

A stylized sunburst graphic in shades of purple and blue, located in the top-left corner of the slide.

Extracción de la Información de la Web

Máster Online en Ciencia de Datos

UCO
ONLINE

Four horizontal bars of equal length, colored yellow, red, yellow, and red from left to right, located at the bottom of the slide.


Dr. José Raúl Romero

Profesor Titular de la Universidad de Córdoba y Doctor en Ingeniería Informática por la Universidad de Málaga. Sus líneas actuales de trabajo se centran en la democratización de la ciencia de datos (*Automated ML* y *Explainable Artificial Intelligence*), aprendizaje automático evolutivo y analítica de software (aplicación de aprendizaje y optimización a la mejora del proceso de desarrollo de software).

Miembro del Consejo de Administración de la *European Association for Data Science*, e investigador senior del Instituto de Investigación Andaluz de *Data Science and Computational Intelligence*.

Director del **Máster Online en Ciencia de Datos** de la Universidad de Córdoba.



A stylized sunburst or fan-like graphic in shades of purple and blue, located in the top-left corner of the slide.

Programación de Extracción de Información de la Web mediante *Web Scraping*

El *toolkit* de Python para *Web Scraping*

Fuentes de datos

Ficheros, APIs, sitios web (**Requests**)

Parseo de datos

Regular-Expressions (re), BeautifulSoup

Estructuras de datos

Listas/diccionarios de Python, PANDAS

Modelos

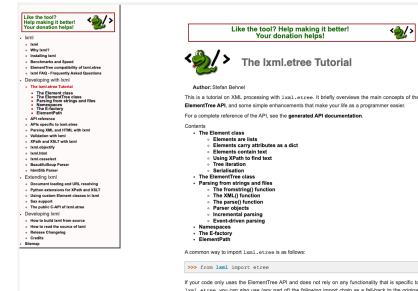
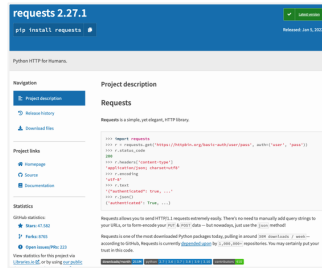
Regresión lineal, clasificación, clustering, etc.

El *toolkit* de Python para *Web Scraping*

1. Paquete **requests**. Recuperar el contenido de la página web
2. Librería **Beautiful Soup** [*bs4*]. Extraer datos de HTML y XML
3. Librería **lxml**. *Parser* que puede ser utilizado junto con *bs4*

```
pip install requests
pip install bs4
pip install lxml
```

- Es la librería parser recomendada en documentación de Beautiful Soup
- lxml es una librería externa a *bs4* (ojo a compatibilidad y actualizaciones) pero destaca por su velocidad



<https://requests.readthedocs.io/en/latest/>

<https://www.crummy.com/software/BeautifulSoup/>

<https://lxml.de/tutorial.html>

Accediendo al contenido de la página Web

Recupera el contenido de la página Web

Requests **hace** todo el trabajo

```
page = requests.get(url)
```

```
page.status_code
```

```
page.content
```

Devuelve el estado de la solicitud HTTP

200 - success

404 - page not found

Devuelve el contenido de la página, en bytes.

Accediendo al contenido de la página Web

```
import requests

url = 'http://www.uco.es/user/in1rosaj/publicas.html'
session = requests.Session()
r = session.get(url)
html = r.text
print(html[:200])
```

BeautifulSoup

bs4 ofrece mecanismos de navegabilidad por los elementos parseados de la página Web

Se puede utilizar con distintos parseadores (p.ej. lxml)

`get_text()` devuelve todo lo que hay contenido en la Web, no solo texto (p.ej. código Javascript)

```
1 import requests
2 from bs4 import BeautifulSoup
3
4 url = 'https://www.cmmedia.es/noticias/castilla-la-mancha/'
5 respuesta = requests.get(url)
6 contenido = BeautifulSoup(respuesta.text, 'lxml')
7
8 noticias = contenido.find('ul', attrs={'class': 'news-list'})
9 articulos = noticias.findChildren('div', attrs={'class': 'media-body'})
10
11 noticias = []
12 for articulo in articulos:
13     noticias.append({
14         'url': articulo.find('h3').a.get('href'),
15         'titulo': articulo.find('h3').get_text().strip()
16     })
17
18 for noticia in noticias:
19     print(noticia)
```

```
{'url': 'https://www.cmmedia.es/noticias/castilla-la-mancha/23-empresas-entidades-y-personas-destacadas-de-castilla-la-mancha-seran-reconocidas-por-su-labor-en-el-medioambiente', 'titulo': '23 empresas, entidades y personas destacadas de Castilla-La Mancha serán reconocidas por su labor en el medioambiente'}
{'url': 'https://www.cmmedia.es/noticias/castilla-la-mancha/un-facultativo-de-toledo-asesora-a-la-oms-sobre-el-diagnostico-de-la-mastocitosis', 'titulo': 'Un facultativo de Toledo asesora a la OMS sobre el diagnóstico de la mastocitosis'}
{'url': 'https://www.cmmedia.es/noticias/castilla-la-mancha/hieren-con-arma-blanca-a-hombre-en-el-puente-de-alcantara-de-toledo', 'titulo': 'Hieren con arma blanca a un hombre en el Puente de Alcántara de Toledo'}
{'url': 'https://www.cmmedia.es/noticias/castilla-la-mancha/una-reyerta-en-hellin-albacete-deja-tres-personas-heridas-una-de-ellas-por-arma-blanca', 'titulo': 'Una reyerta en Hellín (Albacete) deja tres personas heridas, una de ellas por arma blanca'}
{'url': 'https://www.cmmedia.es/noticias/castilla-la-mancha/campo-de-criptana-debera-pagar-a-un-socorrista-los-salarios-que-no-percibio-por-el-cierre-de-la-piscina-durante-la-pandemia', 'titulo': 'Campo de Criptana deberá pagar a un socorrista los salarios que no percibió por el cierre de la piscina durante la pandemia'}
{'url': 'https://www.cmmedia.es/noticias/castilla-la-mancha/un-incendio-en-un-bloque-de-pisos-en-socuellamos-ciudad-real-deja-siete-personas-afectadas-dos-de-ellas-guardias-civiles', 'titulo': 'Un incendio en un bloque de pisos de Socuéllamos (Ciudad Real) deja siete personas afectadas, dos de ellas guardias civiles'}
```


BeautifulSoup

bs4 permite parsear el contenido de la página Web

```
soup = BeautifulSoup(pagina.content, "html.parser")
```

```
soup.title
```

```
soup.title.text
```

Devuelve el contexto completo del elemento, incluyendo la etiqueta:

```
<title data-rh="true">El Diario  
Córdoba</title>
```

Devuelve la parte de contenido de texto de la etiqueta:

El Dario Córdoba

BeautifulSoup

- bs4 hace más llevadero trabajar sobre estructuras HTML
- Ofrece funciones de acceso rápido a elementos:
 - Se facilita mucho la labor si el documento está bien etiquetado y marcado con identificadores, clases, etc.
- Por ejemplo: encontrar todos los enlaces que hay en la página

```
link_list = l.get('href') for l in soup.findAll('a')
```

La página Web es un árbol que, aunque no es navegable directamente:

```
tree = bs4.BeautifulSoup(source)

## get html root node
root_node = tree.html

## get head from root using contents
head = root_node.contents[0]

## get body from root
body = root_node.contents[1]

## could directly access body
tree.body
```

Antes de continuar...

- Abre la siguiente dirección:

<https://www.geeksforgeeks.org/implementing-web-scraping-python-beautiful-soup/>

- Lee el tutorial e implementa los ejemplos
- Te tomará poco tiempo y resulta muy ilustrativo

A stylized sunburst graphic in shades of purple and blue, located in the top-left corner of the slide. It features a semi-circle on the left with several rays extending outwards to the right.

¡Gracias!

UCO
ONLINE

A horizontal bar at the bottom of the slide consisting of three equal-width rectangles of yellow, red, and blue, representing the colors of the Spanish flag.