WEB SCRAPING CON PYTHON

LEONARDO TAMIANO

Created: 2023-12-31 Sun 18:38

TABLE OF CONTENTS

- What is Web Scraping?
- Beautiful Soup
- Real (Life) Example

WHAT IS WEB SCRAPING?

Per "web scraping" si intende l'utilizzo di una serie di tecnologie al fine di estrarre dati dal web (tipicamente da pagine HTML) in modo da poterli poi processare come si vuole.

Un tipico esempio di web scraping consiste nell'estrarre dei dati di interesse da una pagina web per poi metterli in un'altra pagina web, andando però a cambiare lo stile utilizzato per mostrare i dati.

DOCUMENT OBJECT MODEL

I file scritti in **HTML** possono essere rappresentati tramite una struttura dati chiamata **Document Object Model** (DOM). II DOM è una struttura alborea che contiene sia la struttura del documento e sia il contenuto del documento.

```
<!DOCTYPE html>
<head>
  <title> Titolo Pagina </title>
  <meta name="viewport" content="width=device-width, initial-scale=1">
  <meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
</head>
<body>
 <div id="content">
   <h1> Headline level 1 </h1>
    This is a paragraph! 
   <div id="footer">
     <b>Author</b>: Leonardo Tamiano
    حديث الماري
```

Esempio file HTML

```
_DOCTYPE: html
-HTML
 _HEAD
   -#text:
   -TITLE
    #text: Titolo Pagina
   #text:
  -#text:
 BODY
   -#text:
   DIV id="content"
    -#text:
     _H1
      └#text: Headline level 1
     -#text:

    #text: This is a paragraph!

     -#text:
      -DIV id="footer"
      -#text:
      P class="author"
         #text: Author
        #text: : Leonardo Tamiano
      -#text:
     #text:
    #text:
```

Esempio DOM generato con live-dom-viewer

Esempio DOM generato con Graphviz

Tipicamente le librerie utilizzate per fare web scraping funzionano in due passi:

- 1. Si costruiscono il DOM rappresentante il documento da analizzare.
- 2. Offrono una serie di APIs per muoversi all'interno del DOM e raccogliere solamente i dati di nostro interesse.

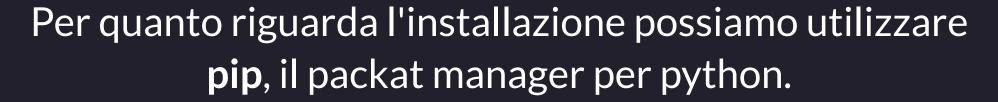
BEAUTIFUL SOUP

Beautiful Soup è una libreria Python che ci permette di fare web scraping.

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.

Source: Beautiful Soup

INSTALLATION



pip install beautifulsoup4

GENERATING THE DOM

Supponiamo di aver salvato il codice html fatto vedere prima nel file **web**pageexample.html. Per generare il DOM possiamo procedere come segue

```
#!/usr/bin/env python3

from bs4 import BeautifulSoup

# -- read file
f = open("./web_page_example.html", "r")
text = f.read()

# -- generate DOM structure
soup = BeautifulSoup(text, 'html.parser')
```

NAVIGATING THE DOM

Una volta che abbiamo generato la struttura DOM la possiamo navigare in vari modi:

Trovare tutti i tags di un certo tipo

```
# -- find all tags of the form  ... 
paragraphs = soup.find_all("p")
```

Trovare tutti i tags con un certo attributo

```
# -- find all tags of the form <div id="footer"> ... </div>
footer_div = soup.find("div", {"id": "footer"})
```

 A partire da un nodo del DOM possiamo ripetere la ricerca per trovare tutti i tags contenuti in quel particolare sotto-albero del DOM.

```
if footer_div:
    # -- find firsts ... inside <div id="footer"> ... </div>
    author_p = footer_div.find("p")
```

 Possiamo anche esplorare il DOM utilizzando la notazione con il punto (.) come segue

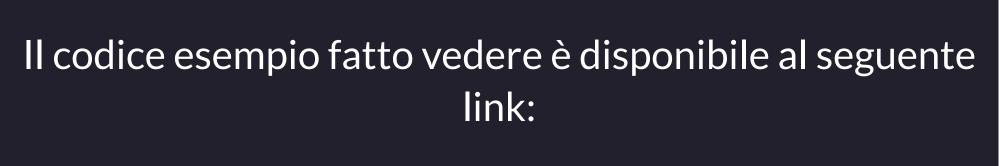
```
author_p = footer_div.p
```

Così facendo però non siamo sicuri se l'elemento a cui stiamo tentando di accedere esiste davvero.

EXTRACTING DATA FROM THE DOM

Una volta che abbiamo i tag di interesse possiamo accedere ai dati veri e propri come segue

```
# -- get all data
print(author_p.decode_contents()) # prints: <b>Author</b>: Leonardo Tamiano
# -- get only text data # prints: Author: Leonardo Tamiano
print(author_p.text)
```



REAL (LIFE) EXAMPLE

Consideriamo la seguente pagina, che mostra gli orari delle lezioni.

9-00			
State Stat			
9.00			
11:00 Comertie et alsgebra Fisica Comertie et alsgebra C	Venerdì		
11-00			
Comparison of a signature of a sig			
Second and Second			
14:00 Ligida er tel logiche Programmazione del calcolatori con laboratorio Logica e reti logiche Programmazione del calcolatori con laboratorio Logica e reti logiche Programmazione del calcolatori con laboratorio Logica e reti logiche Programmazione del calcolatori con laboratorio Logica e reti logiche Programmazione del calcolatori con laboratorio Logica e reti logiche Programmazione del calcolatori con laboratorio Logica e reti logiche Programmazione del calcolatori con laboratorio Logica e reti logiche Programmazione del calcolatori con laboratorio Logica e reti logiche Logica e reti logiche Programmazione del calcolatorio con laboratorio Logica e reti logiche Logica Programmazione Logica Logic			
Caula 13			
Agent Agen	e dei calcolatori con laboratorio		
17:00	e dei calcolatori con laboratorio		
Secondo anno Seco	e dei calcolatori con laboratorio		
Secondo amo			
Secondo anno Lezioni in autia online se non diversamente specificato Sistemi operativi e reti			
9:00 Signature dati Sistemi operativi e reti Fondamenti di informatica Ajporitmi e strutture dati Sistemi operativi e reti Fondamenti di informatica Ajporitmi e strutture dati O			
Algoritmi e strutture dati () () () () () () () () () () () () ()	Vanandi		
10:00 0	Veneral		
11:00 C			
12:00 Clinguaggi e metodologie di programmazione Clinguaggi e metodologie di endologia Clinguaggi e metodologie Clinguaggi e etodologie Clingu			
13:00			
14:00 Basi di dati e di conoscenza Ricerca operativa Radi dati e di conoscenza Ruserca operativa Ruserca operativa Basi di dati e di conoscenza Ruserca operativa Ruserc			
Aula 3PP2 (Aula 3PP2 C)			
Aula 3PP2 ()			
17:00			
Company Comp			
Terzo anno Lezioni in aula online se non diversamente specificato Vene Giovedi Vene			
Company Comp			
Ora Lunedi Martedi Mercoledi Giovedi Vene			
9:00 10:00 [Ingegneria del software (9:30) Frogrammazione Java per dispositivi mobili (10:30) Frogrammazione Web (12:30) Frogrammazione Java per dispositivi mobili (12:30) Frogrammazione Web (12:30) Frogrammazione Web (12:30) Frogrammazione Java per dispositivi mobili (13:30) Frogrammazione Java per dispositivi mobili (13:30) Frogrammazione Jav	Venerdi		
10:00 Ingegneria del software (9:30) Ingegneria del software (9:30) Ingegneria del software (9:30) Ingegneria del software (9:30) Ingegneria del software (10:30) In			
11:00	oftware		
12:00 Modelli e linguaggi di simulazione (11:30) Programmazione Java per dispositivi mobili Modelli e linguaggi di simulazione (12:30) Programmazione Web (12:30) Programmazione Java per dispositivi mobili (12:30) Programmazione Java per dispositivi mobili (13:30) Programma	oftware		
13:00 Modelli e linguaggi di simulazione (12:30) Programmazione Web (10:30) Programmazione Web (10:30) Programmazione Web (10:30) Programmazione Java per dispositivi mobili (10:30) Programmazione Web (10:30) Programmazione Web (10:30) Programmazione Java per dispositivi mobili (10:30) Programmazione Web (10:30) Progr			
14:00 Programmazione Web Programmazione Java per dispositivi mobili () Programmazione Java per dispositivi mobili () Programmazione Java per dispositivi mobili ()			
15:00 Programmazione Web Programmazione Java per dispositivi mobili			
16:00			
17:00			
17:00			
10.00			

Il nostro obiettivo è scaricare il file .html contenente le informazioni degli orari e salvare i dati in un file .csv, in modo poi da poterli processare a nostro piacimento.

