

Web Scraping with rvest

Ways to scrape data

- **Text pattern matching:** Another simple yet powerful approach to extract information from the web is by using regular expression matching facilities of programming languages. You can learn more about regular expressions.
- **API Interface:** Many websites like Facebook, Twitter, LinkedIn, etc. provides public and/ or private APIs which can be called using standard code for retrieving the data in the prescribed format.
- **DOM Parsing:** By using the web browsers, programs can retrieve the dynamic content generated by client-side scripts. It is also possible to parse web pages into a DOM tree, based on which programs can retrieve parts of these pages.

HTML DOMS

- **Document** object model.

The DOM is the way Javascript sees its containing pages' data. It is an object that includes how the HTML/XHTML/XML is formatted, as well as the browser state.

- A DOM element is something like a DIV, HTML, BODY element on a page. You can add classes to all of these using CSS, or interact with them using JS.

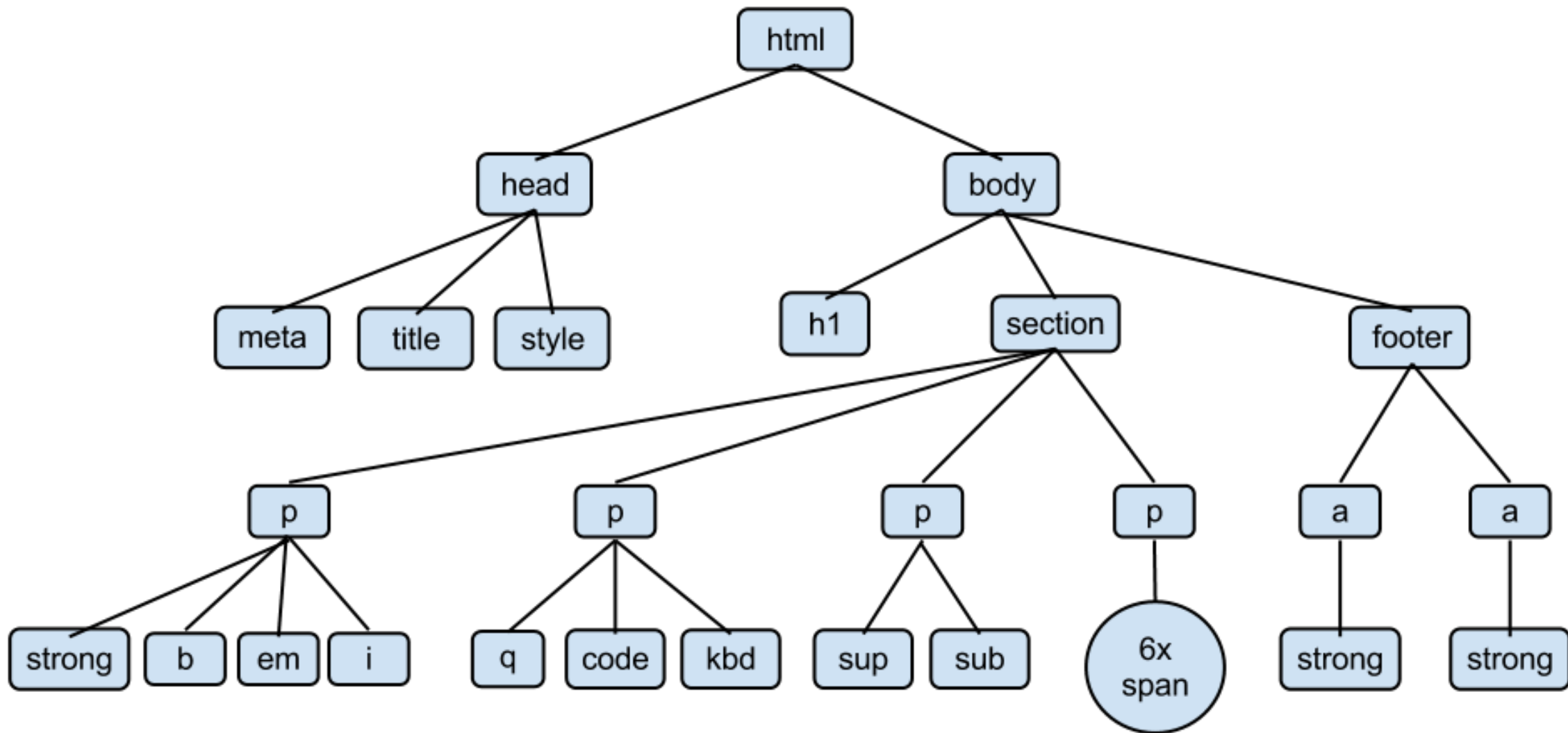


Fig.

An example of an HTML DOM tree

Web scraping

So far we have used data that can be downloaded in a structured, tabular format (such as CSV).

However, sometimes data is not available in an easily downloadable and importable form.

Consider <http://www.imdb.com>, which compiles a great deal of information about movies in a searchable way, but doesn't make the information easy to export to a format that can be read into R. How can we utilize IMDB's enormous database of movie data then?

Today, we will discuss how to harvest and tidy unstructured data from the web using the `rvest` package.

Web scraping with rvest

The `rvest` package is designed with the same conventions as the `tidyverse` packages.

The data is always the first argument, so it plays nicely with the `%>%` piping operator.

To scrape web data in R, we will need to do the following:

- ▶ Parse the HTML from the web page using `read_html`
- ▶ Identify the CSS selectors that correspond to the fields of interest
- ▶ Extract the data using `html_nodes()` and `html_text()`
- ▶ Perform any additional data wrangling necessary to tidy the final dataset

SelectorGadget

After reading in a web page's HTML source, we still have to do some processing to extract the data we want.

We will use an additional web tool called SelectorGadget to identify the CSS selectors that correspond to the data fields of interest.

- ▶ Bookmark tool link or available as a Chrome extension
- ▶ Visually identify the CSS associated with page elements
- ▶ Available at <http://selectorgadget.com>

(If you are experienced with CSS, you can also simply view the page source and identify the elements this way.)

IMDB cast example

Suppose we want to create a dataset for the cast of *The Last Jedi*.

We can scrape data from <http://www.imdb.com/title/tt2527336/>:

```
library(rvest)
jedi <- read_html("http://www.imdb.com/title/tt2527336/")
jedi

## {xml_document}
## <html xmlns:og="http://ogp.me/ns#" xmlns:fb="http://www.faceb
## [1] <head>\n<meta http-equiv="Content-Type" content="text/htm
## [2] <body id="styleguide-v2" class="fixed">\n<script>\n    if
```


Select cast (attempt #1)

The screenshot shows the IMDb website for the movie "Star Wars: The Last Jedi". The "Cast overview, first billed only:" section lists the following cast members:

Cast Member	Character
Mark Hamill	Darth Vader
Carrie Fisher	Lia Organa
Adam Driver	Yoda
Daisy Ridley	Rey
John Boyega	Finn
Oscar Isaac	Poe Dameron
Andy Serkis	Snoke
Lupita Nyong'o	Maz Kanata
Samuel L. Jackson	General Hux

The name "Lupita Nyong'o" is highlighted with an orange box, and the character name "Maz Kanata" is missing. The search bar at the bottom contains the text "#titleCast a".

Additional details from the screenshot include:

- Browser: Chrome, URL: imdb.com
- Page Title: Cast overview, first billed only:
- Movie Title: Star Wars: The Last Jedi
- IMDb Rating: 7.5/10
- Follow @imdb: 4.24M followers
- Share this Rating: Use the HTML below.
- User Polls: Most Anticipated Blockbuster of

Figure 1: Maz Kanata is missing...

Select actors (attempt #1)

The screenshot shows a web browser window with the URL `imdb.com`. The page displays the cast overview for `Star Wars: The Last Jedi`. A yellow vertical bar highlights the first actor, Luke Skywalker / Dobbu Scay. The cast list includes:

Actor	Character
Luke Skywalker / Dobbu Scay	Luke Skywalker / Dobbu Scay
Carrie Fisher	Leia Organa
Adam Driver	Kylo Ren
Daisy Ridley	Rey
John Boyega	Finn
Oscar Isaac	Poe Dameron
Andy Serkis	Snoke
Lupita Nyong'o	Maz Kanata
Domhnall Gleeson	General Hux

At the bottom of the browser window, a search bar contains the text `#titleCast .itemprop`. To the right of the search bar are buttons for `Clear (30)`, `Toggle Position`, `XPath`, `Help`, and `X`.

Figure 2: Select actors first...

```
jedi %>% html_nodes("#titleCast .itemprop") %>% html_text()
```

```
## [1] "\n Mark Hamill\n          "  
## [2] "Mark Hamill"  
## [3] "\n Carrie Fisher\n          "  
## [4] "Carrie Fisher"  
## [5] "\n Adam Driver\n          "  
## [6] "Adam Driver"  
## [7] "\n Daisy Ridley\n          "  
## [8] "Daisy Ridley"  
## [9] "\n John Boyega\n          "  
## [10] "John Boyega"  
## [11] "\n Oscar Isaac\n          "  
## [12] "Oscar Isaac"  
## [13] "\n Andy Serkis\n          "  
## [14] "Andy Serkis"  
## [15] "\n Lupita Nyong'o\n          "  
## [16] "Lupita Nyong'o"  
## [17] "\n Domhnall Gleeson\n          "  
## [18] "Domhnall Gleeson"  
## [19] "\n Anthony Daniels\n          "  
## [20] "Anthony Daniels"
```

Select actors (attempt #2)

The screenshot shows the IMDb website for the movie "Star Wars: The Last Jedi". The "Cast" section is visible, listing actors and their roles. A red box highlights the first actor's name, "Dylan McDermott", which is selected by the XPath selector ".itemprop .itemprop" shown in the bottom toolbar. The browser's address bar shows the URL "www.imdb.com/title/tt2527336/". The right sidebar contains social media links and a rating of 7.5.

Actor	Role
Dylan McDermott	Luke Skywalker / Dobbu Scay
Marrie Fisher	Lela Organa
Adam Driver	Kylo Ren
Daisy Ridley	Rey
John Boyega	Finn
Oscar Isaac	Poe Dameron
Andy Serkis	Snoke
Lupita Nyong'o	Maz Kanata

Figure 3: Select actors w/out framing cells...

```
jedi %>% html_nodes(".itemprop .itemprop") %>% html_text()
```

```
## [1] "Mark Hamill"           "Carrie Fisher"       "Adam Driver"
## [4] "Daisy Ridley"         "John Boyega"         "Oscar Isaac"
## [7] "Andy Serkis"          "Lupita Nyong'o"     "Domhnall Gleeson"
## [10] "Anthony Daniels"     "Gwendoline Christie" "Kelly Marie Brown"
## [13] "Laura Dern"           "Benicio Del Toro"    "Frank Oz"
```

Select characters (attempt #1)

The screenshot shows a web browser window with the URL `imdb.com`. The page displays the cast overview for `Star Wars: The Last Jedi`. A selection tool is active, showing a list of characters on the left and a corresponding list of selected characters on the right. The selected characters are highlighted in yellow. The selection tool includes a search bar with the text `#titleCast .character`, a `Clear (15)` button, and buttons for `Toggle Position`, `XPath`, `Help`, and `X`.

Character	Selected
Mark Hamill	Luke Skywalker / Palpathe Slave
Carrie Fisher	Leia Organa
Adam Driver	Kylo Ren
Daisy Ridley	Rey
John Boyega	Finn
Oscar Isaac	Poe Dameron
Andy Serkis	Snoke
Lupita Nyong'o	Maz Kanata
Domhnall Gleeson	General Hux

Additional page elements include a navigation menu (Wikipedia, Northeastern, Technology, Developer, Statistics, Teaching, Funding, Running, Writing, Anime), a social media follow button for `@imdb` (4.24M followers), a rating section for `Star Wars: The Last Jedi` (IMDb 7.5/10), and a `User Polls` section for `Most Anticipated Blockbuster of`.

Figure 4: Select characters ...

```
jedi %>% html_nodes("#titleCast .character") %>% html_text()
```

```
## [1] "\n           \n           Luke Skywalker / \n\n## [2] "\n           \n           Leia Organa \n\n## [3] "\n           \n           Kylo Ren \n\n## [4] "\n           \n           Rey \n\n## [5] "\n           \n           Finn \n\n## [6] "\n           \n           Poe Dameron \n\n## [7] "\n           \n           Snoke \n\n## [8] "\n           \n           Maz Kanata \n\n## [9] "\n           \n           General Hux \n\n## [10] "\n           \n           C-3PO \n\n## [11] "\n           \n           Captain Phasma \n\n## [12] "\n           \n           Rose Tico \n\n## [13] "\n           \n           Vice Admiral Holdo \n\n## [14] "\n           \n           DJ \n\n## [15] "\n           \n           Yoda \n           \n           \n           (voice)\n
```

Select characters (attempt #2)

Cast overview, first billed only:

	Mark Hamill	... Darth Vader / Dobbu Scar
	Carrie Fisher	... Leia Organa
	Adam Driver	... Kylo Ren
	Daisy Ridley	... Rey
	John Boyega	... Finn
	Oscar Isaac	... Han Solo
	Andy Serkis	... Snoke
	Lupita Nyong'o	... Maz Kanata
	Domhnall Gleeson	... General Hux

.character a

Clear (15) Toggle Position XPath Help X

Figure 5: Maz Kanata is missing again...


```
jedi %>% html_nodes(".character a") %>% html_text()
```

```
## [1] "Luke Skywalker"      "Dobbu Scay"          "Leia Organa"  
## [4] "Kyro Ren"            "Rey"                 "Finn"  
## [7] "Poe Dameron"        "Snoke"               "General Hux"  
## [10] "C-3PO"              "Captain Phasma"     "Rose Tico"  
## [13] "Vice Admiral Holdo" "DJ"                  "Yoda"
```

Clean up attempt #1 instead

```
library(stringr)
jedi %>% html_nodes("#titleCast .character") %>% html_text() %>%
  str_replace_all("  ", "") %>%
  str_replace_all("\n", "") %>%
  str_replace_all("/", "/ ") %>%
  str_trim()
```

```
## [1] "Luke Skywalker / Dobbu Scay" "Leia Organa"
## [3] "Kylo Ren"                    "Rey"
## [5] "Finn"                        "Poe Dameron"
## [7] "Snoko"                       "Maz Kanata"
## [9] "General Hux"                 "C-3PO"
## [11] "Captain Phasma"             "Rose Tico"
## [13] "Vice Admiral Holdo"        "DJ"
## [15] "Yoda (voice)"
```

```
library(tidyverse)
actors <- jedi %>%
  html_nodes(".itemprop .itemprop") %>%
  html_text()
characters <- jedi %>%
  html_nodes("#titleCast .character") %>%
  html_text() %>%
  str_replace_all(" ", "") %>%
  str_replace_all("\n", "") %>%
  str_replace_all("/", "/ ") %>%
  str_trim()
cast <- tibble(actors=actors, characters=characters)
```

cast

A tibble: 15 x 2

##	actors	characters
##	<chr>	<chr>
## 1	Mark Hamill	Luke Skywalker / Dobbu Scay
## 2	Carrie Fisher	Leia Organa
## 3	Adam Driver	Kylo Ren
## 4	Daisy Ridley	Rey
## 5	John Boyega	Finn
## 6	Oscar Isaac	Poe Dameron
## 7	Andy Serkis	Snoke
## 8	Lupita Nyong'o	Maz Kanata
## 9	Domhnall Gleeson	General Hux
## 10	Anthony Daniels	C-3P0
## 11	Gwendoline Christie	Captain Phasma
## 12	Kelly Marie Tran	Rose Tico
## 13	Laura Dern	Vice Admiral Holdo
## 14	Benicio Del Toro	DJ
## 15	Frank Oz	Yoda (voice)