

Programación. Python

Web scraping





Introducción: HTML

```
pequenno codigo html =
<html>
   <head>
      <title>
         A web page
      </title>
   </head>
   <body>
      Joel Grus
      Data Science
   </body>
</html>
```

```
pequenno_codigo_html =
<html>
   <head>
      <title>
         A web page
      </title>
   </head>
   <body>
      Joel Grus
      Data Science
   </body>
</html>
.....
```

Introducción: HTML... y BeautifulSoup

```
from bs4 import BeautifulSoup
soup_pequenno_codigo = BeautifulSoup(pequenno_codigo html, "lxml")
primer parrafo = soup pequenno codigo.(find('p'))
                                                    pequenno codigo html =
                                                   <html>
print(primer parrafo.text)
                                                       <head>
print("-
                                                           <title>
print(primer parrafo.text(split(
print(primer parrafo["id"])
                                                               A web page
todos_los_parrafos = soup_pequenno_codigo.find_all('p'
                                                           </title>
print(todos los parrafos)
                                                       </head>
print(todos los parrafos[0].text)
                                                       <body>
                                                           Joel Grus
Joel Grus
                                                           Data Science
                                                       </body>
 'Joel', 'Grus'
                                                    </html>
aut hor
|Joel Grus, Data Science]
Joel Grus
```

Una página web real y su fuente



Cristóbal Pareja Flores, página web X

cristóbal

pareja flores

docencia

Curso 2013-14

antares.sip.ucm.es/cpareja

Actualizado: 2014-abril-06

Asignaturas impartidas durante el curso 2013-14: Programación Declarativa (Grado en CC. Matemáticas), Programación II y Taller de Algoritmos (Grado en Estadística). Los siguientes enlaces serán útiles a los estudiantes de estas asignaturas.

Campus virtual UCM

Mooshak (manual)

FLOP

Grupo

investigación

Hasta 2014

Actualizado: 2014-dic-23

Áreas e intereses principales: Programación funcional, métodos formales, transformación de programas. Entornos de programación y herramientas de visualización. Testing and performance evaluation, y Cloud computing. Enseñanza de la Informática. Historia de las matemáticas.

He aquí una pequeña selección de publicaciones, sin actualizar. (Google académico y DBLP recogen su propia selección de mis contribuciones.)

- L. Llana Díaz, E. Martín Martín, C. Pareja Flores y J. Á. Velázquez Iturbide, "FLOP: A User-Friendly System for Automated Program Assessment", J.UCS, Journal of Universal Computer Science, 20(9), págs. 1304–1326, octubre 2014. (http://www.jucs.org/jucs_20_9/flop_a_user_friendly)
- M. Rubio, Păivi Kinnunen, C. Pareja and J. Á. Velázquez, "Student perception and

view-source:http://antares.sip.ucm.es/cpareja/

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
 4 <!--
    Design by Free CSS Templates
    http://www.freecsstemplates.org
    Released for free under a Creative Commons Attribution 2.5 License
               : C. Pareia Flores
    Description: Home page
     Version
             : 1.0
  <html xmlns="http://www.w3.org/1999/xhtml">
15 <head>
    <meta name="keywords" content="" />
    <meta name="description" content="" />
    <meta http-equiv="content-type" content="text/html; charset=utf-8" />
    <title>Cristóbal Pareja Flores, página web</title>
    <link href="./style.css" rel="stylesheet" type="text/css" media="screen" />
21 </head>
  <body>
    <div id="wrapper">
      <div id="page">
        <div id="page-bgtop">
          <div id="page-bgbtm">
            <div id="content">
              <div class="post">
                <h2 class="title"><a href="#">Docencia</a></h2>
                <span class="date">Curso 2013-14</span><span class="posted">Actualizado: 2014-abril-06</span>
```

Una página web real y su fuente

```
mi_url = "http://antares.sip.ucm.es/cpareja/"
mi_codigo_html = requests.get(mi_url).text
print(mi_codigo_html[:250])
print(".......")
print(mi_codigo_html[-250:])

<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<!--
Design by Free CSS Templates</pre>
```

ga.src = ('https:' == document.location.protocol ? 'https://ssl' : 'http://www') + '.google-anal

var s = document.getElementsByTagName('script')[0]; s.parentNode.insertBefore(ga, s);

import requests

vtics.com/ga.js';

})();

</script> </body>

</html>

http://www.freecsstemplates.org

Released for free under a Creative Commons Attribution 2.5 L

```
mi pag web = BeautifulSoup(mi_codigo_html, "lxml")
print(mi pag web.find('head'))
print("----")
print(mi pag web.find('p'))
<head>
<meta content="" name="keywords"/>
<meta content="" name="description"/>
<meta content="text/html; charset=utf-8" http-equiv="content-type"/>
<title>Cristóbal Pareja Flores, página web</title>
<link href="./style.css" media="screen" rel="stylesheet" type="text/css"/>
</head>
<span class="date">Curso 2013-14</span><span class="posted">Actualizado: 20
14-abril-06</span>
```

```
S
        mi pag web = Beau
                             print(len(mi_pag_web.find_all('p')))
elemento
        print(mi pag web.
                             print("-----
        print("----
                             print(mi pag web.find all('p')[8])
                             print("-----
        print(mi pag web.
                             print(mi_pag_web.find_all('p', {'class': 'meta'}))
                             print("--
        <head>
                             anclas o hiperenlaces = mi pag web.find all('a')
        <meta content=""</pre>
                             print(len(anclas o hiperenlaces))
                             print(anclas o hiperenlaces[3])
        <meta content=""
                             print(anclas o hiperenlaces[3].get("href"))
        <meta content="te print(anclas_o_hiperenlaces[3].get_text())</pre>
        <title>Cristóbal
        <link href="./stv</pre>
                             S
        </head>
                             <span class="date">2015</span><span class="posted">Actualizado: 2015-abril-6</span>
SU
                             <span class="date [<p class="meta">
                             <span class="date">Curso 2013-14/span><span class="posted">Actualizado: 2014-abril-06</span>
        14-abril-06</span
a
                             , 
        <span class="date">Hasta 2014</span><span class="posted">Actualizado: 2014-dic-23</span>
                             , 
                             <span class="date">2015</span><span class="posted">Actualizado: 2015-abril-6</span>
                             , 
S
                             <span class="date">Hasta 2014/span class="posted">Actualizado: 2014-abril-6
NCCe
                             ]
                             <a href="http://eprints.ucm.es/8705/1/CUADERNO DE TRABAJO 8.pdf">manual</a>
                             http://eprints.ucm.es/8705/1/CUADERNO DE TRABAJO 8.pdf
                            manual
```

Referencias:

- Del libro "Data Science from Scratch", de Joel Grus, capítulo 9 ("Getting Data"), el apartado "Scraping the Web".
- La página oficial de documentación de esta librería:

https://pypi.org/project/beautifulsoup4/