# Translating Beautiful Soup to C++

Diksha

TalentSprint WE

June 29, 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)
```

# Parsing HTML: Beautiful Soup

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

# About Beautiful Soup

# About Beautiful Soup

- A Python library

# About Beautiful Soup

- A Python library
- Extracts data from HTML and XML

# About Beautiful Soup

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

# About Beautiful Soup

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

Code available at:
https://code.launchpad.net/beautifulsoup (5000 lines)

# Motivation

# Motivation

- Past experience

# Motivation

- Past experience
- Good learning exercise

# Motivation

- Past experience
- Good learning exercise
  - Learn working with open source

# Motivation

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

# Motivation

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

# Challenges

# Challenges

- Read BS4 code

# Challenges

- Read BS4 code
- No good HTML parsers in C++

# Challenges

- Read BS4 code
- No good HTML parsers in C++
  - Using htmlcxx as of now

# Challenges

- Read BS4 code
- No good HTML parsers in C++
  - Using htmlcxx as of now
  - No proper documentation

# Challenges

- Read BS4 code
- No good HTML parsers in C++
  - Using htmlcxx as of now
  - No proper documentation
- Design decisions

# My Plan

# My Plan

- Read documentation - completed

# My Plan

- Read documentation - completed
- Read code base - completed

# My Plan

- Read documentation - completed
- Read code base - completed
- Convert - completed

# My Plan

- Read documentation - completed
- Read code base - completed
- Convert - completed
- Package into library

# Learning

# Learning

- Working with Bazaar

# Learning

- Working with Bazaar
- Read and understood code

# Learning

- Working with Bazaar
- Read and understood code
- Fixed some documentation (merged)

# Learning

- Working with Bazaar
- Read and understood code
- Fixed some documentation (merged)
- Building docs with Sphinx

# Learning

- Working with Bazaar
- Read and understood code
- Fixed some documentation (merged)
- Building docs with Sphinx
- Tree.hh API and htmlcxx parser

# Discussion