



Programación. Python

Web scraping

Introducción: HTML

```
pequenno_codigo_html = """
<html>
  <head>
    <title>
      A web page
    </title>
  </head>
  <body>
    <p id="author">Joel Grus</p>
    <p id="subject">Data Science</p>
  </body>
</html>
"""
```

```
pequenno_codigo_html = """
<html>
  <head>
    <title>
      A web page
    </title>
  </head>
  <body>
    <p id="author">Joel Grus</p>
    <p id="subject">Data Science</p>
  </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')
```

```
print(primer_parrafo.text)  
print("-----")  
print(primer_parrafo.text.split())  
print(primer_parrafo["id"])  
todos_los_parrafos = soup_pequenno_codigo.find_all('p')  
print(todos_los_parrafos)  
print(todos_los_parrafos[0].text)
```

Joel Grus

['Joel', 'Grus']

author

[<p id="author">Joel Grus</p>, <p id="subject">Data Science</p>]

Joel Grus

```
pequenno_codigo_html = """
```

```
<html>
```

```
<head>
```

```
<title>
```

```
    A web page
```

```
</title>
```

```
</head>
```

```
<body>
```

```
<p id="author">Joel Grus</p>
```

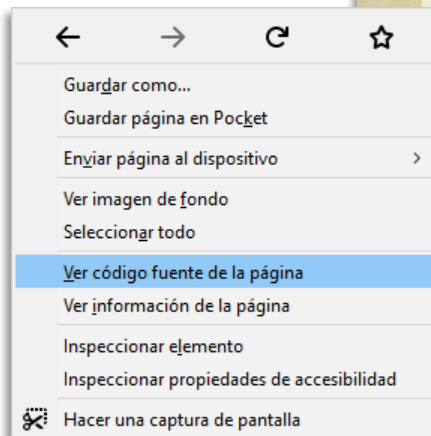
```
<p id="subject">Data Science</p>
```

```
</body>
```

```
</html>
```

```
"""
```

Una página web real y su fuente



```
1
2 <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
3
4 <!--
5   Design by Free CSS Templates
6   http://www.freecsstemplates.org
7   Released for free under a Creative Commons Attribution 2.5 License
8   Name       : C. Pareja Flores
9   Description: Home page
10  Version    : 1.0
11 -->
12
13 <html xmlns="http://www.w3.org/1999/xhtml">
14
15 <head>
16   <meta name="keywords" content="" />
17   <meta name="description" content="" />
18   <meta http-equiv="content-type" content="text/html; charset=utf-8" />
19   <title>Cristóbal Pareja Flores, página web</title>
20   <link href="/style.css" rel="stylesheet" type="text/css" media="screen" />
21 </head>
22
23 <body>
24   <div id="wrapper">
25     <div id="page">
26       <div id="page-bgtop">
27         <div id="page-bgbtm">
28
29           <div id="content">
30
31             <div class="post">
32               <h2 class="title"><a href="#">Docencia</a></h2>
33               <p class="meta">
34                 <span class="date">Curso 2013-14</span><span class="posted">Actualizado: 2014-abril-06</span>
35               </p>

```

Una página
web real
y su fuente

```
1
2 <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
3
4 <!--
5   Design by Free CSS Templates
6   http://www.freecsstemplates.org
7   Released for free under a Creative Commons Attribution 2.5 License
8   Name       : C. Pareja Flores
9   Description: Home page
10  Version    : 1.0
11 -->
12
13 <html xmlns="http://www.w3.org/1999/xhtml">
14
15 <head>
16   <meta name="keywords" content="" />
17   <meta name="description" content="" />
18   <meta http-equiv="content-type" content="text/html; charset=utf-8" />
19   <title>Cristóbal Pareja Flores, página web</title>
20   <link href="./style.css" rel="stylesheet" type="text/css" media="screen" />
21 </head>
22
23 <body>
24   <div id="wrapper">
25     <div id="page">
26       <div id="page-bgtop">
27         <div id="page-bgbtm">
28           <div id="content">
29             <div class="post">
30               <h2 class="title"><a href="#">Docencia</a></h2>
31               <p class="meta">
32                 <span class="date">Curso 2013-14</span><span class="posted">Actualizado: 2014-abril-06</span>
33               </p>
34             </div>
35           </div>
36         </div>
37       </div>
38     </div>
39   </div>
40 </body>
41 </html>
```

Una página
web real
y su fuente

Descarga en Python

```
import requests
```

```
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
```

```
http://www.freecsstemplates.org
```

```
Released for free under a Creative Commons Attribution 2.5 L
```

```
.....  
ga.src = ('https:' == document.location.protocol ? 'https://ssl' : 'http://www') + '.google-anal  
ytics.com/ga.js';  
var s = document.getElementsByTagName('script')[0]; s.parentNode.insertBefore(ga, s);  
})();
```

```
</script>
```

```
</body>
```

```
</html>
```

Acceso a sus elementos

```
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>Crist bal Pareja Flores, p gina web</title>
<link href="./style.css" media="screen" rel="stylesheet" type="text/css"/>
</head>
-----
<p class="meta">
<span class="date">Curso 2013-14</span><span class="posted">Actualizado: 20
14-abril-06</span>
</p>
```


Acceso a sus elementos

```
mi_pag_web = BeautifulSoup(html)
print(mi_pag_web.prettify())
print("-----")
print(mi_pag_web.find_all('p'))
print(mi_pag_web.find_all('p', {'class': 'meta'}))
```

```
<head>
<meta content="">
<meta content="">
<meta content="te">
<title>CristÃ³bal
<link href="./sty
</head>
-----
<p class="meta">
<span class="date"
14-abril-06</span>
</p>
```

```
print(len(mi_pag_web.find_all('p')))
print("-----")
print(mi_pag_web.find_all('p')[8])
print("-----")
print(mi_pag_web.find_all('p', {'class': 'meta'}))
print("-----")
anclas_o_hiperenlaces = mi_pag_web.find_all('a')
print(len(anclas_o_hiperenlaces))
print(anclas_o_hiperenlaces[3])
print(anclas_o_hiperenlaces[3].get("href"))
print(anclas_o_hiperenlaces[3].get_text())
```

```
22
-----
<p class="meta">
<span class="date">2015</span><span class="posted">Actualizado: 2015-abril-6</span>
</p>
-----
[<p class="meta">
<span class="date">Curso 2013-14</span><span class="posted">Actualizado: 2014-abril-06</span>
</p>, <p class="meta">
<span class="date">Hasta 2014</span><span class="posted">Actualizado: 2014-dic-23</span>
</p>, <p class="meta">
<span class="date">2015</span><span class="posted">Actualizado: 2015-abril-6</span>
</p>, <p class="meta">
<span class="date">Hasta 2014</span><span class="posted">Actualizado: 2014-abril-6</span>
</p>]
-----
46
<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:

1. Del libro "Data Science from Scratch", de Joel Grus, capítulo 9 ("Getting Data"), el apartado "Scraping the Web".
2. La página oficial de documentación de esta librería:
<https://pypi.org/project/beautifulsoup4/>