

# CSS Locators

WEB SCRAPING IN PYTHON



**Thomas Laetsch**  
Data Scientist, NYU

# Rosetta CSStone

- `/` replace by `>` (except first character)
  - **XPath:** `/html/body/div`
  - **CSS Locator:** `html > body > div`
- `//` replaced by a blank space (except first character)
  - **XPath:** `//div/span//p`
  - **CSS Locator:** `div > span p`
- `[N]` replaced by `:nth-of-type(N)`
  - **XPath:** `//div/p[2]`
  - **CSS Locator:** `div > p:nth-of-type(2)`

# Rosetta CSStone

## XPATH

```
xpath = '/html/body//div/p[2]'
```

## CSS

```
css = 'html > body div > p:nth-of-type(2)'
```

# Attributes in CSS

- To find an element by class, use a period `.`
  - Example: `p.class-1` selects all paragraph elements belonging to `class-1`
- To find an element by id, use a pound sign `#`
  - Example: `div#uid` selects the `div` element with `id` equal to `uid`

# Attributes in CSS

Select paragraph elements within class `class1` :

```
css_locator = 'div#uid > p.class1'
```

Select all elements whose class attribute belongs to `class1` :

```
css_locator = '.class1'
```

# Class Status

```
css = '.class1'
```

 `<p class="class-1"> ... </p>`


 `<div class="class-1 class-2"> ... </div>`

 `<p class="class-1 2"> ... </p>`

# Class Status

```
xpath = '//*[@class="class1"]'
```

 `<p class="class-1"> ... </p>`

 `<div class="class-1 class-2"> ... </div>`

 `<p class="class-1 2"> ... </p>`

# Class Status

```
xpath = '//*[@contains(@class, "class1")]'
```

☒ `<p class="class-1"> ... </p>`

☒ `<div class="class-1 class-2"> ... </div>`

☒ `<p class="class-1 2"> ... </p>`



# Selectors with CSS

```
from scrapy import Selector

html = '''
<html>
  <body>
    <div class="hello datacamp">
      <p>Hello World!</p>
    </div>
    <p>Enjoy DataCamp!</p>
  </body>
</html>
'''

sel = Selector( text = html )
```

```
>>> sel.css("div > p")
out: [<Selector xpath='...' data='<p>Hello World!</p>'>]

>>> sel.css("div > p").extract()
out: [ '<p>Hello World!</p>' ]
```

# C(SS) You Soon!

WEB SCRAPING IN PYTHON

# Attribute and Text Selection

WEB SCRAPING IN PYTHON



Thomas Laetsch  
Data Scientist, NYU

# You Must have Guts to use your Colon

- Using XPath: `<xpath-to-element>/@attr-name`

```
xpath = ' //div[@id="uid"]/a/@href'
```

- Using CSS Locator: `<css-to-element>::attr(attr-name)`

```
css_locator = 'div#uid > a::attr(href)'
```

# Text Extraction

```
<p id="p-example">
  Hello world!
  Try <a href="http://www.datacamp.com">DataCamp</a> today!
</p>
```

- In XPath use `text()`

```
sel.xpath(' //p[@id="p-example"]/text()').extract()
# result: ['\n Hello world!\n Try ', ' today!\n']
```

```
sel.xpath(' //p[@id="p-example"]//text()').extract()
# result: ['\n Hello world!\n Try ', 'DataCamp', ' today!\n']
```

# Text Extraction

```
<p id="p-example">
  Hello world!
  Try <a href="http://www.datacamp.com">DataCamp</a> today!
</p>
```

- For CSS Locator, use `::text`

```
sel.css('p#p-example::text').extract()
# result: ['\n Hello world!\n Try ', ' today!\n']
```

```
sel.css('p#p-example ::text').extract()
# result: ['\n Hello world!\n Try ', 'DataCamp', ' today!\n']
```

# Scoping the Colon

WEB SCRAPING IN PYTHON

# Getting Ready to Crawl

WEB SCRAPING IN PYTHON



**Thomas Laetsch**  
Data Scientist, NYU



# Let's Respond

## Selector vs Response:

- The Response has all the tools we learned with Selectors:
  - `xpath` and `css` methods followed by `extract` and `extract_first` methods.
- The Response also keeps track of the url where the HTML code was loaded from.
- The Response helps us move from one site to another, so that we can "crawl" the web while scraping.

# What We Know!

- `xpath` method works like a Selector

```
response.xpath( '//div/span[@class="bio"]' )
```

- `css` method works like a Selector

```
response.css( 'div > span.bio' )
```

- Chaining works like a Selector

```
response.xpath( '//div' ).css( 'span.bio' )
```

- Data extraction works like a Selector

```
response.xpath( '//div' ).css( 'span.bio' ).extract()  
response.xpath( '//div' ).css( 'span.bio' ).extract_first()
```

# What We Don't Know

- The `response` keeps track of the URL within the response url variable.

```
response.url  
>>> 'http://www.DataCamp.com/courses/all'
```

- The `response` lets us "follow" a new link with the `follow()` method

```
# next_url is the string path of the next url we want to scrape  
response.follow( next_url )
```

- We'll learn more about `follow` later.

# In Response

## WEB SCRAPING IN PYTHON

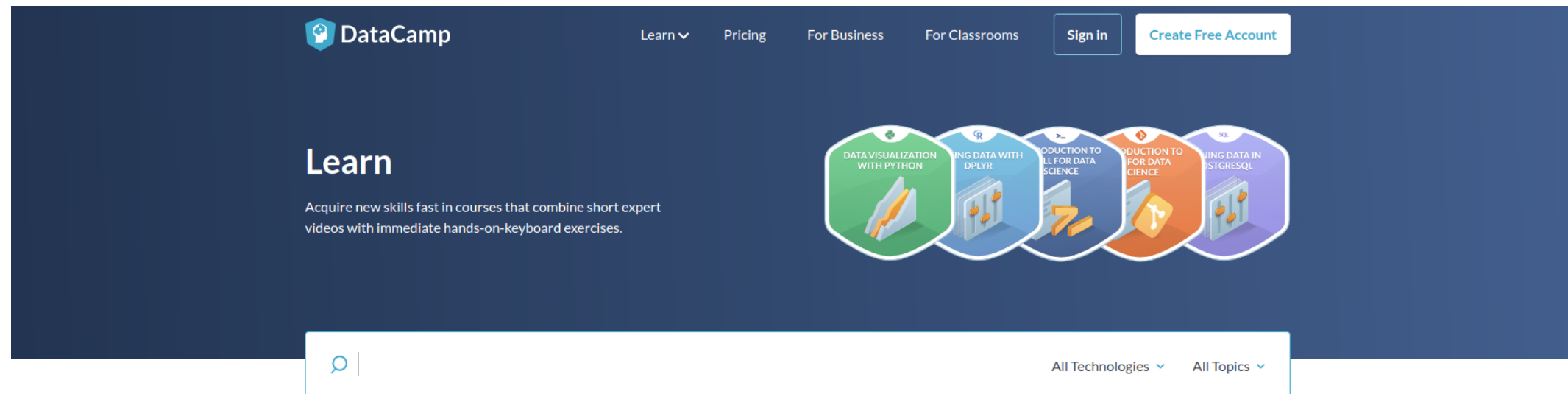
# Scraping For Reals

WEB SCRAPING IN PYTHON









Thomas Laetsch  
Data Scientist, NYU

# DataCamp Site



## All Data Science Courses

|   |  |   |
|---|--|---|
| <br><b>Introduction to R</b><br><br>Master the basics of data analysis by manipulating common data structures such as vectors, matrices and data frames.<br><br>⌚ 4 hours<br><br> <b>JONATHAN CORNELISSEN</b><br>Co-founder and CEO of DataCamp | <br><b>Data Analysis in R, the data.table Way</b><br><br>Master core concepts in data manipulation such as subsetting, updating, indexing and joining your data using data.table.<br><br>⌚ 4 hours<br><br> <b>MATT DOWLE</b><br>Author of data.table | <br><b>Data Manipulation in R with dplyr</b><br><br>Master techniques for data manipulation using the select, mutate, filter, arrange, and summarise functions in dplyr.<br><br>⌚ 4 hours<br><br> <b>GARRETT GROLEMUND</b><br>Data Scientist at RStudio |
|---|--|---|

<https://www.datacamp.com/courses/all>

# What's the Div, Yo?

```
# response loaded with HTML from https://www.datacamp.com/courses/all
```

```
course_divs = response.css('div.course-block')
```

```
print( len(course_divs) )
```

```
>>> 185
```

# Inspecting course-block



## Introduction to R

Master the basics of data analysis by manipulating common data structures such as vectors, matrices and data frames.

🕒 4 hours



**JONATHAN CORNELISSEN**

Co-founder and CEO of  
DataCamp

```
first_div = course_divs[0]
children = first_div.xpath('./*')
print( len(children) )
>>> 3
```



# The first child



## Introduction to R

Master the basics of data analysis by manipulating common data structures such as vectors, matrices and data frames.

🕒 4 hours



JONATHAN CORNELISSEN

Co-founder and CEO of  
DataCamp

```
first_div = course_divs[0]
children = first_div.xpath('./*')
```

```
first_child = children[0]
print( first_child.extract() )
>>> <a class=... />
```

# The second child



## Introduction to R

Master the basics of data analysis by manipulating common data structures such as vectors, matrices and data frames.

⌚ 4 hours



**JONATHAN CORNELISSEN**

Co-founder and CEO of  
DataCamp

```
first_div = course_divs[0]
children = first_div.xpath('./*')
```

```
second_child = children[1]
print( second_child.extract() )
>>> <div class=... />
```

# The forgotten child



## Introduction to R

Master the basics of data analysis by manipulating common data structures such as vectors, matrices and data frames.

🕒 4 hours



**JONATHAN CORNELISSEN**

Co-founder and CEO of  
DataCamp

```
first_div = course_divs[0]
children = first_div.xpath('./*')
```

```
third_child = children[2]
print( third_child.extract() )
>>> <span class=... />
```

# Listful

- In one CSS Locator

```
links = response.css('div.course-block > a::attr(href)').extract()
```

- Stepwise

```
# step 1: course blocks
course_divs = response.css('div.course-block')

# step 2: hyperlink elements
hrefs = course_divs.xpath('./a/@href')

# step 3: extract the links
links = hrefs.extract()
```

# Get Schooled

```
for l in links:  
    print( l )
```

```
>>> /courses/free-introduction-to-r
```

```
>>> /courses/data-table-data-manipulation-r-tutorial
```

```
>>> /courses/dplyr-data-manipulation-r-tutorial
```

```
>>> /courses/ggvis-data-visualization-r-tutorial
```

```
>>> /courses/reporting-with-r-markdown
```

```
>>> /courses/intermediate-r
```

```
...
```

# Links Achieved

WEB SCRAPING IN PYTHON