>DevByExample_

A Python Web Scraping How-To Guide

Web Scraping with Python/BeautifulSoup/Requests

Install

```
$ pip install requests beautifulsoup4
```

BeautifulSoup on Text

```
from bs4 import BeautifulSoup

text = '''<div><h1>My Header</h1></div>'''

soup = BeautifulSoup(text, 'html.parser')
print(soup.prettify())
```

```
<div>
 <h1>
  My Header
 </h1>
</div>
```

Fetch Webpage and Create Soup

```
import requests
from bs4 import BeautifulSoup

url = 'https://devbyexample.com/test-scraping'
r = requests.get(url)

soup = BeautifulSoup(r.text, 'html.parser')
```

Find By ID

```
<h1 id="article-title">Hello Everyone</h1>
```

```
header = soup.find(id="article-id")
print(header)
```

```
<h1 id="article-title">Hello Everyone</h1>
```

```
print(header.string)
```

```
Hello Everyone
```

Find By Class

```
<div id="articles">
    <div class='article'>...</div>
    <div class='article'>...</div>
    <div class='article'>...</div>
    <div class='article'>...</div>
    <div class='end'><button>Next Page</button></div>
</div>
```

```
articles = soup.select('.article')
print(articles)
```

```
[   <div class="article">...</div>,
    <div class="article">...</div>,
    <div class="article">...</div>,
    <div class="article">...</div>]
```

Navigating Elements in Tree

```
<ul>
    <li><a href="https://google.com">Google</a></li>
    <li><a href="https://bing.com">Bing</a></li>
    <li><a href="https://apple.com">Apple</a></li>
</ul>
```

```
# Get First Link
print(soup.a)
```

```
<a href="https://google.com">Google</a>
```

```
# Get all Link elements on page
print(soup.find_all("a"))
```

```
[   <a href="https://google.com">Google</a>,
    <a href="https://bing.com">Bing</a>,
    <a href="https://apple.com">Apple</a>]
```

```
# Print all hrefs on page
for link in soup.find_all("a"):
    print(link['href'])
```

```
https://google.com
https://bing.com
https://apple.com
```

Element Attributes

```
<div id="article-10" class="article">
    <h3>Header</h3>
    <p>First Paragraph</p>
    <p>Second Paragraph</p>
</div>
```

```
print(soup.div.name)
```

```
div
```

```
print(soup.div.contents)
```

```
[   '\n',
    <h3>Header</h3>,
    '\n',
    <p>First Paragraph</p>,
    '\n',
    <p>Second Paragraph</p>,
    '\n']
```

```
for strings in div.strings:
    print(repr(strings))
```

```
'\n'
'Header'
'\n'
'First Paragraph'
'\n'
'Second Paragraph'
'\n'
```

```
for strings in soup.div.stripped_strings:
    print(repr(strings))
```

```
'Header'
'First Paragraph'
'Second Paragraph'
```

Find By Regex

```
<div>
    <head><title>Sample Title</title></head>
    <h1>Title Header</h1>
    <hr>
    <div>A description of something</div>
    <h2>Section Header</h2>
    <p>...</p>
    <h2>Another Header</h2>
    <p>...</p>
</div>
```

```
import re

headers = soup.find_all(re.compile('^h[1-6]'))
print(headers)
```

```
[   <h1>Title Header</h1>,
    <h2>Section Header</h2>,
    <h2>Another Header</h2>]
```

Search with CSS Select

```
<div>
    <h3><a href="/sites">Sites</a></h3>
    <ul class="site-list">
        <li><a href="https://google.com">Google</a></li>
        <li><a href="https://bing.com">Bing</a></li>
        <li><a href="https://apple.com">Apple</a></li>
    </ul>
</div>
```

```
print(soup.select('div a'))
```

```
[   <a href="/sites">Sites</a>,
    <a href="https://google.com">Google</a>,
    <a href="https://bing.com">Bing</a>,
    <a href="https://apple.com">Apple</a>]
```

```
print(soup.select('div > h3 > a'))
```

```
[<a href="/sites">Sites</a>]
```

```
print(soup.select('li:nth-child(odd)'))
```

```
[   <li><a href="https://google.com">Google</a></li>,
    <li><a href="https://apple.com">Apple</a></li>]
```

```
print(soup.select('a[href*="http"]'))
```

```
[   <a href="https://google.com">Google</a>,
    <a href="https://bing.com">Bing</a>,
    <a href="https://apple.com">Apple</a>]
```

Parent, Children and Siblings

```
<div>
    <ul>
        <li><a href="https://google.com">Google</a></li>
        <li><a href="https://bing.com">Bing</a></li>
        <li><a href="https://apple.com">Apple</a></li>
    </ul>
</div>
```

```
# Get Parent Name
ul_element = soup.find('ul')
print(ul_element.parent.name)
```

```
div
```

```
# Print all text in children
for child in ul_element.children:
    print(child.string)
```

```
Google
Bing
Apple
```

```
# Siblings
first_li_element = soup.find('li')
print(first_li_element)
for sibling in first_li_element.next_siblings:
    print(sibling)
```

```
<li><a href="https://google.com">Google</a></li>
<li><a href="https://bing.com">Bing</a></li>
<li><a href="https://apple.com">Apple</a></li>
```