UNIVERSITY of WASHINGTON

Data Science UW Methods for Data Analysis

More on Hypothesis Testing, The Central Limit Theorem,
And an introduction to Regression
Lecture 4

Nick McClure

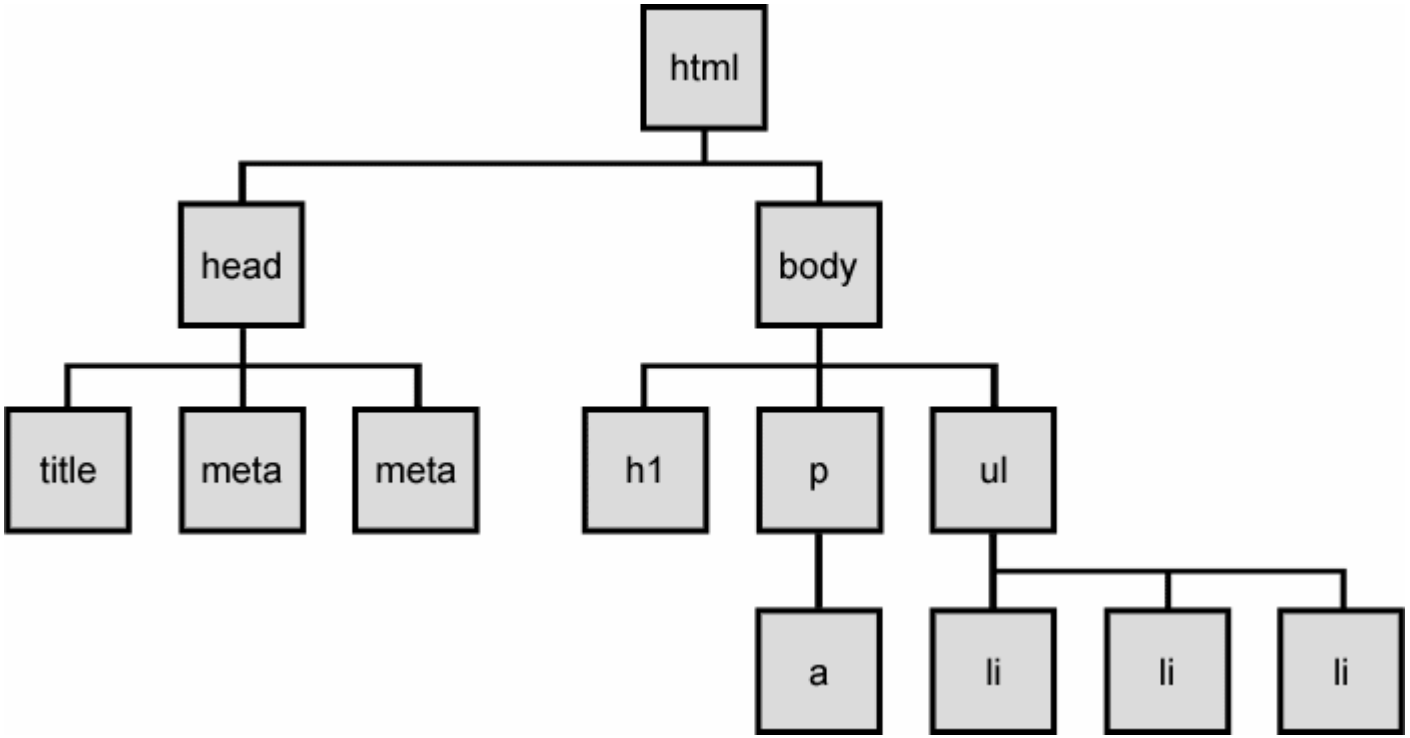
Topics

- X-Path and HTML
- Getting Python Started
- Urllib, Requests, BeautifulSoup

Note: Always check out the server's /robots.txt for what is allowed for scraping.

Xpath

> Xpath is a way to navigate XML documents.
> All XML documents are basically a large tree.



XML vs. HTML

- HTML is a kind of XML focused on making text presentable in a browser.
- You can navigate HTML with similar commands as navigating XML.

Xpath

```
<html>
   <head>
      <title>A List of Beers</title>
   </head>
   <body>
      <h1 class="beer">The Abyss</h1>
      <p style="text-align:right">Deschutes Brewing</p>
      <div class="introduction">
         <p>Imperial Stout.</p>
      </div>

      <h2 class="ABV">12.20%e</h2>
      <p>Slightly high on the ABV side.</p>

      <h2 class="Beer Advocate Rating">99%</h2>
   </body>
</html>
```

h1	= selecting all nodes with name ‘h1’
/body/h1	= selecting from root node
//h1	= selects all nodes from current node that match ‘h1’
.	= select current node
..	= select parent node
@class	=select attributes/filter by attributes

Xpath

```
<html>
   <head>
      <title>A List of Beers</title>
   </head>
   <body>
      <h1 class="beer">The Abyss</h1>
      <p style="text-align:right">Deschutes Brewing</p>
      <div class="introduction">
         <p>Imperial Stout.</p>
      </div>

      <h2 class="ABV">12.20%e</h2>
      <p>Slightly high on the ABV side.</p>

      <h2 class="Beer Advocate Rating">99%</h2>
   </body>
</html>
```

body = selecting all nodes with name ‘h1’

/body/h1[1] = selects the first h1 node

//h1[@class=“beer”] = selects an h1 node where class=“beer”

Python Libraries

- ‘request’ is a library that retrieves HTML requests and has some nice functions to retrieve information from the trees.
- ‘urllib2’ is a Xpath navigating library.
- ‘BeautifulSoup4’ is a library that makes parsing XML slightly easier
  - Html_soup=BeautifulSoup(document, ‘html.parser’)
  - Html_soup.title
  - Html_soup.title.name
  - Html_soup.find_all(‘beer’)