ntic master
revolucionamos la comunicación

UNIVERSIDAD COMPLUTENSE DE MADRID

Programación. Python

Web scraping

Cristóbal Pareja Flores

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>
"""
```

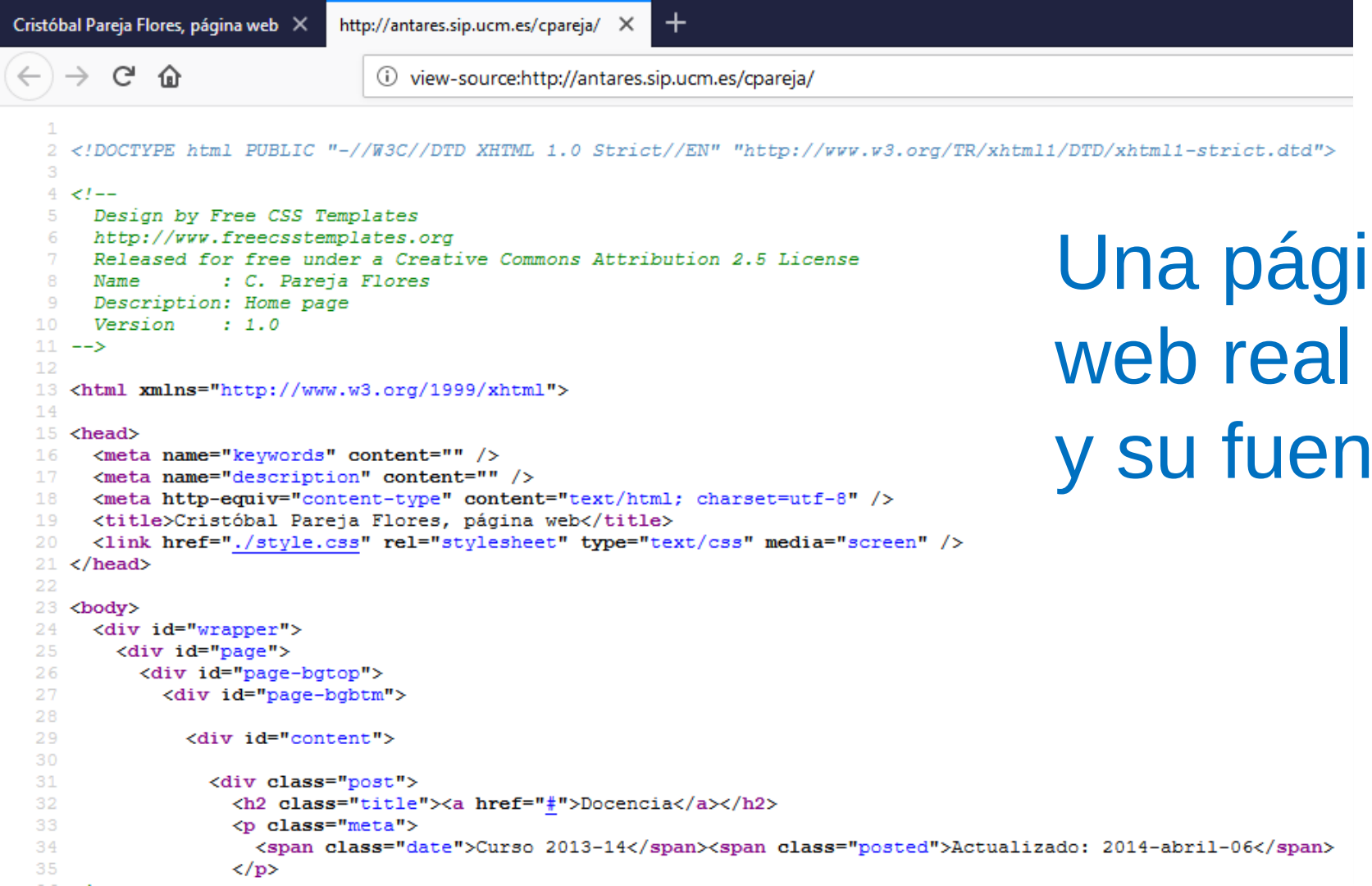
Introducción: HTML… y BeautifulSoup

```
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



Una página web real y su fuente

Cristóbal Pareja Flores, página web | http://antares.sip.ucm.es/cpareja/

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

Una página web real y su fuente

```
<!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 License
  Name        : C. Pareja Flores
  Description: Home page
  Version     : 1.0
-->

<html xmlns="http://www.w3.org/1999/xhtml">

<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" />
</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>
              <p class="meta">
                <span class="date">Curso 2013-14</span><span class="posted">Actualizado: 2014-abril-06</span>
              </p>
```

Descarga en Python

```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-s
trict.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>CristÃ³bal Pareja Flores, pÃ¡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 = Beau
print(mi_pag_web.
print("----------
print(mi_pag_web.
```

```
<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/