SISTEMAS WEB
CURSO 2022/2023

HTTP - Web Scraping
Beautiful Soup

WEB SCRAPING

- Se utilizará la librería **Beautiful Soup** de Python que permite la extracción de datos de archivos HTML y XML.
- Instalación de la libreria:

  python -m pip install BeautifulSoup4

- **Documentación:**

  https://www.crummy.com/software/BeautifulSoup/bs4/doc/

WEB SCRAPING – PROBANDO BEAUTIFUL SOUP

- Prueba la librería con la pagina de www.google.es

```
# -*- coding: UTF-8 -*-
import requests
from bs4 import BeautifulSoup

response = requests.get('http://google.com/')

print ("    STATUS: " + str(response.status_code)
print ("    CABECERAS: " + str(response.headers)

cuerpo _respuesta= response.content

if response.status_code == 200:

   # Pasamos el contenido HTML de la web a un objeto BeautifulSoup()
   html = BeautifulSoup(cuerpo_respuesta , "html.parser")

   print ("-------------------------------------“)
   print (response.content)
```

En **cuerpo_respuesta** tenemos la pagina html de www.google.es/

Cont ….

WEB SCRAPING – PROBANDO BEAUTIFUL SOUP

```
print ("---------------------------------------")
print (html.head)

print("---------------------------------------")
for meta in html.find_all('meta'):
    print(meta)
print (html.head.name)
print (html.head.meta)
print (html.head.meta['content'])

print("---------------------------------------")
print(html.head.title)
print(html.head.title.string)

print("---------------------------------------")
print (html.head.p)
print (html.body.p)
print (html.p)
print (html.p.name)
print (html.p.string)
print (html.p.text)
print (html.p.a)
print (html.p.a.string)
print (html.p.a['href'])
print (html.p['style'])
print (html.p.get('style'))
```

```
print("---------------------------------------")
print (html.a)
print (html.a.parent)
print (html.a.parent.name)
print (html.p.a.parent.name)

print("---------------------------------------")
# Todos los enlaces
for enlace in html.find_all('a'):
    print(enlace)
    print(enlace.string)
    print(enlace['href'])
    print(enlace.get('href'))

print("---------------------------------------")
# Todos los textos
print(html.get_text())

print("---------------------------------------")
for div in html.find_all('div'):
    print(div)
print("---------------------------------------")
for div in html.find_all('div', {'id':'guser'}):
    print(div)
print("---------------------------------------")
for div in html.find_all('div'):
    if div.has_attr('class'):
        clase = div['class']
        print(div)
        print(clase)
        print (clase[0])
```

Introduce las instrucciones poco a poco y observa que imprime

EJERCICIO

- Partiendo de programa ***sending_form_directorio_es.py*** que realiza consultas en el directorios de la *ehu* y utilizando la librería *BeautifulSoup*, conseguir que muestre el nombre y apellidos junto con el **enlace** de las personas encontradas en el directorio de la *ehu*.