Curso de Python para astrónomos aficionados... ¡o no!

Introducción al Web Scraping

S. Alonso [a.k.a. Zerjillo] - zerjioi@ugr.es y
Javier Flores - javierfloresmartin1992@gmail.com

Abril - Mayo 2022

Contenidos

- Introducción al Web-Scraping
- Conceptos básicos: la `WWW`, `HTML`, `http`, `AJAX`
- Formatos de intercambio de datos: `XML`, `json`, `CSV`
- Asuntos “legales” del web-scraping

Introducción al Web-Scraping

Definición

Web scraping es una técnica utilizada para extraer información de sitios web mediante programas de software

Wikipedia

Términos relacionados

- *Web harvesting*, *web data extraction*: sinónimos
- *Crawler*, *spider*, *spiderbot*: Robot (software) que navega por la *WWW*

Posibles usos del web-scraping

- Análisis de oferta de la competencia
- Comparación de precios
- Rastreo de los intereses de los usuarios en las redes sociales
- Marketing de contenidos, rastreo de keywords para SEO
- Registro y análisis de datos de clientes
- Lo que se te ocurra...

La World Wide Web

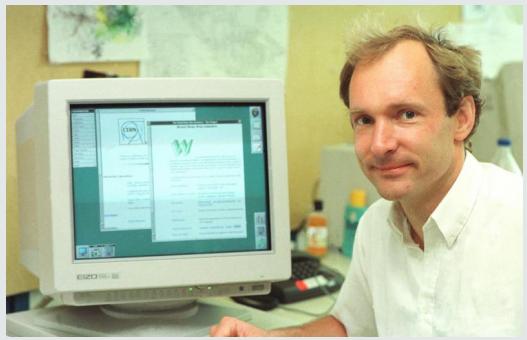
¿Qué es?

Un sistema de distribución de documentos de **hipertexto** interconectados y accesibles a través de Internet.

Wikipedia

Tim Berners-Lee

- "Padre" de la **WWW**
- Primera conexión: 1989
- Propuso las ideas fundamentales: **HTML**, **HTTP**, **URL**



HTML: Hypertext Markup Language

¿Qué es?

Es el lenguaje que define el contenido y estructura de la página web. Originalmente los contenidos eran estáticos y el aspecto también se definía con este lenguaje.

Hoy en día

Se complementa con:

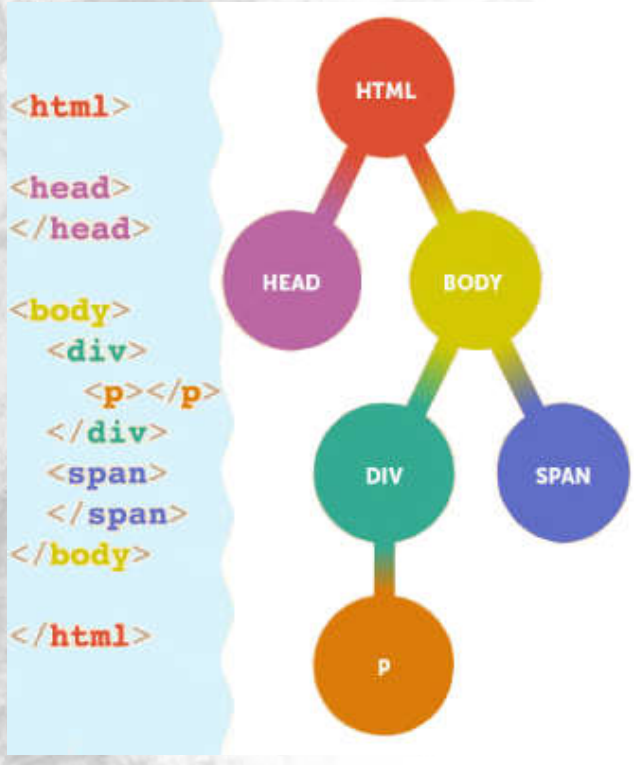
- CSS: *Cascade Style Sheets*. Permiten definir el aspecto de la página web.
- `JavaScript`: Lenguaje de programación que permite interactividad en la página web.

HTML: Características principales

- Formato de texto plano: puede escribirse *hasta* con el bloc de notas
- La información de la página se estructura mediante **marcas**, **etiquetas** o *tags*
- La extensión típica de archivo es `.html` (en ocasiones y por herencia de sistemas operativos inferiores aún puede verse como `.htm`)
- Los elementos no textuales (imágenes, vídeos, estilos, scripts...) se guardan en ficheros distintos
- Cada página de un sitio web suele escribirse en un fichero diferente

Marcas de HTML

- Comienzan por `<` y terminan por `>`
- Apertura de una marca: `<marca>`.
  Cierre: `</marca>`.
  Autocierre: `<marca />` $\Rightarrow$ `<marca></marca>`
- Anidación: Dentro de cada marca puede haber “cualquier” contenido (incluyendo otras marcas)
- Un documento HTML es en última instancia un árbol de etiquetas



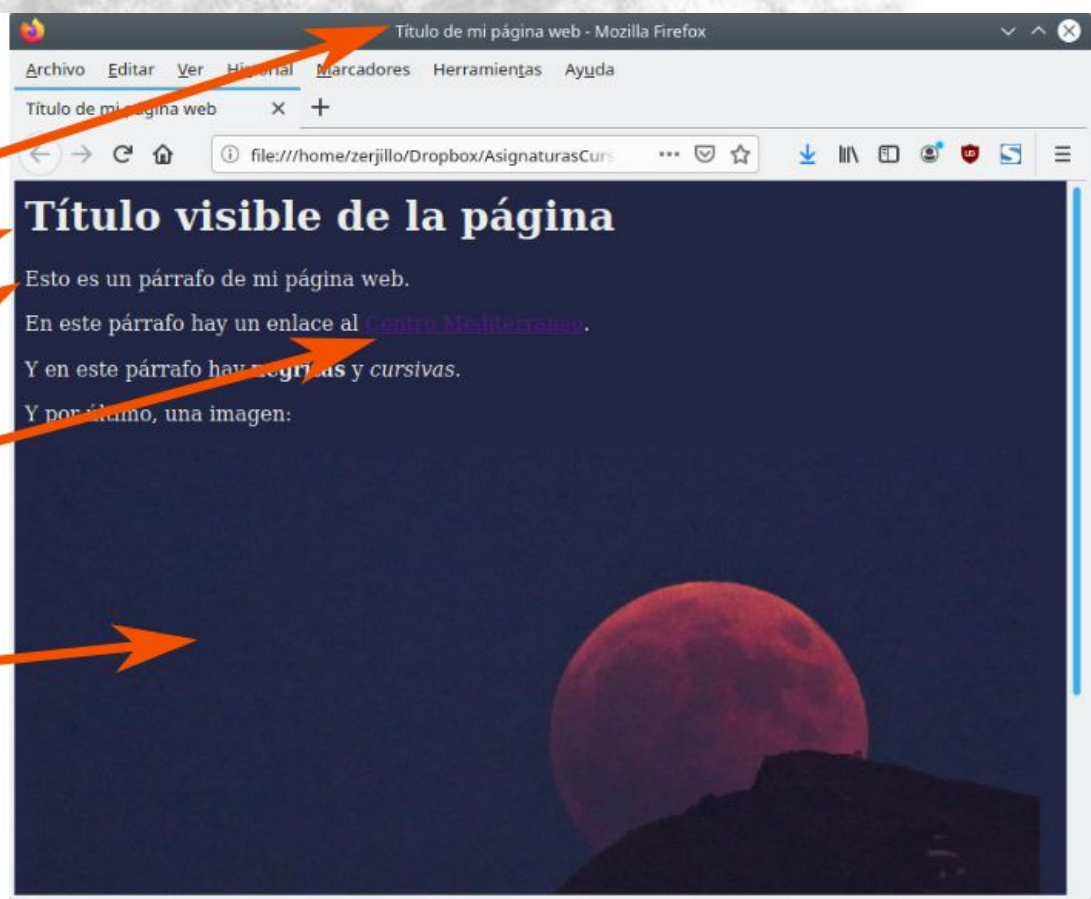
Fuente de la imagen

Ejemplos de marcas

- `<html>...</html>`: Comienzo y fin de la página
- `<head>...</head>`: Cabecera de la página
- `<body>...</body>`: Cuerpo (contenidos)
- `<p>...</p>`: Un párrafo
- `<strong>...</strong>`: “Negrita”
- `<a>...</a>`: Un enlace

Ejemplo práctico y minimalista de página web

```html
<html>
  <head>
    <title>Título de mi página web</title>
  </head>
  <body style="background-color: #222544; color: #eee;">
    <h1>Título visible de la página</h1>

    <p>Esto es un párrafo de mi página web.</p>

    <p>En este párrafo hay un enlace al <a
href="https://cemed.ugr.es/">Centro Mediterraneo</a>.</p>

    <p>Y en este párrafo hay <strong>negritas</strong> y
<em>cursivas</em>.</p>

    <p>Y por último, una imagen:</p>

    <img src="eclipse.jpg" />
  </body>
</html>
```



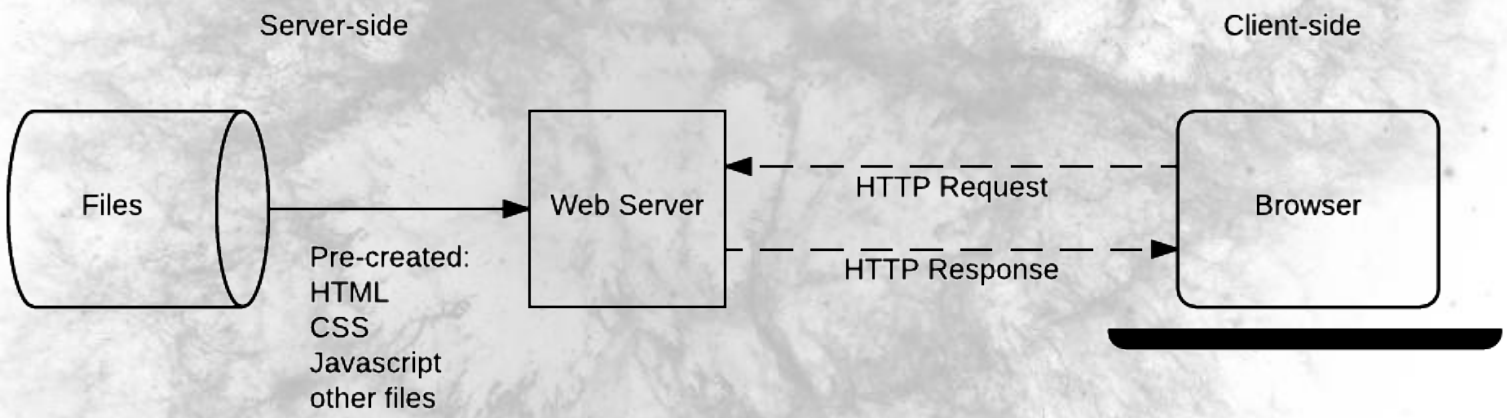
HTML: Herramientas y enlaces de interés

- Primeros pasos en la web
- HTML Tutorial
- CSS Tutorial
- JavaScript Tutorial
- Navegador → botón derecho → Ver código fuente de la página
- Navegador → Herramientas → Desarrollador web → Alternar Herramientas → Inspector

¿Cómo se transmiten las páginas web?

La **WWW** sigue una arquitectura **cliente-servidor**:

- **Cliente:** El navegador (Firefox, Chrome...) hace peticiones de páginas (*request*)
- **Servidor:** Un programa que tras una petición decide que página hay que devolver (*response*)



Fuente de la imagen

HTTP: Hypertext transfer protocol

¿Qué es?

El *protocolo de transferencia de hipertexto* es el protocolo de comunicación que permite las transferencias de información en la **WWW**

Wikipedia

Características

- Protocolo de solo texto (tanto la petición como la respuesta)
- "Inseguro" (no se encripta). Solución: usar **HTTPS**

Ejemplo HTTP

Petición

```
GET https://cemed.ugr.es

Host: cemed.ugr.es
User-Agent: Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:74.0) Gecko/20100101 Firefox/74.0
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8
Accept-Language: es-ES,es;q=0.8,en-US;q=0.5,en;q=0.3
Accept-Encoding: gzip, deflate, br
DNT: 1
Connection: keep-alive
Upgrade-Insecure-Requests: 1
```

Respuesta

```
HTTP/1.1 200 OK
Date: Mon, 13 Apr 2020 23:48:36 GMT
Server: Apache/2.4.41 (codeit) OpenSSL/1.1.1c PHP/7.3.9
X-Powered-By: PHP/7.3.9
Vary: Accept-Encoding,Cookie
Cache-Control: max-age=3, must-revalidate
Content-Encoding: gzip
Content-Length: 14293
Last-Modified: Mon, 13 Apr 2020 23:37:47 GMT
Connection: close
Content-Type: text/html; charset=UTF-8
```

HTTP: Herramientas y enlaces de interés

- Cómo funciona la web
- Visión General Cliente-Servidor
- Navegador → Herramientas → Desarrollador web → Alternar Herramientas → Red
- Extensión para Firefox: **HTTP** *request maker*
- Extensión para Firefox: *Disable JavaScript*

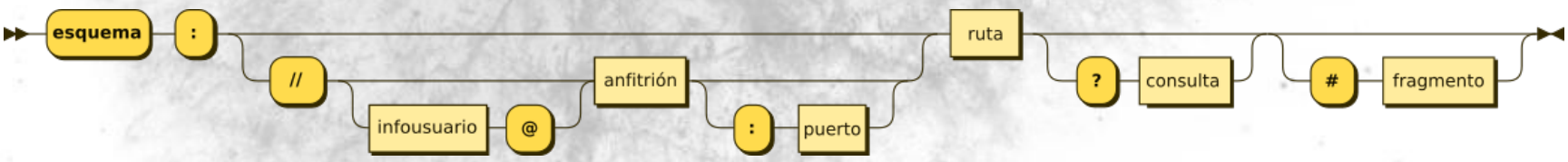
URL: Localizador de recursos uniforme

¿Qué es?

Un localizador de recursos uniforme es una cadena de caracteres que identifica los recursos (que pueden variar en el tiempo) de una red de forma unívoca.

Sintaxis

- **Esquema:** En nuestro caso `http:` o `https:`
- **Autoridad:** Dominio o `IP`
- **Ruta:** organizada jerárquicamente
- **Consulta:** posibles parámetros
- **Fragmento:** localizador de parte del documento



Fuente de la imagen

AJAX: Asynchronous JavaScript And XML

¿Qué es?

Técnica de desarrollo web para crear aplicaciones interactivas que se ejecutan en el cliente mientras se mantiene la comunicación asíncrona con el servidor en segundo plano. De esta forma es posible realizar cambios sobre las páginas sin necesidad de recargarlas

Wikipedia

¿Qué implica?

Que todas las páginas que usen esta técnica (o incluso solo `JavaScript`) no se podrán "escrapear" de manera sencilla

AJAX: Asynchronous JavaScript And XML

¿Qué es?

Técnica de desarrollo web para crear aplicaciones interactivas que se ejecutan en el cliente mientras se mantiene la comunicación asíncrona con el servidor en segundo plano. De esta forma es posible realizar cambios sobre las páginas sin necesidad de recargarlas

Wikipedia

¿Qué implica?

Que todas las páginas que usen esta técnica (o incluso solo `JavaScript`) no se podrán "escrapear" de manera sencilla

¡Pero no te preocupes!

¡Lo resolveremos!

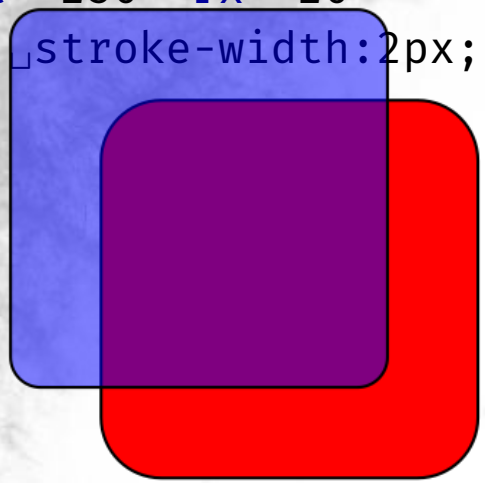
Otros formatos de intercambio de información habituales

- `XML`: *eXtensible Markup Language*. Metalenguaje de marcado. Actualmente `HTML` es un lenguaje derivado de `XML`
- `JSON`: *JavaScript Object Notation*. Su conversión a diccionarios y arrays es trivial en muchos lenguajes (incluido Python).
- `CSV`: *Comma-Separated Values*. Para datos tabulares: cada columna se separa de la siguiente por una coma (`,`). Cada fila se separa con un salto de línea (`\n`)

Ejemplo XML

Archivo SVG: Scalable Vector Graphics

```
1  <?xml version="1.0"?>
2  <!DOCTYPE svg PUBLIC "-//W3C//DTD␣SVG␣1.1//EN"
3    "http://www.w3.org/Graphics/SVG/1.1/DTD/svg11.dtd">
4
5  <svg xmlns="http://www.w3.org/2000/svg" width="467" height="462">
6  ␣␣<rect x="140" y="120" width="250" height="250" rx="40"
7      style="fill:#ff0000;stroke:#000000;stroke-width:2px;" />
8
9  ␣␣<rect x="80" y="60" width="250" height="250" rx="20"
10     style="fill:#0000ff;␣stroke:#000000;␣stroke-width:2px;
11 ␣␣␣␣␣fill-opacity:0.5;" />
12 </svg>
```



Ejemplo JSON

```
{
"mates": {
    "p1": {
      "pregunta": "5 + 7 = ?",
      "opciones": [
        "10",
        "11",
        "12",
        "13"
      ],
      "respuesta": "12"
    },
    "p2": {
      "pregunta": "12 - 8 = ?",
      "opciones": [
        "1",
        "2",
        "3",
        "4"
      ],
      "respuesta": "4"
    }
  }
}
```

Ejemplo CSV

```
"LatD", "LatM", "LatS", "NS", "LonD", "LonM", "LonS", "EW"
    41,      5,     59,  "N",     80,     39,      0,  "W"
    42,     52,     48,  "N",     97,     23,     23,  "W"
    46,     35,     59,  "N",    120,     30,     36,  "W"
    42,     16,     12,  "N",     71,     48,      0,  "W"
    43,     37,     48,  "N",     89,     46,     11,  "W"
```

LAtD	LatM	LatS	NS	LonD	LonM	LonS	EW
41	5	59	”N”	80	39	0	”W”
42	52	48	”N”	97	23	23	”W”
46	35	59	”N”	120	30	36	”W”
42	16	12	”N”	71	48	0	”W”
43	37	48	”N”	89	46	11	”W”

Legalidad del web-scraping

¿Es legal scrapear una web?

En principio **SÍ**: ¿Es legal el scraping?

Según el Reglamento General de Protección de Datos (RGPD) y la Ley Orgánica de Protección de Datos y Garantía de Derechos Digitales (LOPDGDD)

Esta técnica solo está permitida en los siguientes supuestos:

- Son fuentes de acceso público o los datos se recaban por un fin de interés público general
- Prevalece el interés del responsable del tratamiento sobre el derecho a la protección de datos
- La persona rastreada lo es bajo su consentimiento

Restricciones al web-scraping

¿Podemos hacer lo que queramos con los datos?

Aquí es donde puede estar el problema: Lo mismo scrapear una web es legal, pero usar los datos obtenidos para algún uso particular puede que no.

Condiciones de cada sitio web

Suelen indicar la prohibición o no del scrapeo.

Mecanismos anti-scraping

- Usar `cookies` o `JavaScript`
- `Captchas`: I'm not a robot
- Establecer límites de peticiones y conexiones
- Ofuscar o esconder los datos
- Detectar y bloquear fuentes maliciosas conocidas
- Localizar y frenar el acceso de site scrapers conocidos
- Actualizar constantemente los HTML tags de la página
- Utilizar contenido web falso para atrapar a los atacantes
- Usar `robots.txt`: Introducción a los archivos `robots.txt`

Ética del web-scraping

- No scrapees si no tienes permiso (respeta `robots.txt` y evita saltarte los `captchas`)
- Cuidado con sobrecargar los servidores scrapeados
- Cuidado con la protección de datos de lo scrapeado
- No hagas “cosas chungas” con los datos: spam, robo de identidades, plagios...
- Artículo interesante: Ethics in Web Scraping

Un último consejo

No reinventes la rueda

Si lo que necesitas es duplicar un página web completa ya hay herramientas preparadas y fáciles para ello: `wget`

- https://www.gnu.org/software/wget/