadget extension. The extension's interface is visible at the top of the browser window.

Rank	Link	Title	Score	Comments
1.	<a href="#">AnandTech</a>	Microsoft Surface Review	anandtech.com	77 points   37 comments
2.	<a href="#">Wired's Review of the Microsoft Surface</a>	wired.com	42 points   16 comments	
3.	<a href="#">Zynga May Have Just Laid Off 100+ Employees From Its Austin Office</a>	techcrunch.com	384 points   10 hours ago	
4.	<a href="#">The Hardware Renaissance</a>	com	366 points   11 hours ago   171 comments	
5.	<a href="#">Don't Call The New Microsoft Surface RT A Tablet, This Is A PC</a>	techcrunch.com	23 points   2 hours ago   16 comments	
6.	<a href="#">Why we buy into ideas: how to convince others of our thoughts</a>	bufferapp.com	6 points   sunplus34 23 minutes ago   discuss	
7.	<a href="#">The rise of the "successful" unsustainable company</a>	asmartbear.com	281 points   yannickmache 12 hours ago   109 comments	
8.	<a href="#">Under the hood of Windows 8, or why desktop users should upgrade from Windows 7</a>	extremetech.com	261 points   eve_ 9:12 hours ago   170 comments	
9.	<a href="#">Marc Andreessen's Productivity Trick to Feeling Marvelously Efficient</a>	idonethis.com	106 points   mikemun 7 hours ago   34 comments	
10.	<a href="#">Show HN: Taurus.io - Create a product tour for your web app in 15 minutes</a>	taurus.io	31 points   edzio 3 hours ago   30 comments	
11.	<a href="#">The PC isn't dead</a>	dendory.net	9 points   dendory 1 hour ago   6 comments	
12.	<a href="#">Ceefax: Final Broadcast: 'Goodbye, cruel world.'</a>	h4ck.in	76 points   leemans 7 hours ago   24 comments	
13.	<a href="#">Show HN: Fact check last night's Presidential debate with Quip</a>	quipvideo.com	32 points   dmvaldman 4 hours ago   12 comments	
14.	<a href="#">Increasing wireless network speed by 1000%, by replacing packets with algebra</a>	extremetech.com	98 points   oliver 7 hours ago   30 comments	
15.	<a href="#">Amazon reopen wiped Kindle account</a>	translate.google.com	258 points   lwanTee 15 hours ago   137 comments	
16.	<a href="#">Zynga CEO Mark Pincus Confirms Layoffs: 5% of Workforce</a>	techcrunch.com	47 points   nlukjanj 6 hours ago   11 comments	
17.	<a href="#">Stanford grad's site nets Southwest 'cease and desist'</a>	paloaltoonline.com	21 points   cb33 4 hours ago   18 comments	
18.	<a href="#">OrderAhead is hiring a Marketing Associate</a>	2 hours ago		
19.	<a href="#">New theory may explain the notorious cold fusion experiment from two decades ago</a>	discovermagazine.com		

# Using the SelectorGadget

The screenshot shows a browser window displaying the IMDb Top 250 chart at <https://www.imdb.com/chart/top>. The page features a large banner for TV premieres on CBS. On the left, there's a sidebar for 'IMDb Charts' with 'Top Rated Movies' selected. The main content area shows the top three movies: 'The Shawshank Redemption' (1994) and 'The Godfather' (1972), both with an IMDb rating of 9.2. A yellow box highlights the first movie. A SelectorGadget overlay is visible at the bottom right, showing the path: 'tbody > tr:nth-child(1)'. The overlay includes buttons for 'Clear', 'Toggle Position', 'XPath', '?', and 'X'.

Rank & Title	IMDb Rating	Your Rating	Action
1. The Shawshank Redemption (1994)	9.2	☆	[+]
2. The Godfather (1972)	9.2	☆	[+]
3. The Godfather: Part II (1974)	No valid path found.		

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

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

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

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

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

# Using the SelectorGadget

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

