Translating Beautiful Soup to C++

Diksha

ID: 2000072

June 20, 2019

Parsing HTML: html.parser

```
<div id="remository">20</div>
```

```
class LinksParser(HTMLParser.HTMLParser):
  def __init__(self):
    HTMLParser.HTMLParser.__init__(self)
    self.recording = 0
    self.data = []

  def handle_starttag(self, tag, attributes):
    if tag != 'div':
      return
    if self.recording:
      self.recording += 1
      return
    for name, value in attributes:
      if name == 'id' and value == 'remository':
        break
    else:
      return
    self.recording = 1
```

```
def handle_endtag(self, tag):
  if tag == 'div' and self.recording:
    self.recording -= 1

def handle_data(self, data):
  if self.recording:
    self.data.append(data)
```

What is Beautiful Soup?

What is Beautiful Soup?

- Python's library

What is Beautiful Soup?

- Python's library
- Extracts data from HTML and XML

What is Beautiful Soup?

- Python's library
- Extracts data from HTML and XML
- Easy navigating through the parse tree

https://code.launchpad.net/beautifulsoup

Parsing HTML: Beautiful Soup

```
from bs4 import BeautifulSoup
soup = BeautifulSoup('<div id="remository">20</div>')
tag=soup.div
print(tag.string)
```

Why care to convert?

- Good learning exercise

Why care to convert?

- Good learning exercise
  - Learn working with open source

Why care to convert?

- Good learning exercise
  - Learn working with open source
  - Learn to read code

Why care to convert?

- Good learning exercise
  - Learn working with open source
  - Learn to read code
  - Learn internal details of BS4

My Plan

My Plan

- Read documentation

My Plan

- Read documentation
- Read codebase

My Plan

- Read documentation
- Read codebase
- Convert

My Plan

- Read documentation
- Read codebase
- Convert
- Package into library

Current Progress

- About 50% documentation read

Discussion