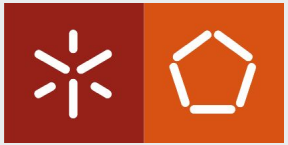
Processamento de Linguagem Natural

Luís Filipe Cunha
lfc@di.uminho.pt

José João Almeida
jj@di.uminho.pt



Introduction to Beautiful Soup

- A Python library used for web scraping and parsing HTML and XML documents.
- Powerful parsing capabilities, easy-to-use syntax, compatibility with different parsers, etc.

Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.

Get Started

```
pip install beautifulsoup4
```

```
from bs4 import BeautifulSoup
html_content = '<html> … </html>'
soup = BeautifulSoup(html_content, 'html.parser')
```

Quick Start

```
<html>
  <head>
    <title>Example Website</title>
  </head>
  <body>
    <h1>Welcome to Beautiful Soup!</h1>
    <p> This is an example HTML document.</p>
    <ul>
      <li>Item 1</li>
      <li class='special'>Item 2</li>
    </ul>
    <a href='https://www.atlasdasaude.pt'> Atlas da Saúde
</a>
  </body>
</html>
'''
```

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(html_content,
'html.parser')

soup.title
#<title>Example Website</title>

soup.title.text
#'Example Website'

soup.a["href"]
#'https://www.atlasdasaude.pt'

soup.find_all("li")
#[<li>Item 1</li>, <li class="special">Item
2</li>]

soup.find_all("li", class_="special")
#[<li class="special">Item 2</li>]
```

Requests

```
pip install requests
```

```
import requests
from bs4 import BeautifulSoup

html_doc = requests.get("https://www.atlasdasaude.pt/doencasaaz/a")

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

Processamento de Linguagem Natural

Luís Filipe Cunha
lfc@di.uminho.pt

José João Almeida
jj@di.uminho.pt