Introduction to Beautiful Soup

- João Vieira
- Manuel Monteiro
- Miguel Dias

What is it?

Beautiful soup is a *Python* library used to extract data from XML/HTML documents. It is capable of creating a parsing tree to provide idiomatic ways of navigating, searching and modifying it . It is most used for **web scraping**.

Installation

Simply open your command prompt and execute:

```
$ pip install beautifulsoup4
```

Utilization - Making a soup

First, we convert a HTML String to Beautiful Soup format

Exemple:

```
1.  <html>
2.  <head><title>Beautiful Soup</title></head>
3.  <body>
4.  <h1 class=”Teste”>Hello World!</h1>
5.  </body>
6.  </html>

soup = BeautifulSoup(html_doc, “html_parser”)
```

Utilization - Extracting

To extract info:

```
1.  soup.title.text
2.  soup.body.h1.text
3.  soup.body.h1['class']
```

Results:

1. “Beautiful Soup”
2. “HelloWorld!”
3. “Teste”

Introduction to Beautiful Soup

- João Vieira
- Manuel Monteiro
- Miguel Dias