

Web Crawling

Beautifulsoup

BeautifulSoup

BeautifulSoup

- What is BeautifulSoup?
 - A Python library for parsing HTML and XML documents.
 - Widely used for web scraping.
- Why Use BeautifulSoup?
 - Simple to learn and implement.
 - Powerful methods for navigating and searching the document tree.
 - Works well with Python's other HTTP libraries to access websites.
- Installation

```
pip install beautifulsoup4
```

```
pip install lxml # optional, recommended parser
```

Basic Example

```
from bs4 import BeautifulSoup

soup = BeautifulSoup("<p>Some<b>bold</b>text.</p>", "html.parser")

print(soup.p.b.string)  # Output: 'bold'
```

Result

```
from bs4 import BeautifulSoup
soup = BeautifulSoup("<p>Some<b>Hello</b>text.</p>", "html.parser")
print(soup.p.b.string) # Output: 'Hello'
```

Hello

Common Uses

- Extracting data from HTML.
- Automating data collection from web sources.
- Cleaning up messy web page HTML.

Example

- To parse HTML using BeautifulSoup, you generally follow these steps:
 - Load the HTML of the webpage.
 - Create a BeautifulSoup object to parse the HTML.
 - Search and extract tags, attributes, text, etc.

```
from bs4 import BeautifulSoup
html_doc = """
<html>
<head>
    <title>The Dormouse's story</title>
</head>
<body>
    <p class="title"><b>The Dormouse's story</b></p>
    <p class="story">Once upon a time there were three little sisters; and their names were
    <a href="http://example.com/elsie" class="sister" id="link1">Elsie</a>,
    <a href="http://example.com/lacie" class="sister" id="link2">Lacie</a> and
    <a href="http://example.com/tillie" class="sister" id="link3">Tillie</a>;
    and they lived at the bottom of a well.</p>
</body>
</html>
"""
```


1. BS object and 2. search and extract a tag

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

```
# Example of accessing an HTML tag
```

```
print(soup.title) # Outputs the <title> tag
```

```
print(soup.head) # Outputs the <head> tag
```

```
print(soup.a) # Outputs the first <a> tag
```

Accessing Tag Contents and Attributes

```
# Accessing tag content
print(soup.title.string)

# Accessing tag attributes
link = soup.a
print(link['href']) # Outputs the href attribute value
```

For more complex conditions when searching for tags, use the `find` or `find_all` methods:

```
# Finding all <a> tags
all_links = soup.find_all('a')
for link in all_links:
    print(link['href'])

# Finding all <a> tags with class "sister"
sisters = soup.find_all('a', class_='sister')
for sister in sisters:
    print(sister.string)
```

Results

```
# Finding all <a> tags
all_links = soup.find_all('a')
for link in all_links:
    print(link['href'])

# Finding all <a> tags with class "sister"
sisters = soup.find_all('a', class_='sister')
for sister in sisters:
    print(sister.string)
```

<http://example.com/elsie>
<http://example.com/lacie>
<http://example.com/tillie>
Elsie
Lacie
Tillie

Example

CS Department Homepage

To practice how to collect data from a real webpage, we'll see how you could write a Python script using BeautifulSoup and the Requests library to scrape data from the specified website, <https://www.utrgv.edu/csci/faculty/index.htm>.

The goal is to collect professor's names.

Step1: Install Required Libraries

```
pip install beautifulsoup4
```

```
pip install requests
```

Step 2: Fetch the Web Page

```
import requests

from bs4 import BeautifulSoup

# URL of the page
url = 'https://www.utrgv.edu/csci/faculty/index.htm'

# Send HTTP request
response = requests.get(url)

# Check if the request was successful
if response.status_code == 200:
    print("Web page fetched successfully!")
else:
    print("Failed to retrieve the web page. Status code:", response.status_code)
```


Step 3: Parse the HTML Content

```
# Parse the HTML content of the page  
soup = BeautifulSoup(response.content, 'html.parser')
```

Search for Position: Search



Emmett Tomai
Professor
Department Chair
2015 UT System Regents'
Outstanding Teaching Award
EIEAB 3.213
956-665-3520
emmett.tomai@utrgv.edu



Zhixiang Chen
Professor
Graduate Program Coordinator
EIEAB 3.241
956-665-2857
zhixiang.chen@utrgv.edu



Andres Figueroa
Professor
Undergraduate Program
Coordinator
EIEAB 3.247
956-665-8744
andres.figueroa@utrgv.edu



Michael Aguilon
Adjunct Lecturer
EIEAB 3.227
michael.aguilon01@utrgv.edu



Marzieh Ayati
Assistant Professor
EIEAB 3.217
956-665-7302
marzieh.ayati@utrgv.edu



Divya Bajaj
Lecturer I
EIEAB 3.228
divya.bajaj@utrgv.edu



Sonya Cirlos
Adjunct Lecturer
sonya.cirlos01@utrgv.edu



Gustavo Dietrich
Senior Lecturer
2013 UT System Regents'
Outstanding Teaching Award
EIEAB 3.245
956-665-2618
gustavo.dietrich@utrgv.edu



Erik Enriquez



Pedro Fonseca



Bin Fu



Yifeng Gao

Browser developer tools showing HTML and CSS. The HTML pane shows the structure of the page with various accessibility-related scripts. The CSS pane shows styles for the body, including font settings and display properties. A diagram in the CSS pane illustrates a box model with a 1180x1921 content area and padding.

... class="page title faculty & staff" ...

- ▶ <div class="department clear" id="showDirectory"> ... </div>
- <hr>
- ▶ <style type="text/css"> ... </style>
- ▼ <div class="images">
- ▼ <div class="listing">
-
- ▼ <p>
- ▼
- Emmett Tomai == \$0
-
-

- " Professor "
-

- " Department Chair "
-

- ▶ <a aria-label="2015 UT System Regents' Outstanding Teaching Award"

Step 5: Extraing tag

```
# Find elements containing professor names
listings = soup.find_all('div',{'class': "listing"})
for l in listings:
    strong = l.find("strong")
    print(strong.string)
```

Emmett Tomai
Zhixiang Chen
Andres Figueroa
Michael Aguilon
Marzieh Ayati
Divya Bajaj
Sonya Cirlos
Gustavo Dietrich
Erik Enriquez
Pedro Fonseca
Bin Fu
Yifeng Gao
Joselito Guardado
Roberto Jimenez
Dong-Chul Kim
Qi Lu
Eric Martinez


Lab 38

Using BeautifulSoup, collect faculty emails from the CS homepage, UTRGV.

<https://www.utrgv.edu/csci/faculty/index.htm>

Results

```
# Find elements containing professor emails
listings = soup.find_all('div',{'class': "listing"})
for l in listings:
```



emmett.tomai@utrgv.edu
zhixiang.chen@utrgv.edu
andres.figueroa@utrgv.edu
michael.aguillon01@utrgv.edu
marzieh.ayati@utrgv.edu
divya.bajaj@utrgv.edu
sonya.cirlos01@utrgv.edu
gustavo.dietrich@utrgv.edu
erik.enriquez01@utrgv.edu
pedro.fonseca01@utrgv.edu
bin.fu@utrgv.edu
yifeng.gao@utrgv.edu
josecito.guardado01@utrgv.edu
roberto.jimenez01@utrgv.edu
dongchul.kim@utrgv.edu
qi.lu@utrgv.edu
eric.m.martinez02@utrgv.edu
askar.nurbekov01@utrgv.edu
carlos.penacaballero01@utrgv.edu
alfredo.ramos02@utrgv.edu
robert.schweller@utrgv.edu
haoteng.tang@utrgv.edu
charlie.ticer01@utrgv.edu
david.torres@utrgv.edu
timothy.wylie@utrgv.edu
li.zhang@utrgv.edu
odette.perez@utrgv.edu