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.

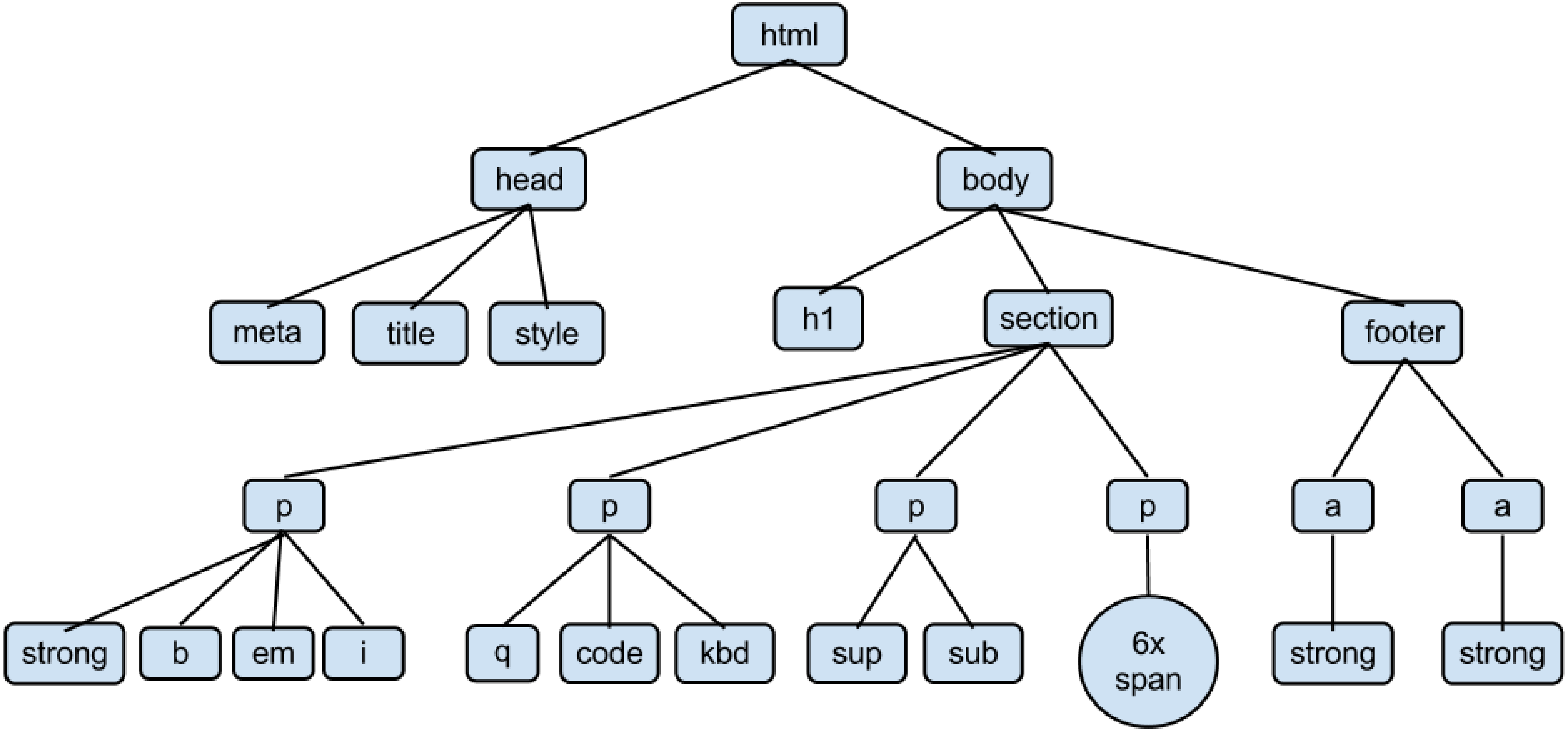


Fig. An example of an HTML DOM tree

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 availabe 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

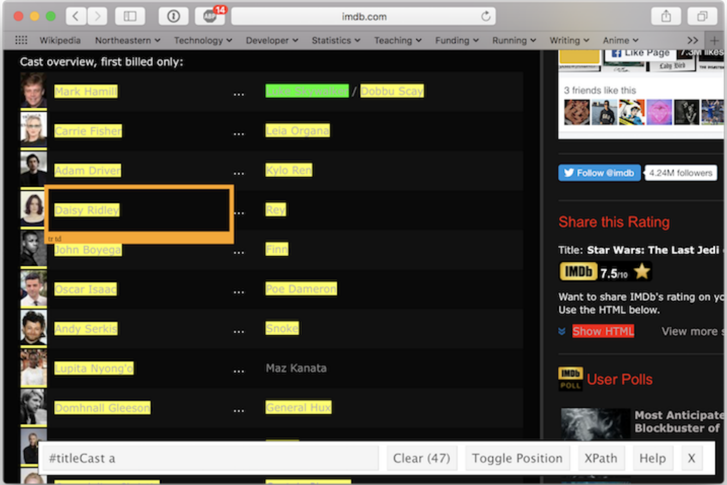
Suppose we want to create a dataset for the cast of *The Last Jedi*.

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

```r
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)



Cast overview, first billed only:

Mark Hamill	...	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

Like Page 7.3M likes

3 friends like this

Follow @imdb 4.24M followers

Share this Rating

Title: Star Wars: The Last Jedi

IMDb 7.5/10

Want to share IMDb's rating on yo
Use the HTML below.

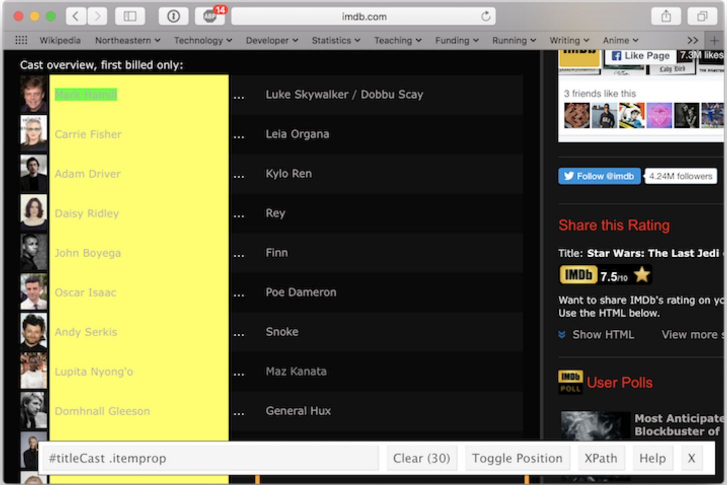
Show HTML View more

User Polls

Most Anticipate
Blockbuster of

Figure 1: Maz Kanata is missing. . .

Select actors (attempt #1)



Cast overview, first billed only:

Mark Hamill	...	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

3 friends like this

Share this Rating

Title: Star Wars: The Last Jedi

IMDb 7.5/10

Want to share IMDb's rating on yo

Use the HTML below.

Show HTML View more

User Polls

Most Anticipate
Blockbuster of

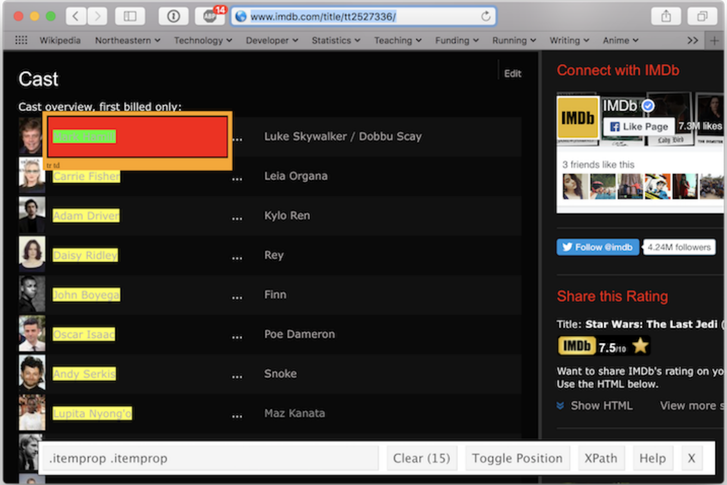
#titleCast .itemprop

Figure 2: Select actors first. . .

```r
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)



Cast

Cast overview, first billed only:

Mark Hamill	...	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

Connect with IMDb

Share this Rating

Title: Star Wars: The Last Jedi (

IMDb 7.5/10

Want to share IMDb's rating on yo
Use the HTML below.

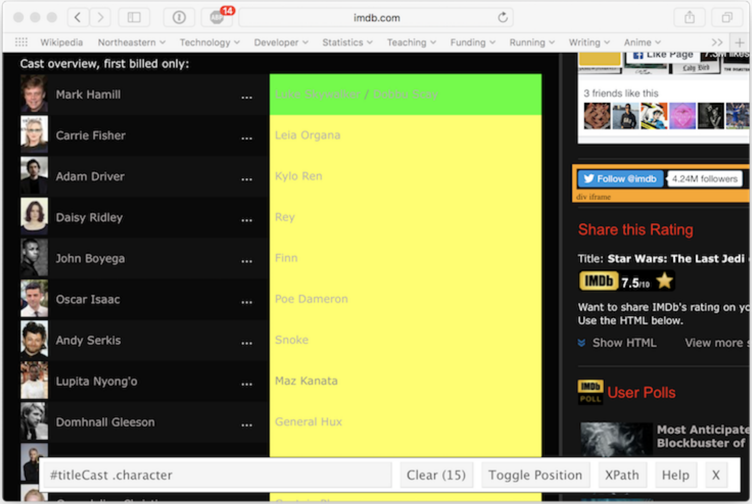
.itemprop .itemprop

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 Gl
## [10] "Anthony Daniels"       "Gwendoline Christie" "Kelly Marie
## [13] "Laura Dern"            "Benicio Del Toro"       "Frank Oz"
```

Select characters (attempt #1)



Cast overview, first billed only:

Mark Hamill	...	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

Share this Rating

Title: **Star Wars: The Last Jedi**

IMDb 7.5/10

Want to share IMDb's rating on yo
Use the HTML below.

Show HTML View more

User Polls

Most Anticipate
Blockbuster of

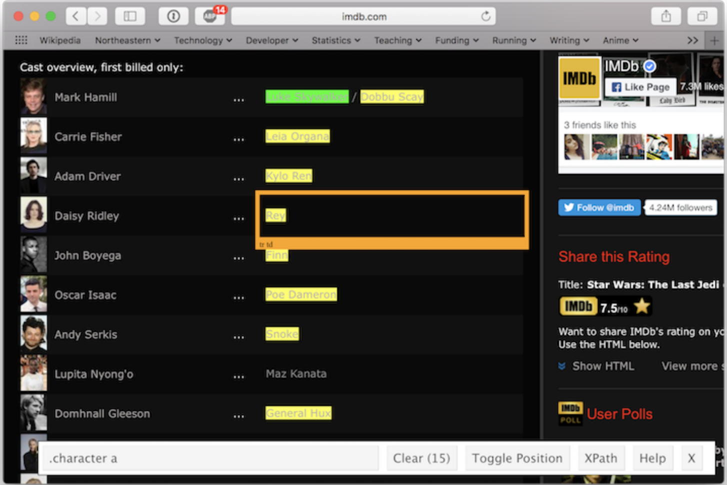
#titleCast .character

Figure 4: Select characters . . .

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

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

Select characters (attempt #2)



Cast overview, first billed only:

Mark Hamill	...	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

Share this Rating

Title: Star Wars: The Last Jedi

Want to share IMDb's rating on yo
Use the HTML below.

.character a

Figure 5: Maz Kanata is missing again. . .

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

```
##  [1] "Luke Skywalker"     "Dobbu Scay"         "Leia Organa"
##  [4] "Kylo 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

```r
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] "Snoke"                       "Maz Kanata"
##  [9] "General Hux"                 "C-3PO"
## [11] "Captain Phasma"              "Rose Tico"
## [13] "Vice Admiral Holdo"          "DJ"
## [15] "Yoda (voice)"
```

```r
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-3PO
## 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)
```