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

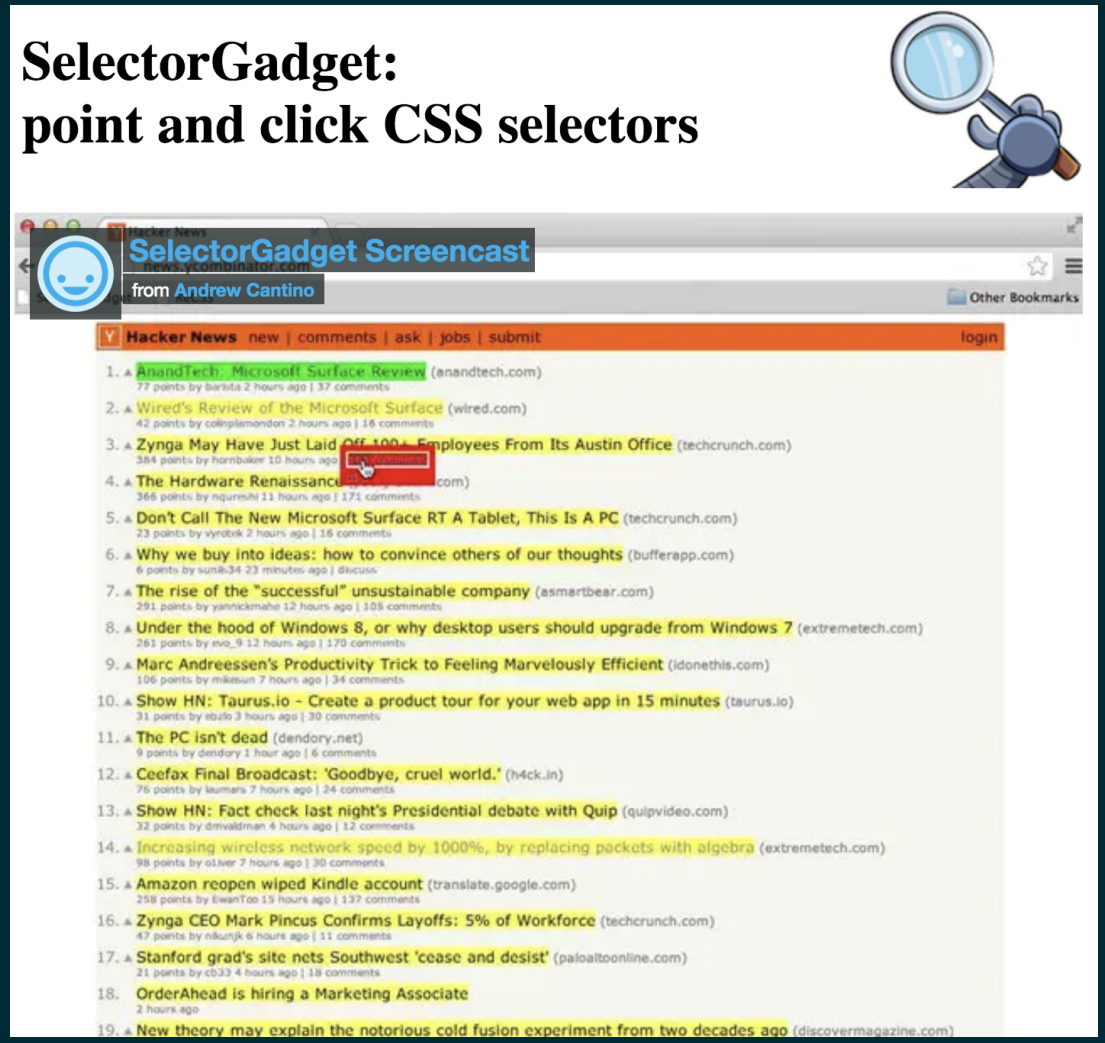
- The **rvest** package makes basic processing and manipulation of HTML data straight forward
- It's designed to work with pipelines built with %>%

Core rvest functions

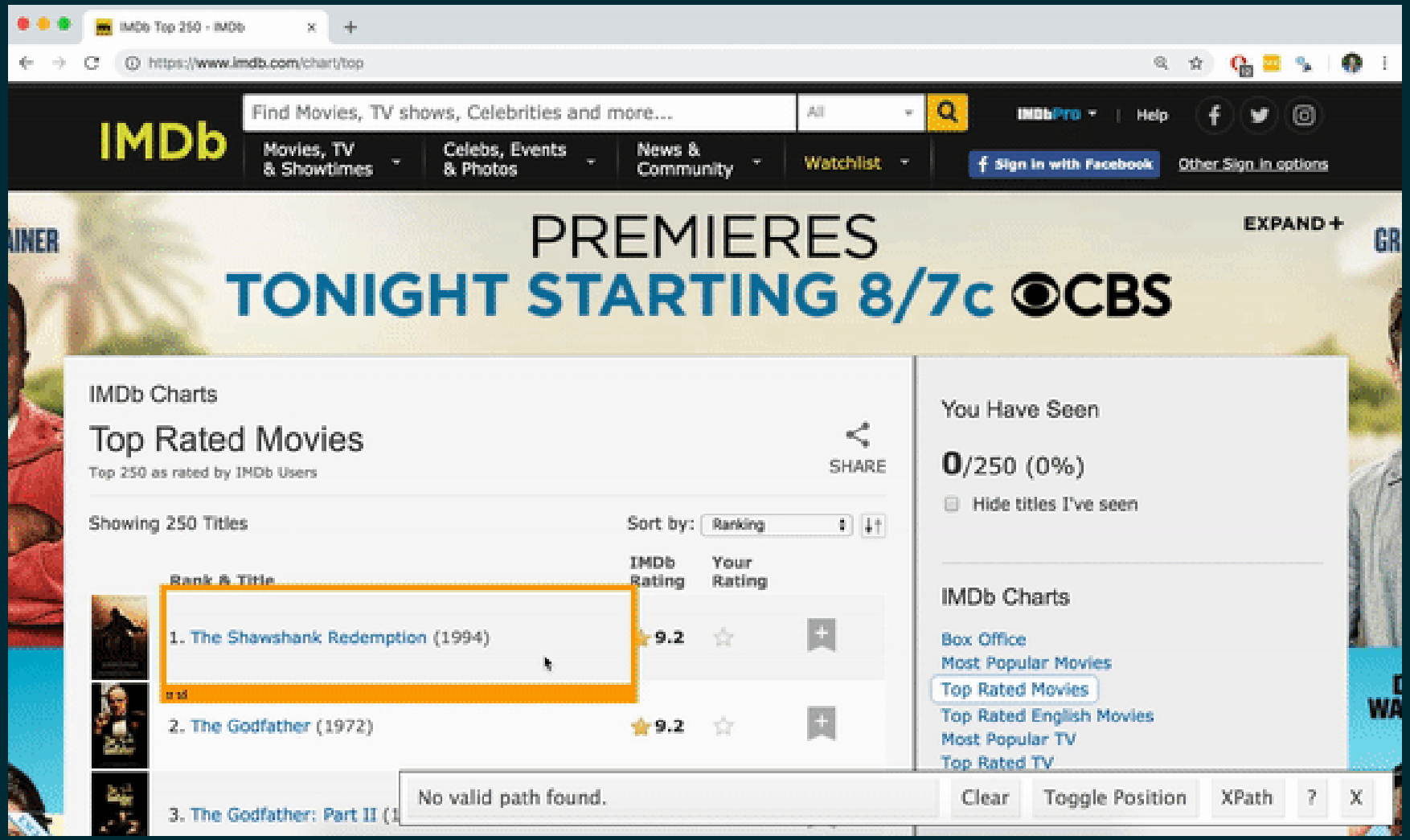
- `read_html` - Read HTML data from a url or character string
- `html_node` - Select a specified node from HTML document
- `html_nodes` - Select specified nodes from HTML document
- `html_table` - Parse an HTML table into a data frame
- `html_text` - Extract tag pairs' content
- `html_name` - Extract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

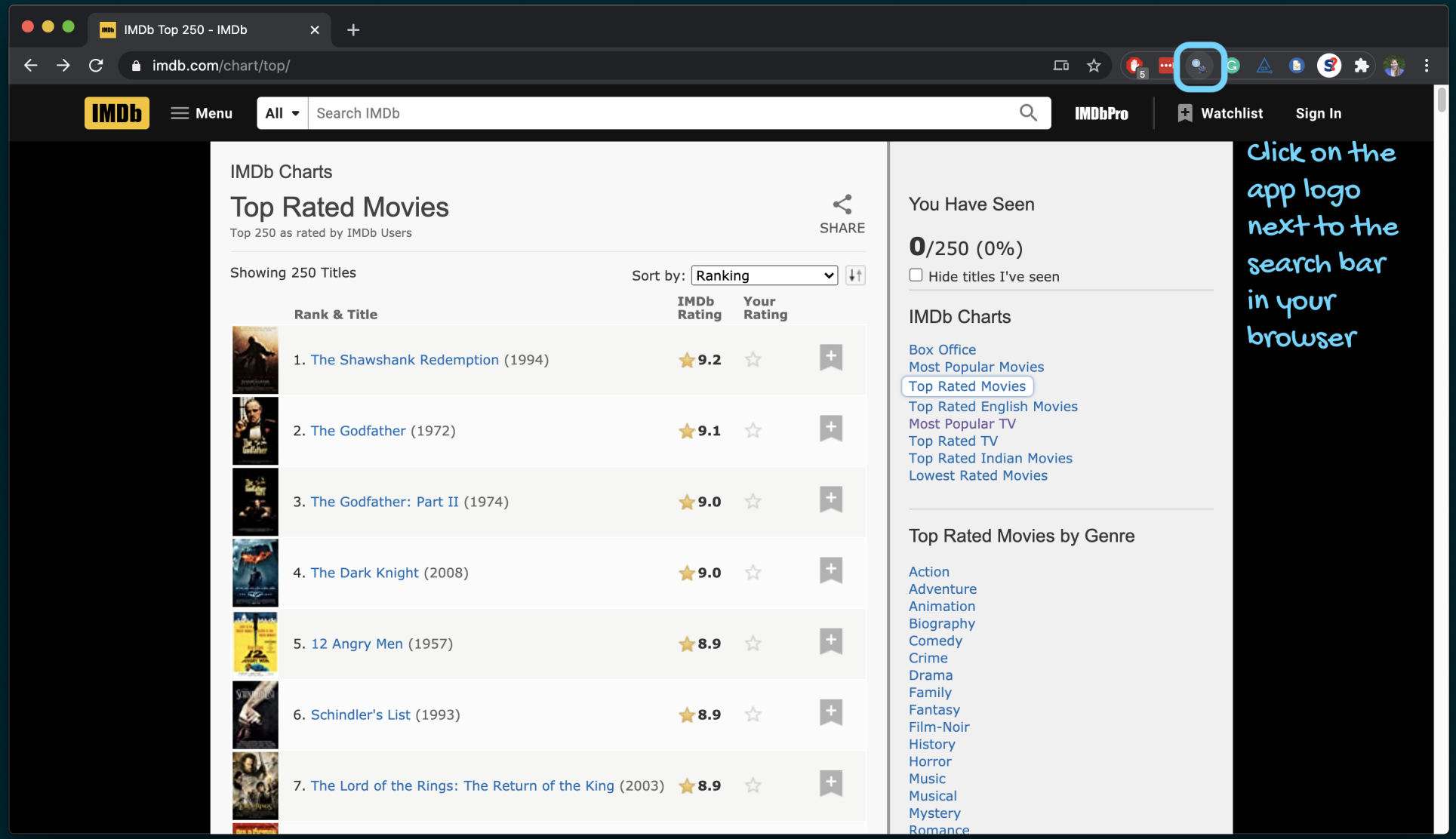
SelectorGadget

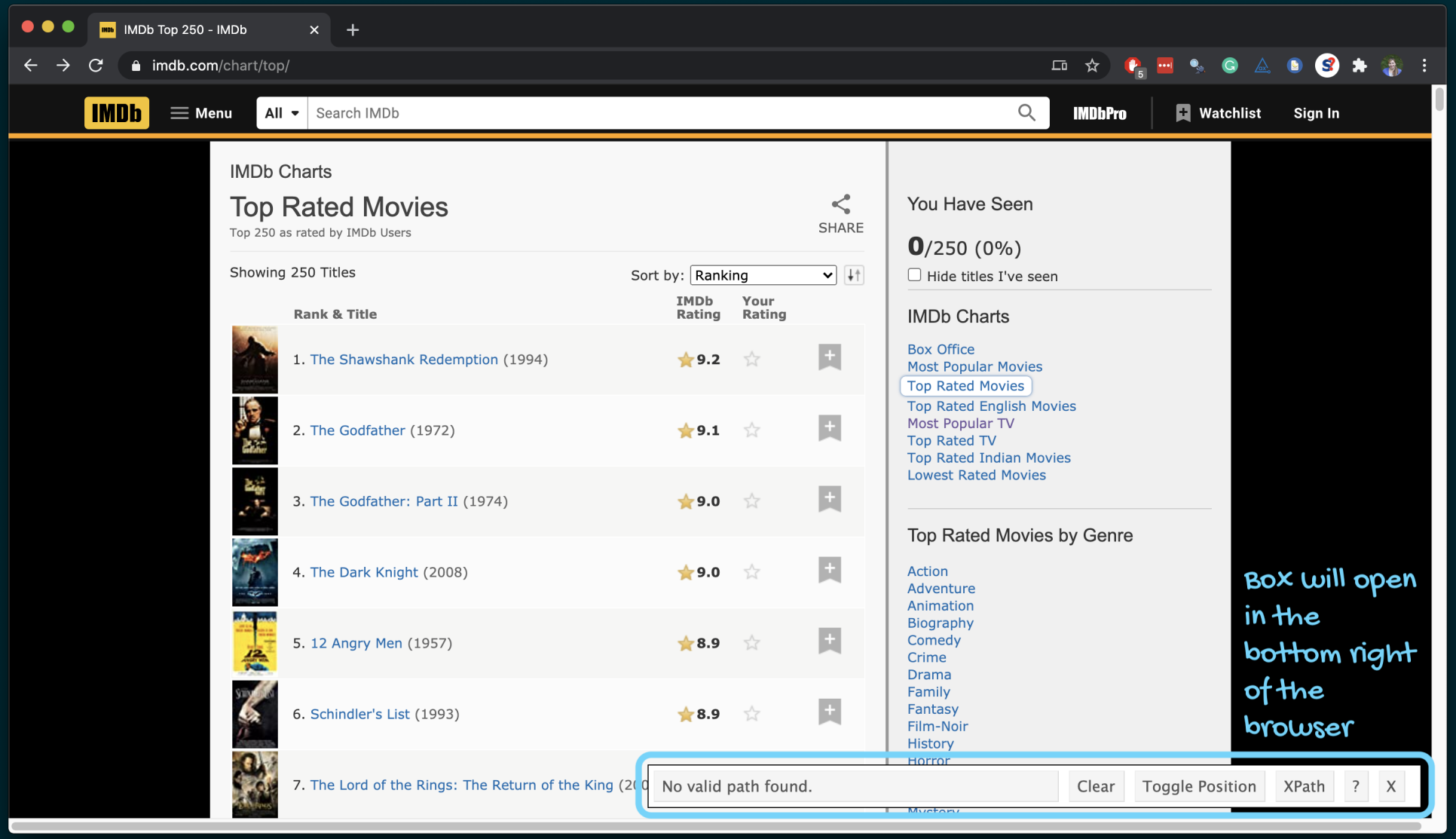
- Open source tool that eases CSS selector generation and discovery
- Easiest to use with the Chrome Extension
- Find out more on the SelectorGadget vignette

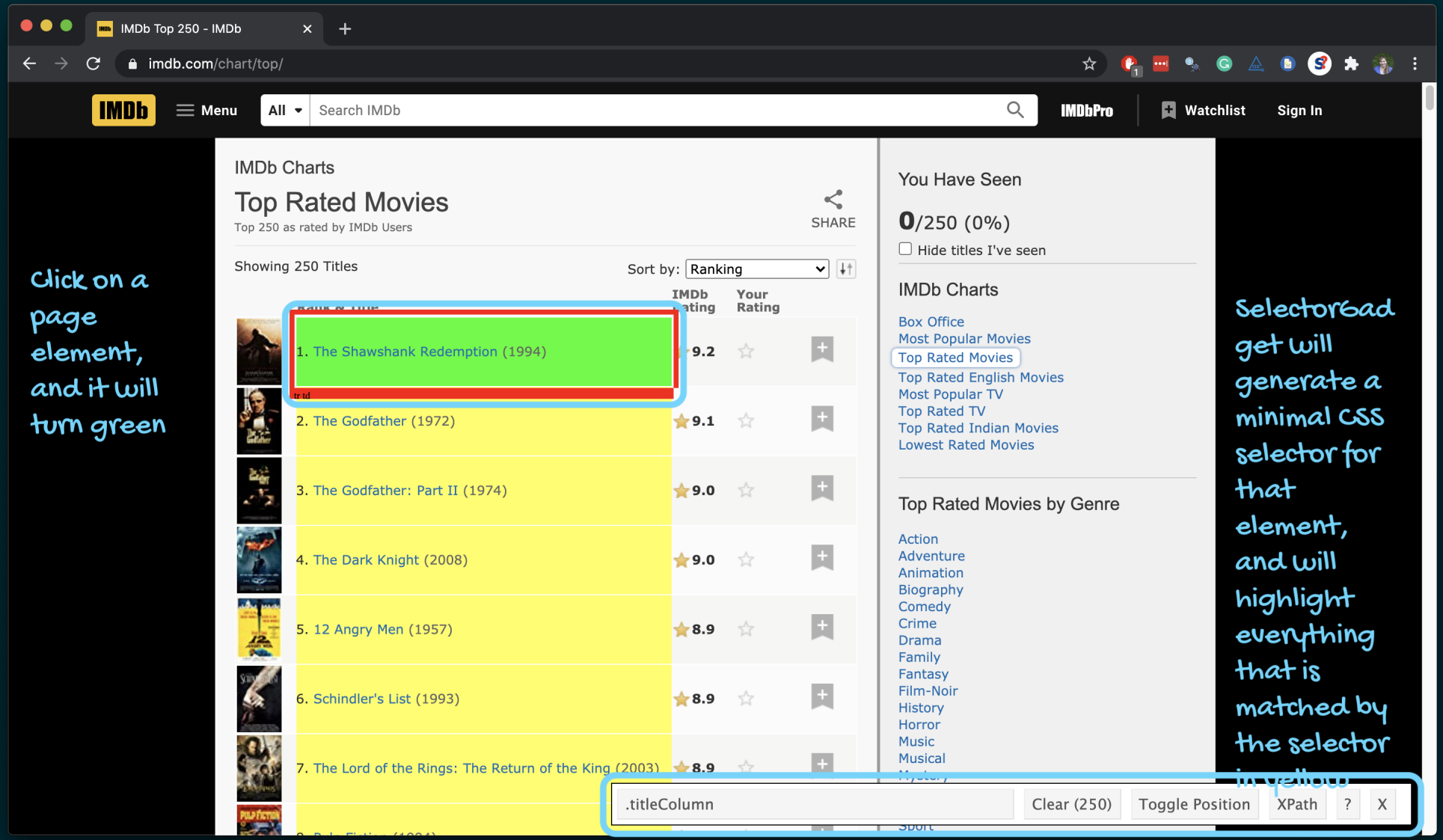


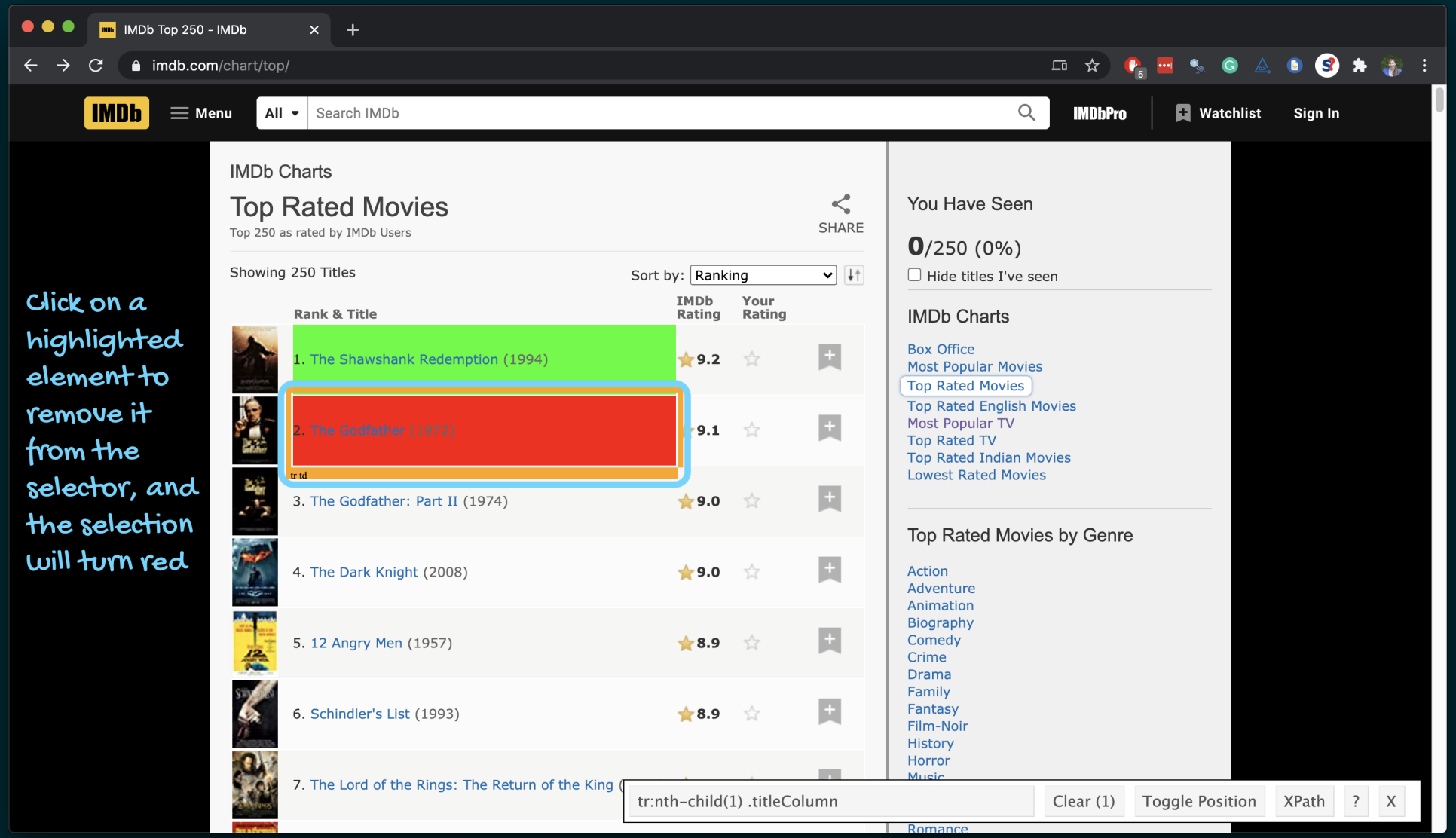
Using the SelectorGadget

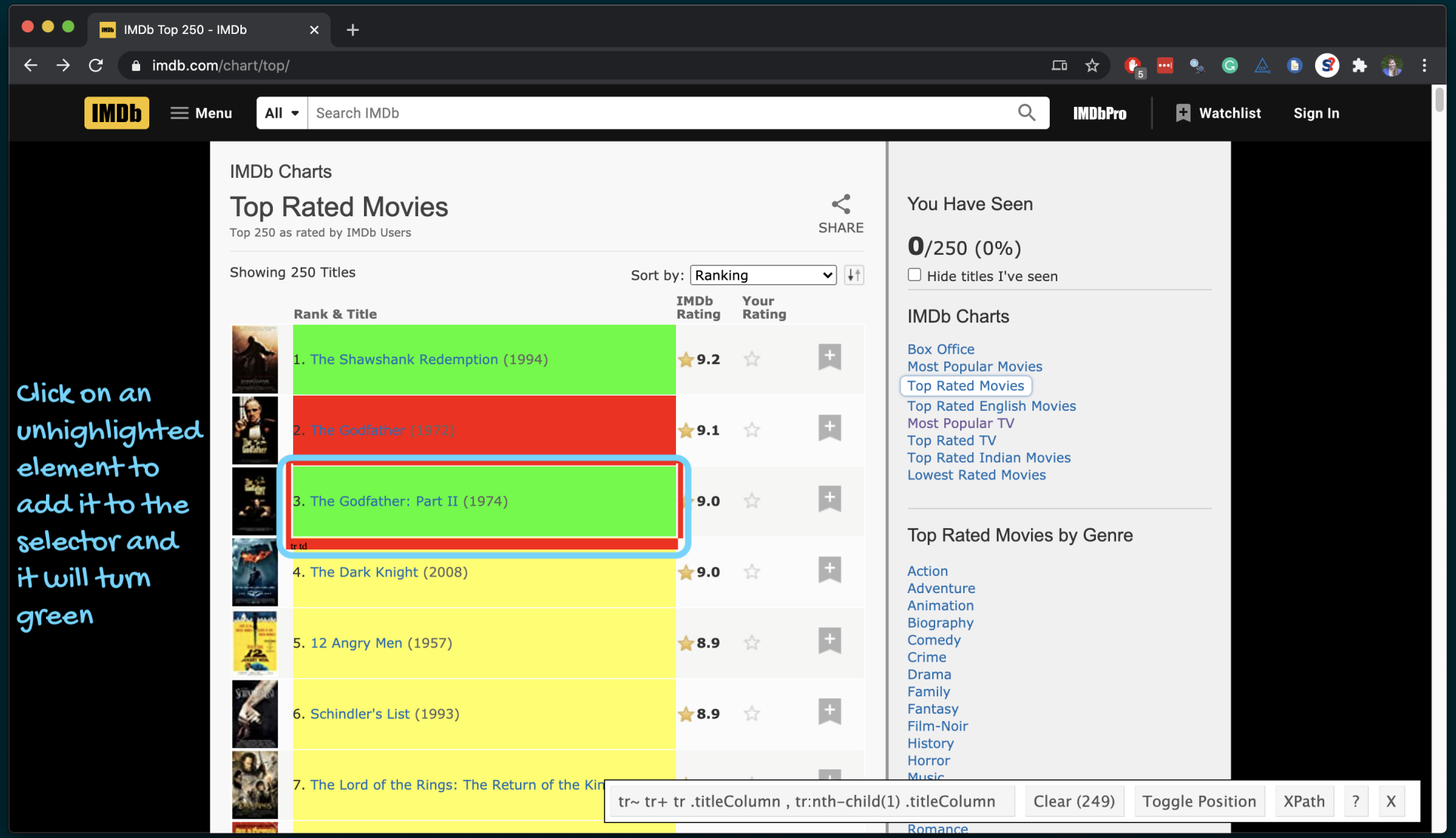



IMDb Top 250 - IMDb
imdb.com/chart/top/
IMDb
Menu
All
Search IMDb
IMDbPro
Watchlist
Sign In
IMDb Charts
Top Rated Movies
Top 250 as rated by IMDb Users
SHARE
Showing 250 Titles
Sort by: Ranking
Rank & Title
IMDb Rating
Your Rating
1. The Shawshank Redemption (1994) 9.2
2. The Godfather (1972) 9.1
3. 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
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
Menu
All
Search IMDb
IMDbPro
Watchlist
Sign In
IMDb Charts
Top Rated Movies
Top 250 as rated by IMDb Users
SHARE
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
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
No valid path found.
Clear
Toggle Position
XPath
?
X


IMDb Top 250 - IMDb
imdb.com/chart/top/
IMDb
Menu
All
Search IMDb
IMDbPro
Watchlist
Sign In
IMDb Charts
Top Rated Movies
Top 250 as rated by IMDb Users
SHARE
Showing 250 Titles
Sort by: Ranking
IMDb Rating
Your Rating
1. The Shawshank Redemption (1994) 9.2
tr td
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
.titleColumn
Clear (250)
Toggle Position
XPath
?
X
Click on a page element, and it will turn green
Selectorgad 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
IMDbPro
Watchlist
Sign In
IMDb Charts
Top Rated Movies
Top 250 as rated by IMDb Users
SHARE
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
tr td
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
Click on a highlighted element to remove it from the selector, and the selection will turn red
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
Romance
tr:nth-child(1) .titleColumn
Clear (1)
Toggle Position
XPath
?
X


IMDb Top 250 - IMDb
imdb.com/chart/top/
IMDb
Menu
All
Search IMDb
IMDbPro
Watchlist
Sign In
IMDb Charts
Top Rated Movies
Top 250 as rated by IMDb Users
SHARE
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
tr td
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 Kin
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
Romance
Click on an unhighlighted element to add it to the selector and it will turn green
tr~ tr+ tr .titleColumn , tr:nth-child(1) .titleColumn
Clear (249)
Toggle Position
XPath
?
X

Using the SelectorGadget

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