

# Web Scraping

Fall 2022

October 19 2022

## Web scraping

the process of downloading, parsing, and extracting data presented in an HTML file and then converting it into a structured format that allows us to analyze it.

## Two different scenarios:

1. Screen scraping: extract data from source code of website, with html parser (easy) or regular expression matching (less easy).
2. Web APIs (application programming interface): website offers a set of structured http requests that return JSON or XML files.



## polite package

- Two main functions `bow` and `scrape` define and realize a web harvesting session
- Builds on awesome toolkits for defining and managing http sessions using `rvest`

<https://www.mncorn.org/corn-facts/>

# Can we scrape this webpage?

## Corn Facts

Minnesota is the 4th largest producer of corn in the United States. Learn more about the largest crop grown in the state that serves as a vital economic driver benefitting all Minnesotans:

2021 Corn Production  
in Minnesota

2021 National Corn  
Production

Components of  
Dent Corn

Minnesota  
Corn Exports

Top Five Markets for  
Minnesota-Grown Corn

Top Uses For Corn In The  
United States

Know Your Corn

polite:: bow()

```
session <- bow("https://www.mncorn.org/corn-facts/",  
               user_agent = "Polite tutorial")
```

```
session  
<polite session> https://www.mncorn.org/corn-facts/  
  User-agent: Polite tutorial  
  robots.txt: 2 rules are defined for 1 bots  
  Crawl delay: 5 sec  
  The path is scrapable for this user-agent
```

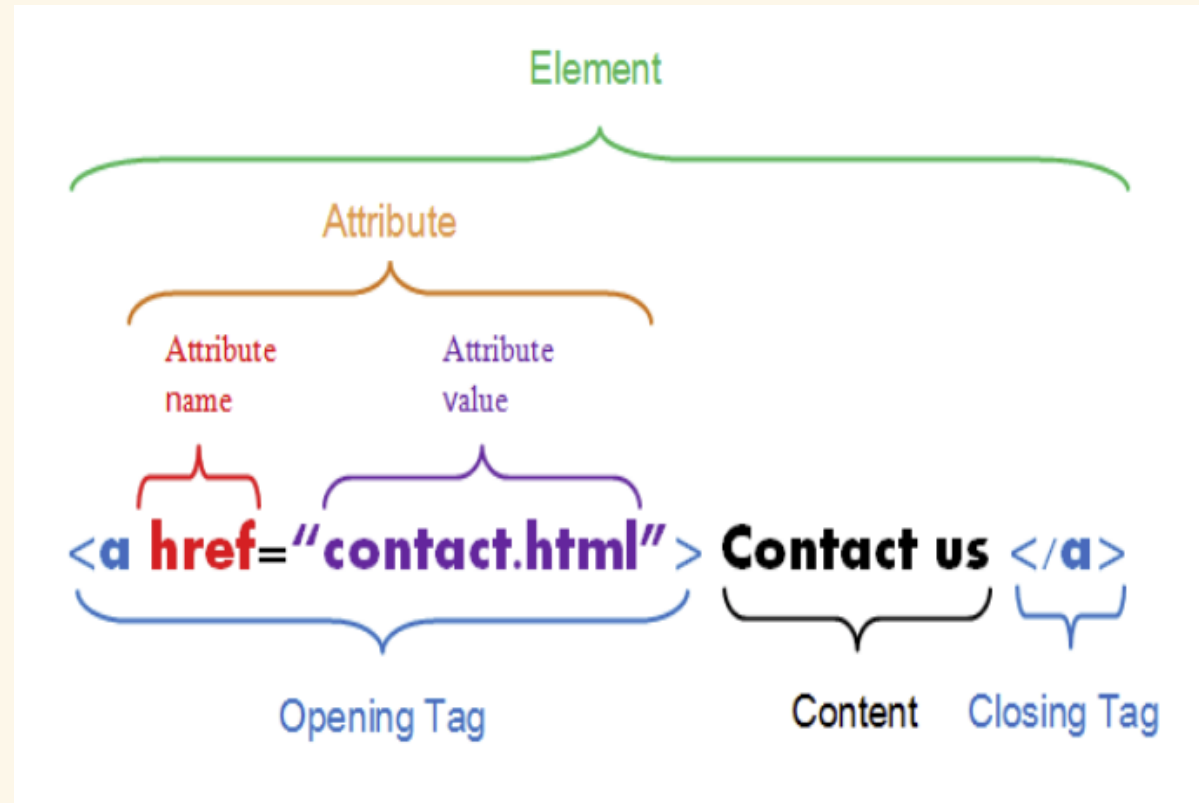
## polite:: scrape()

```
session <- session %>%  
  scrape()
```

```
session  
{html_document}  
<html class="no-js" lang="en-US">  
[1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-8 ...  
[2] <body data-rsssl="1" class="page-template page-template-page-cornfacts pa ...
```

# HyperText Markup Language (HTML)

HTML page consists of series of elements which browsers use to interpret how to display the content





# HyperText Markup Language (HTML)

While it is structured (hierarchical/tree based) it often is 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>
```

# HTML tags

HTML uses **tags** to describe different aspects of document content

| Tag                     | Example  |
|-------------------------|--|
| heading                 | <code>&lt;h1&gt;My Title&lt;/h1&gt;</code>   |
| paragraph               | <code>&lt;p&gt;A paragraph of content...&lt;/p&gt;</code>                          |
| table                   | <code>&lt;table&gt; ... &lt;/table&gt;</code>                                      |
| anchor (with attribute) | <code>&lt;a href="http://www.ratebeer.com"&gt;click here for link&lt;/a&gt;</code> |



Makes basic processing and manipulation of HTML data straight forward.

## Core `rvest` functions

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

# Demo: read tables into R

```
bow("http://www.imdb.com/chart/top/")
  scrape() %>%
  html_nodes("table") %>%
  purrr::pluck(1) %>%
  html_table()
```

```
# A tibble: 250 × 5
  `Rank & Title`
  <lg1> <chr>
1 NA      "1.\n      The Shawshank Redemption\n      (19..
2 NA      "2.\n      The Godfather\n      (1972)"
3 NA      "3.\n      The Dark Knight\n      (2008)"
4 NA      "4.\n      The Godfather Part II\n      (1974)"
5 NA      "5.\n      12 Angry Men\n      (1957)"
6 NA      "6.\n      Schindler's List\n      (1993)"
7 NA      "7.\n      The Lord of the Rings: The Return of ...
8 NA      "8.\n      Pulp Fiction\n      (1994)"
9 NA      "9.\n      The Lord of the Rings: The Fellowship...
10 NA     "10.\n      The Good, the Bad and the Ugly\n      ...
# ... with 240 more rows, and abbreviated variable names 1`I
# 2`Your Rating`
```

# CSS

- CSS (Cascading Style Sheets) is a language that describes how HTML elements should be displayed.
- CSS selectors:
  - shortcuts for selecting HTML elements to style
  - can also be used to extract the content of these elements

# SelectorGadget

SelectorGadget is a point-and-click CSS selector, specifically for Chrome, and it comes as a [Chrome Extension](#) (Click to install!)


The screenshot shows the IMDb website interface. At the top, there's a navigation bar with the IMDb logo, a menu icon, and a search bar. Below the navigation bar, there's a section titled "Best Picture-Winning (Sorted by Year Descending)". This section includes a "View Mode" selector set to "Detailed" and a "Sort by" dropdown menu set to "Year". The list of titles is displayed, with the first item being "1. Nomadland (2020)". The entry for "Nomadland" includes a movie poster, the title "Nomadland (2020)", the rating "R | 107 min | Drama", a star rating of 7.3, a "Rate this" link, a Metascore of 93, a brief description, the director "Chloé Zhao", the stars "Frances McDormand, David Strathairn, Linda May, Gay DeForest", and the total votes "144,701".

IMDb Menu All Search IMDb .lister-item-header a Clear (94)

### Best Picture-Winning (Sorted by Year Descending)

94 titles. View Mode: Compact | Detailed

Sort by: Popularity | A-Z | User Rating | Number of Votes | US Box Office | Runtime | Year ▼ | Release Date | Date of Your Rating | Your Rating

1. **Nomadland** (2020) 

R | 107 min | Drama

★ 7.3 ☆ Rate this **93** Metascore

A woman in her sixties, after losing everything in the Great Recession, embarks on a journey through the American West, living as a van-dwelling modern-day nomad.

Director: Chloé Zhao | Stars: Frances McDormand, David Strathairn, Linda May, Gay DeForest

Votes: 144,701

SelectorGadget: select all elements that are related to that object. Next, de-select anything in yellow you do not want

Top 250 Movies - IMDb

imdb.com/chart/top/

IMDb Menu All Search IMDb IMDb Pro Watchlist Sign In EN

LIMITED-TIME OFFER  
**SAVE OVER 40% WHEN YOU PREPAY FOR A YEAR**  
HBO MAX SIGN UP NOW

IMDb Charts  
IMDb Top 250 Movies  
IMDb Top 250 as rated by regular IMDb voters.

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 Dark Knight (2008)                               | 9.0         |             |
| 4. The Godfather Part II (1974)                         | 9.0         |             |
| 5. 12 Angry Men (1957)                                  | 9.0         |             |
| 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 250 Movies  
Top Rated English Movies  
Most Popular TV Shows  
Top 250 TV Shows  
Top Rated Indian Movies  
Lowest Rated Movies

No valid path found. Clear Toggle Position XPath ? X

Top Rated Movies by Genre



## Read HTML into R

```
webpage <- bow('https://www.imdb.com/search/title/?groups=best_picture_winner&sort=year,desc&count=100&scrape()')
```

```
webpage
{html_document}
<html xmlns:og="http://ogp.me/ns#" xmlns:fb="http://www.facebook.com/2008/fbml">
[1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-8 ...
[2] <body id="styleguide-v2" class="fixed">\n                <img height="1" width ...
```

# Extract titles

Use `html_nodes()` to extract pieces out of HTML documents

```
title_data <- webpage %>% html_nodes(".list-item-header a") %>% html_text()
```

```
title_data
[1] "CODA"
[2] "Nomadland"
[3] "Parasite"
[4] "Green Book"
[5] "The Shape of Water"
[6] "Moonlight"
[7] "Spotlight"
[8] "Birdman or (The Unexpected Virtue of Ignorance)"
[9] "12 Years a Slave"
[10] "Argo"
[11] "The Artist"
```

# Scrapped table

Show  entries

Search:

|   | Year | Title      | Description   | Runtime | Rating | Votes  |
|---|------|------------|---|---------|--------|--------|
| 1 | 2021 | CODA       | As a CODA (Child of Deaf Adults) Ruby is the only hearing person in her deaf family. When the family's fishing business is threatened, Ruby finds herself torn between pursuing her passion at Berklee College of Music and her fear of abandoning her parents. | 111     | 8      | 128025 |
| 2 | 2020 | Nomadland  | A woman in her sixties, after losing everything in the Great Recession, embarks on a journey through the American West, living as a van-dwelling modern-day nomad.  | 107     | 7.3    | 159215 |
| 3 | 2019 | Parasite   | Greed and class discrimination threaten the newly formed symbiotic relationship between the wealthy Park family and the destitute Kim clan.   | 132     | 8.5    | 783773 |
| 4 | 2018 | Green Book | A working-class Italian-American bouncer becomes the driver for an African-American classical pianist on a tour of venues through the 1960s American South.   | 130     | 8.2    | 484175 |

# Group Activity 1

15:00



- Let's go over to maize server/ local Rstudio and our class [moodle](#)
- Get the class activity 16.Rmd file
- Work on activity 1
- Ask me questions

## IMDb Charts

# Most Popular TV Shows

As determined by IMDb Users



Showing 100 Titles

Sort by: Ranking

| Rank & Title   |   | IMDb Rating   | Your Rating   |   |
|--|---|---|---|---|
|    | <a href="#">House of the Dragon</a> (2022)<br>1 (no change)   |  8.6   |    |    |
|    | <a href="#">The Lord of the Rings: The Rings of Power</a> (2022)<br>2 (no change)   |  6.9   |    |    |
|   | <a href="#">The Midnight Club</a> (2022)<br>3 (  11)                           |  6.7   |    |    |
|  | <a href="#">Dahmer - Monster: The Jeffrey Dahmer Story</a> (2022)<br>4 (  1) |  8.1 |  |  |
|  | <a href="#">She-Hulk: Attorney at Law</a> (2022)<br>6 (  1)                  |  5.1 |  |  |

You Have Seen

0/100 (0%)

☐ Hide titles I've seen

## IMDb Charts

[Box Office](#)

[Most Popular Movies](#)

[Top 250 Movies](#)

[Top Rated English Movies](#)

[Most Popular TV Shows](#)

[Top 250 TV Shows](#)

[Top Rated Indian Movies](#)

[Lowest Rated Movies](#)

## Popular TV by Genre

[Action](#)

[Adventure](#)

[Animation](#)

[Biography](#)

[Comedy](#)

[Crime](#)

[Documentary](#)

# Scrapped table

Show 

10 ▾

 entries

Search:

|    | rank ▾ | name ▾                                     | score ▾ | year ▾ |
|----|--------|--|---------|--------|
| 1  | 1      | House of the Dragon                        | 8.6     | 2022   |
| 2  | 2      | The Lord of the Rings: The Rings of Power  | 6.9     | 2022   |
| 3  | 3      | The Midnight Club                          | 6.7     | 2022   |
| 4  | 4      | Dahmer - Monster: The Jeffrey Dahmer Story | 8.1     | 2022   |
| 5  | 6      | She-Hulk: Attorney at Law                  | 5.1     | 2022   |
| 6  | 7      | The Watcher                                | 6.7     | 2022   |
| 7  | 8      | Andor                                      | 8.1     | 2022   |
| 8  | 9      | The Walking Dead                           | 8.1     | 2010   |
| 9  | 10     | Game of Thrones                            | 9.2     | 2011   |
| 10 | 11     | Wednesday                                  |         | 2022   |

## Group Activity 2

10:00



- Work on activity 2
- Ask me questions