Data Scraping

Byteflow Dynamics

10/21/2017

Contents

1 Data scraping using Selectorgadget

Selectorgadget

Interactive bookmarklet which allows you to interactively figure out which CSS selector you need to scrape the desired data from a page.

Installation

Add Selectorgadget to your bookmark

Using Selectorgate

- Click on the element you want to select.

Selectorgadget will make a first guess at what css selector you want. It's likely to be bad since it only has one example to learn from, but it's a start. Elements that match the selector will be highlighted in yellow.

- Click on elements that shouldn't be selected.

They will turn red. Click on elements that should be selected. They will turn green.

- Iterate until only the elements you want are selected. Selectorgadget isn't perfect and sometimes won't be able to find a useful css selector. Sometimes starting from a different element helps.

2 Example

```r
library(rvest)

# add the link from where you want to scrape the data
html <- read_html("http://www.imdb.com/title/tt1490017/")

# copy the CSS into the html_nodes
cast <- html %>%
  html_nodes("#titleCast .itemprop")
cast
```

```
## {xml_nodeset (30)}
##  [1] <td class="itemprop" itemprop="actor" itemscope itemtype="http://sc ...
##  [2] <span class="itemprop" itemprop="name">Will Arnett</span>
##  [3] <td class="itemprop" itemprop="actor" itemscope itemtype="http://sc ...
##  [4] <span class="itemprop" itemprop="name">Elizabeth Banks</span>
##  [5] <td class="itemprop" itemprop="actor" itemscope itemtype="http://sc ...
##  [6] <span class="itemprop" itemprop="name">Craig Berry</span>
##  [7] <td class="itemprop" itemprop="actor" itemscope itemtype="http://sc ...
##  [8] <span class="itemprop" itemprop="name">Alison Brie</span>
##  [9] <td class="itemprop" itemprop="actor" itemscope itemtype="http://sc ...
## [10] <span class="itemprop" itemprop="name">David Burrows</span>
## [11] <td class="itemprop" itemprop="actor" itemscope itemtype="http://sc ...
## [12] <span class="itemprop" itemprop="name">Anthony Daniels</span>
## [13] <td class="itemprop" itemprop="actor" itemscope itemtype="http://sc ...
## [14] <span class="itemprop" itemprop="name">Charlie Day</span>
## [15] <td class="itemprop" itemprop="actor" itemscope itemtype="http://sc ...
## [16] <span class="itemprop" itemprop="name">Amanda Farinos</span>
## [17] <td class="itemprop" itemprop="actor" itemscope itemtype="http://sc ...
## [18] <span class="itemprop" itemprop="name">Keith Ferguson</span>
## [19] <td class="itemprop" itemprop="actor" itemscope itemtype="http://sc ...
## [20] <span class="itemprop" itemprop="name">Will Ferrell</span>
## ...
```

```r
# to obtain the names use html_text

cast %>%
  html_text()
```

```
##  [1] "\n Will Arnett\n          "     "Will Arnett"
##  [3] "\n Elizabeth Banks\n          " "Elizabeth Banks"
##  [5] "\n Craig Berry\n          "     "Craig Berry"
##  [7] "\n Alison Brie\n          "     "Alison Brie"
##  [9] "\n David Burrows\n          "   "David Burrows"
## [11] "\n Anthony Daniels\n          " "Anthony Daniels"
## [13] "\n Charlie Day\n          "     "Charlie Day"
## [15] "\n Amanda Farinos\n          "  "Amanda Farinos"
## [17] "\n Keith Ferguson\n          "  "Keith Ferguson"
## [19] "\n Will Ferrell\n          "    "Will Ferrell"
## [21] "\n Will Forte\n          "      "Will Forte"
## [23] "\n Dave Franco\n          "     "Dave Franco"
## [25] "\n Morgan Freeman\n          "  "Morgan Freeman"
## [27] "\n Todd Hansen\n          "     "Todd Hansen"
## [29] "\n Jonah Hill\n          "      "Jonah Hill"
```

```r
# we see the cast names twice.
# thats because we have selected the cell and the text inside the cell
# We can digg in one more layer to select just the names

cast <- html %>%
  html_nodes("#titleCast span.itemprop")

# to get the text

html_text(cast)
```

```
##  [1] "Will Arnett"     "Elizabeth Banks" "Craig Berry"
##  [4] "Alison Brie"     "David Burrows"   "Anthony Daniels"
##  [7] "Charlie Day"     "Amanda Farinos"  "Keith Ferguson"
## [10] "Will Ferrell"    "Will Forte"      "Dave Franco"
## [13] "Morgan Freeman"  "Todd Hansen"     "Jonah Hill"
```

3 Ratings

TaskSelect the rating of the movie Lego

```
lego_movie <- html("http://www.imdb.com/title/tt1490017/")

lego_movie %>%
  html_node("strong span") %>%
  html_text() %>%
  as.numeric()
```

```
## [1] 7.8
```