Beautiful Soup Cheat Sheet

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

Import Resources

```
import requests
from bs4 import Beauti fulSoup
```

Make a soup object out of a website

```
// 1. The HTTP request
webpage = reques t.get(URL', 'html.p arser');
// 2. Turn the website into a soup object
soup = Beauti ful Sou p(w ebpage.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 = Beauti ful Sou p(' <div id=" exa mpl e">An
example div</d iv> <p>An example p tag</p >');
print( soup.div);
--> <div id=" exa mpl e">An example div</d iv>
--> gets the first tag of that type on the page
print( sou p.di v.name)
print( sou p.di v.attrs)
--> div
--> {'id': 'example'}
//2. Navigable Strings: Piece of text inside of
HTML Tags
print( sou p.di v.string)
--> An example div
```

Navigating by Tags

```
Example Code:
<h1 >Wo rld's Best Chocolate Chip Cookie s</ h1>
<div class= " ban ner ">
    <h1 >In gre die nts </h 1>
</d iv>
<ul>
    <li> 1 cup flour </l i>
    <li> 1/2 cup sugar </l i>
    <li> 2 tbsp oil </l i>
    <li> 1/2 tsp baking soda </l i>
    <li> ½ cup chocolate chips </l i>
    <li> 1/2 tsp vanilla <li>
    <li> 2 tbsp milk </l i>
```

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.f ind _al l(re.com pil e("[ ou] )l")
--> Outputs all <ul >...</ ul> and <ol >...</ ol>
soup.f ind _al l(re.com pil e("h [1- 9]")
--> Outputs all headings
//1.2. find_all() with lists
soup.f ind _al l([ 'h1', 'a', 'p'])
//1.3 find_all() with attributes
soup.f ind _al l(attrs= {'c las s': 'ba nner',
'id':' jum bot ron'});
//1.4 find_all() with functions
def has_ba nne r_c las s_a nd_ hel lo_tag():
        retu rn tag.at tr( 'cl ass') == " ban -
ner " and tag.string == " Hello world"
soup.f ind _al l(h as_ ban ner _cl ass _an d_h -
ell o_w orld)
```

CSS Selectors

```
//1. grab CSS classes with .select("class_name")
soup.select(".r eci peL ink ")
//*2. grab CSS IDs with .selec t("# id_ nam e")
soup.select("#s ele cte d")
//3. using a loop
for link in soup.s ele ct( ".re cip eLink > a"):
    webpage = reques ts.g et (link)
    new _soup = Beauti ful Sou p(w ebpage)
```