Cheatography

Beautiful Soup Cheat Sheet

by Justin1209 (Justin1209) via cheatography.com/101982/cs/21428/

Import Resources

```
import requests
from bs4 import BeautifulSoup
```

Make a soup object out of a website

```
// 1. The HTTP request
webpage = request.get('URL', 'html.parser');
// 2. Turn the website into a soup object
soup = BeautifulSoup(webpage.content);
```

"html.parser" is one option for parsers we could use. There are other options, like "lxml" and "html5lib" that have different advantages and disadvantages.

Object Types

```
//1. Tags correspond to HTML tags
Example Code:
soup = BeautifulSoup('<div id="example">An example div</div><p>An example p tag</p>');
print(soup.div);
--> <div id="example">An example div</div>
--> gets the first tag of that type on the page
print(soup.div.name)
print(soup.div.attrs)
--> div
--> {'id': 'example'}
//2. Navigable Strings: Piece of text inside of HTML Tags
print(soup.div.string)
--> An example div
```

Navigating by Tags

```
Example Code:
<h1>World's Best Chocolate Chip Cookies</h1>
<div class="banner">
  <h1>Ingredients</h1>
</div>
<ul>
  <li> 1 cup flour </li>
  <li> 1/2 cup sugar </li>
  <li> 2 tbsp oil </li>
  <li> 1/2 tsp baking soda </li>
  <li> ½ cup chocolate chips </li>
  <li> 1/2 tsp vanilla <li>
  <li> 2 tbsp milk </li>
```

Navigating by Tags (cont)

```
</ul>
//1. Get the children of a tag:
for child in soup.ul.children:
    print(child)
--> <li> 1 cup flour </li>
--> <li> 1/2 cup sugar </li>
...
//2. Get the parent of a tag:
for parent in soup.li.parents:
    print(parent)
```

Find All

```
//1. find_all()
print(soup.find_all("h1"))
--> Outputs all <h1>...</h1> on the website
//1.1. find_all() with regex
import re
soup.find_all(re.compile("[ou]l"))
--> Outputs all <ul>...</ul> and <ol>...</ol>
soup.find_all(re.compile("h[1-9]"))
--> Outputs all headings
//1.2. find_all() with lists
soup.find_all(['h1', 'a', 'p'])
//1.3 find_all() with attributes
soup.find_all(attrs={'class':'banner', 'id':'jumbotron'});
//1.4 find_all() with functions
def has_banner_class_and_hello_world(tag):
    return tag.attr('class') == "banner" and tag.string == "Hello world"
soup.find_all(has_banner_class_and_hello_world)
```

CSS Selectors

```
//1. grab CSS classes with .select("class_name")
soup.select(".recipeLink")
//*2. grab CSS IDs with .select("#id_name")
soup.select("#selected")
//3. using a loop
for link in soup.select(".recipeLink > a"):
  webpage = requests.get(link)
  new_soup = BeautifulSoup(webpage)
```