XPath Navigation

WEB SCRAPING IN PYTHON



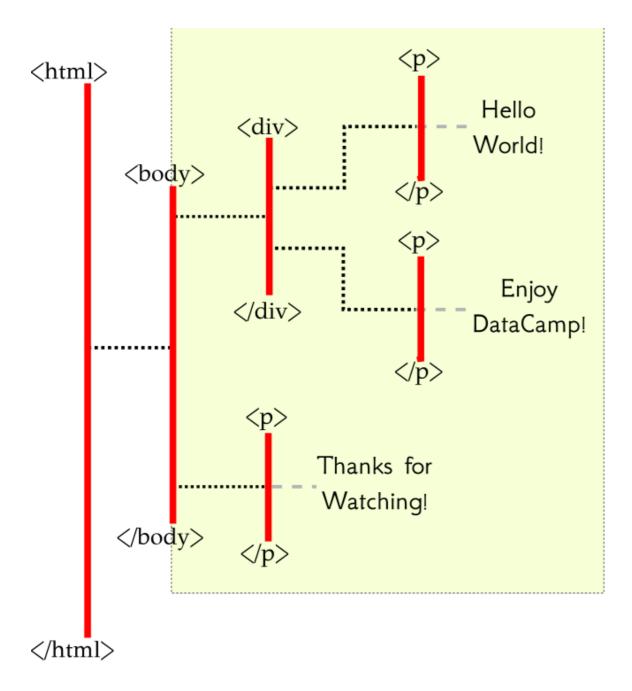
Thomas Laetsch
Data Scientist, NYU



Slashes and Brackets

- Single forward slash / looks forward one generation
- Double forward slash // looks forward all future generations
- Square brackets [] help narrow in on specific elements

To Bracket or not to Bracket



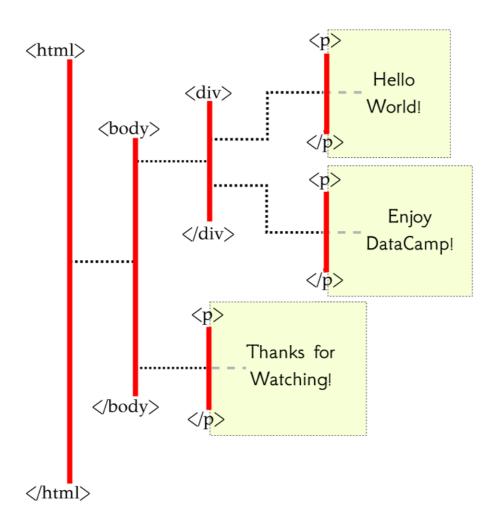
```
xpath = '/html/body'

xpath = '/html[1]/body[1]'
```

Give the same selection

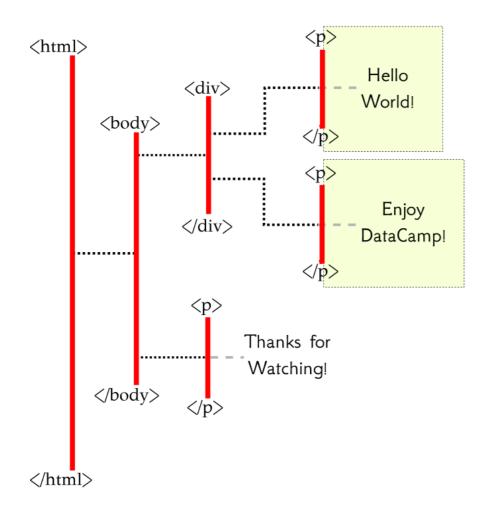
A Body of P

```
xpath = '//p'
```

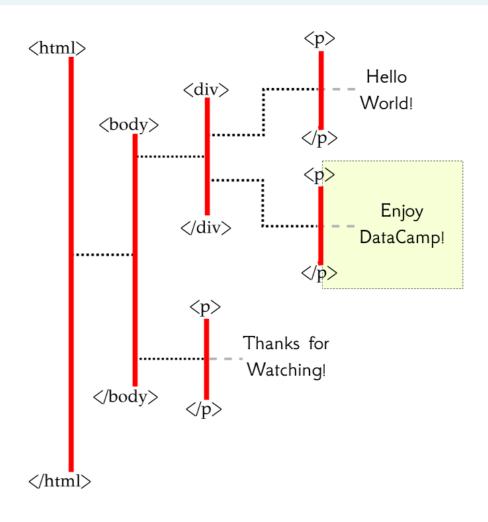


The Birds and the Ps

```
xpath = '/html/body/div/p'
```



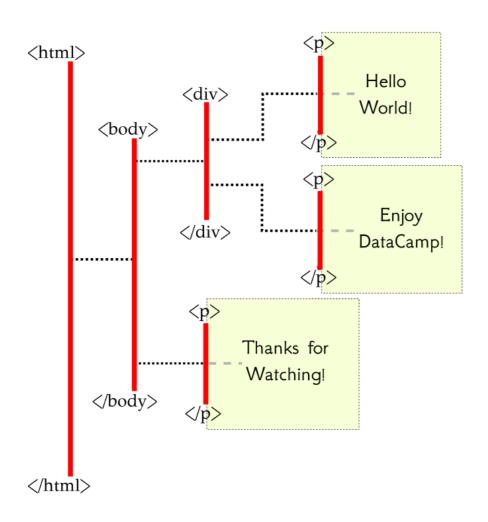
xpath = '/html/body/div/p[2]'

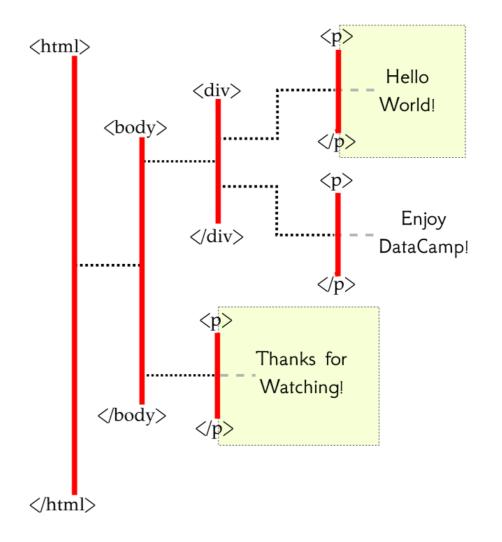


Double Slashing the Brackets

```
xpath = '//p'
```

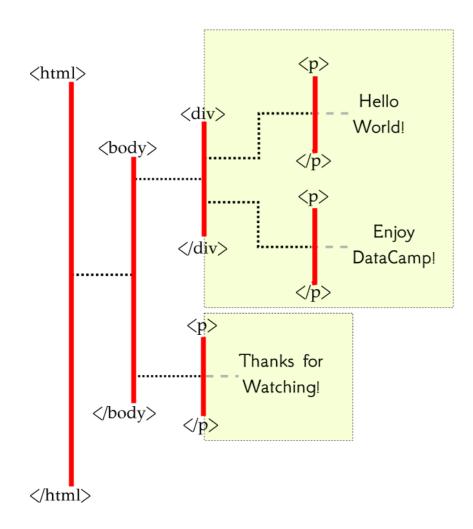






The Wildcard

```
xpath = '/html/body/*'
```



The asterisks * is the "wildcard"

Xposé WEB SCRAPING IN PYTHON



Off the Beaten XPath

WEB SCRAPING IN PYTHON

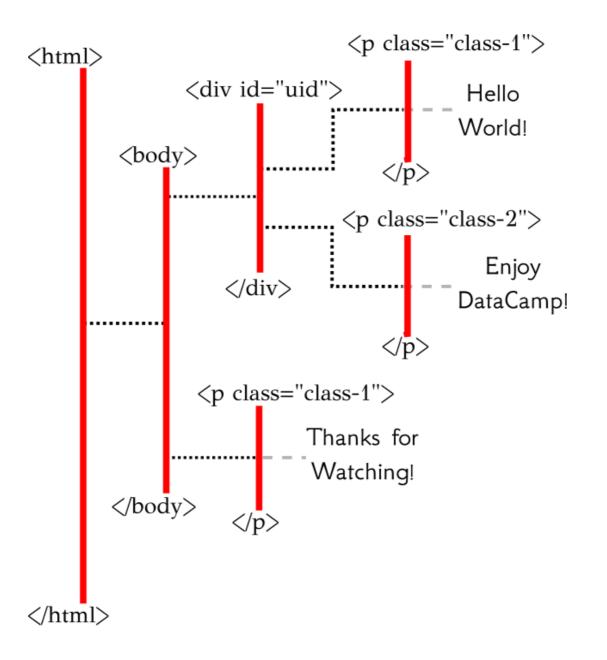


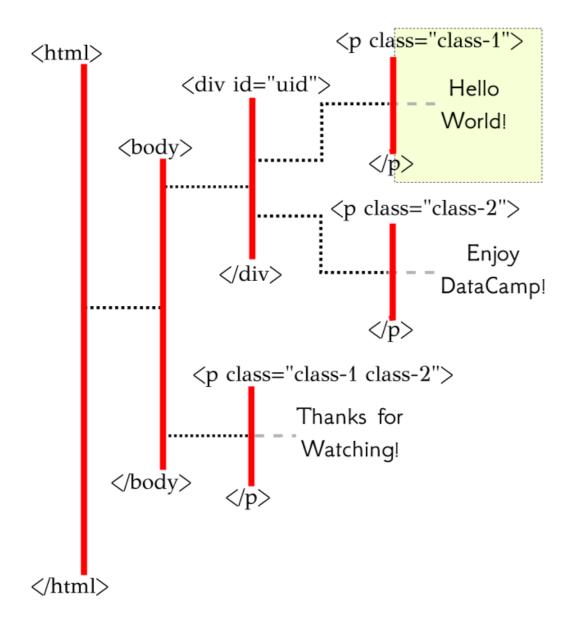
Thomas Laetsch
Data Scientist, NYU



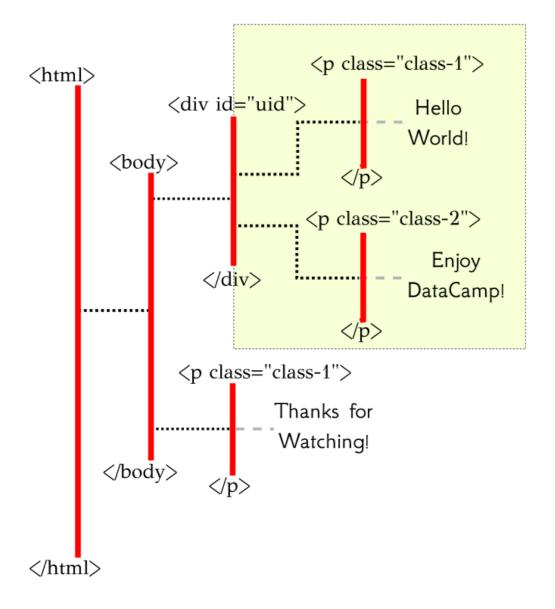
(At)tribute

- @ represents "attribute"
 - o @class
 - o @id
 - o @href

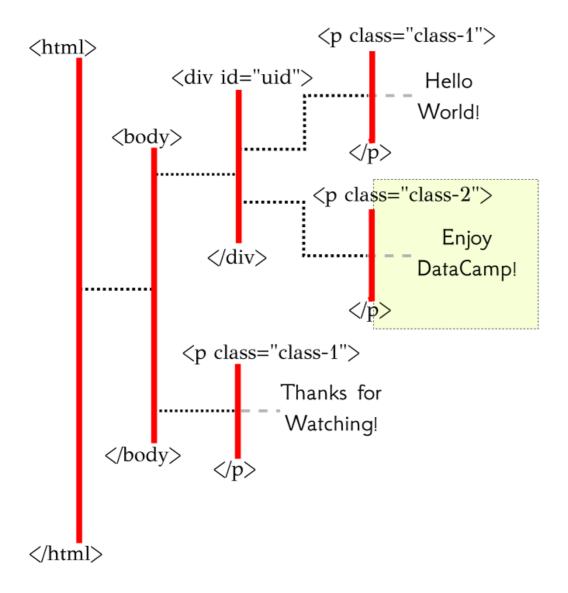




```
xpath = '//p[@class="class-1"]'
```



```
xpath = '//*[@id="uid"]'
```



xpath = '//div[@id="uid"]/p[2]'

Content with Contains

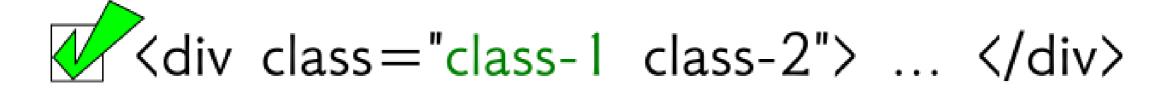
Xpath Contains Notation:

contains(@attri-name, "string-expr")

Contain This

xpath = '//*[contains(@class,"class-1")]'





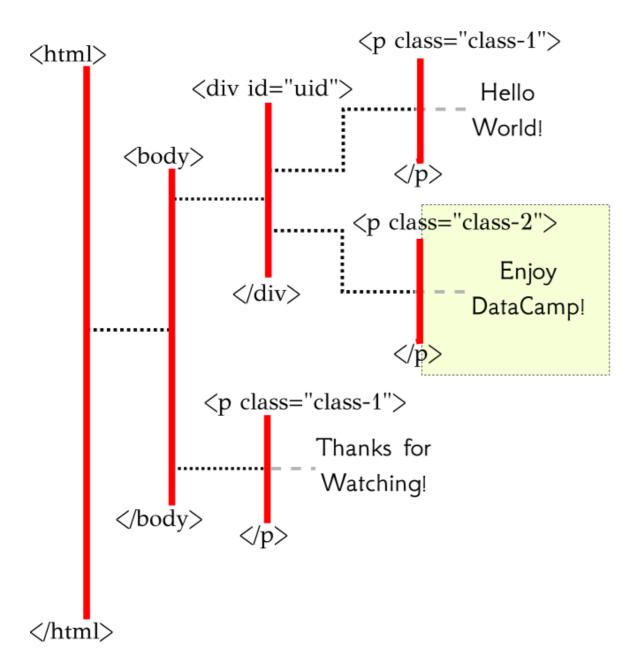
Contain This

```
xpath = '//*[@class="class-1"]'
```



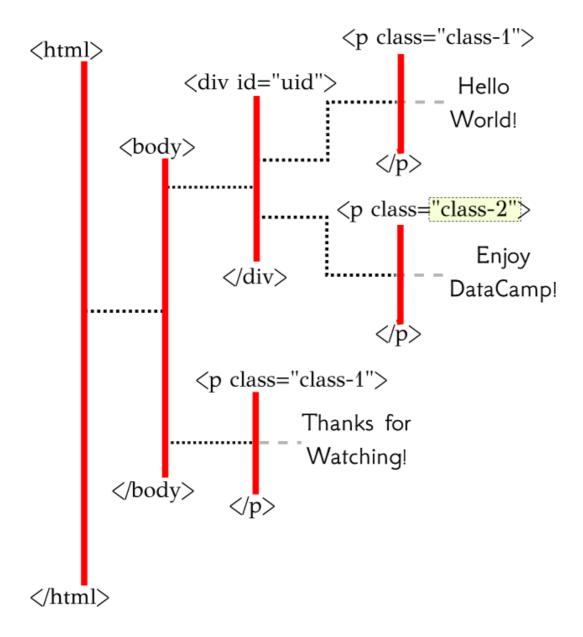


Get Classy



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

Get Classy



xpath = '/html/body/div/p[2]/@class'

End of the Path

WEB SCRAPING IN PYTHON



Introduction to the scrapy Selector

WEB SCRAPING IN PYTHON



Thomas Laetsch
Data Scientist, NYU



Setting up a Selector

```
from scrapy import Selector
html = '''
<html>
  <body>
   <div class="hello datacamp">
     Hello World!
   </div>
   Enjoy DataCamp!
  </body>
</html>
sel = Selector( text = html )
```

- Created a scrapy Selector object using a string with the html code
- The selector sel has selected the entire html document

Selecting Selectors

- We can use the xpath call within a Selector to create new Selector s of specific pieces of the html code
- The return is a SelectorList of Selector objects

Extracting Data from a SelectorList

Use the extract() method

• We can use extract_first() to get the first element of the list

```
>>> sel.xpath("//p").extract_first()
out: 'Hello World!'
```

Extracting Data from a Selector

```
ps = sel.xpath('//p')
second_p = ps[1]

second_p.extract()
out: 'Enjoy DataCamp!'
```

Select This Course!

WEB SCRAPING IN PYTHON



"Inspecting the HTML"

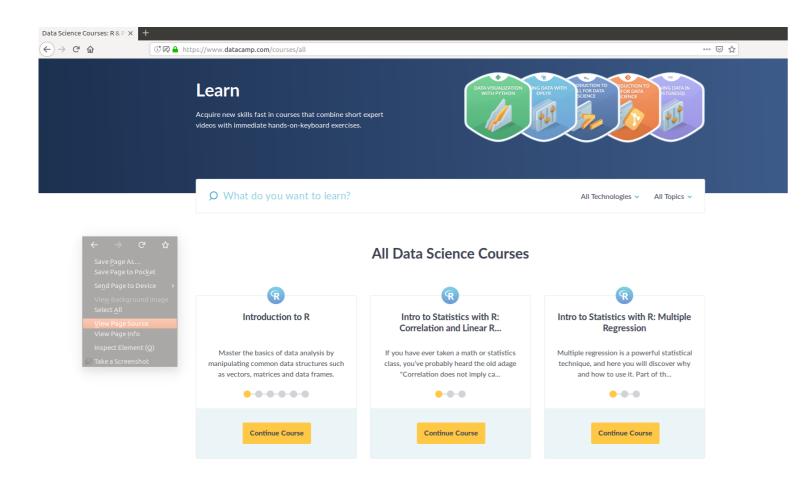
WEB SCRAPING IN PYTHON



Thomas Laetsch, PhD
Data Scientist, NYU

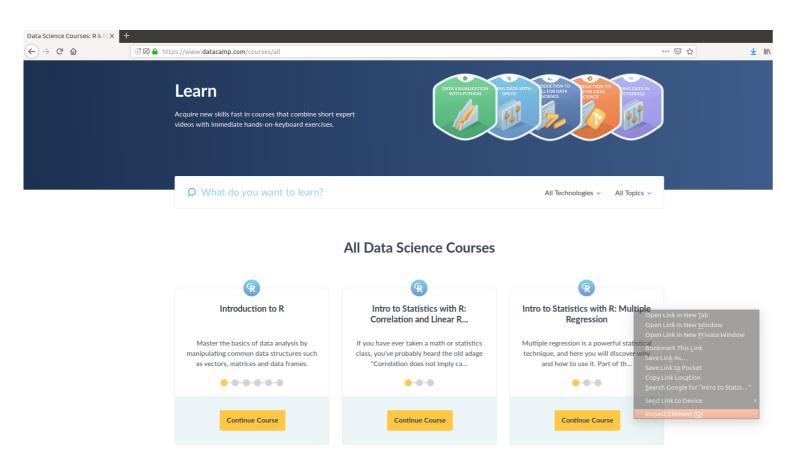


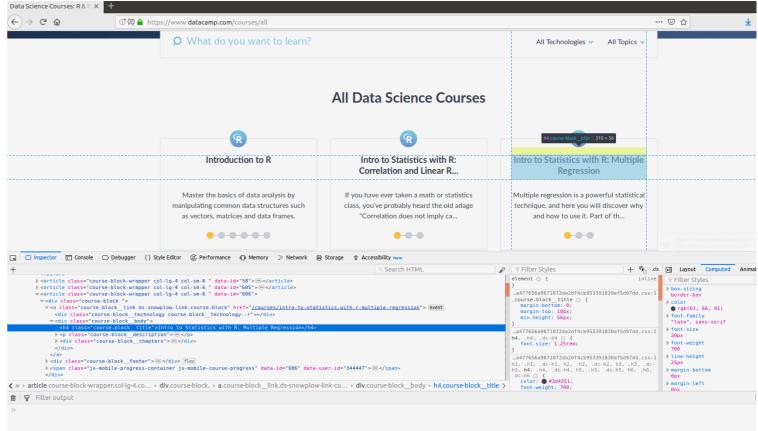
"Source" = HTML Code



```
Data Science Courses: R & P X https://www.datacamp.com X
(-) → C 0
                                                                                                                                                                              ... ☑ ☆
                             i view-source:https://www.datacamp.com/courses/all
    <html class="no-is">
     <meta charset="utf-8">
      cmeta name="google-site-verification" content="ao3s4PdjisD2QsfTbldo7YJx7VX2QLkPEtlDpyFTjo8" />
cmeta name="apple-itunes-app" content="app-id=1263413087">
      < rel="stylesheet" media="all" href="https://cdn.datacamp.com/main-app/assets/application_v2-566f39ae18a65951b6d46fe6a477656a9871872da2df4cb953391830af5d97dd.css" />
   current sylesheet media= att nren= https://com.datat
<script>
(function(h,o,t,j,a,r){
    h,h=h,h][function(){(h,h),q=h,h,q|[]]).push(arguments)};
    h,hjsettings={hjid:484663,hjsv:6};
    a-o.getElement(Syraghae('head')[0];
    r-o.createElement('script');.rasync=1;
    r,src=t-h,hjsettings.hjid+j+h.hjSettings.hjsv;
    a.appendChild(r);
    })(window,document,'https://static.hotjar.com/c/hotjar-','.js?sv=');
    <div class="site-wrap">
  <div class="dc-flash-wrapper" id="flash messages">
   </div>
```

Inspecting Elements





HTML text to Selector

```
import requests
url = 'https://www.datacamp.com/courses/all'
html = requests.get( url ).content

sel = Selector( text = html )
```

You Know Our Secrets

WEB SCRAPING IN PYTHON

