Web Scraping

Gustavo Juantorena [github.com/GEJ1]

Hoja de ruta

1. ¿Qué es el web scraping?
2. ¿Qué son las APIs?
3. ¿Cuándo conviene usar cada cosa?
4. Breve introducción a las tecnologías web.
5. Web Scraping con Python.

¿Qué es el web scraping?

La práctica de recopilar datos mediante un programa automatizado que consulta un servidor web

¿Qué es el web scraping?

La práctica de recopilar datos mediante un programa automatizado que consulta un servidor web

*El **web crawling** o indexación, se utiliza para indexar la información de la página mediante bots*

Un bot rastrea un sitio web, recorre todas las páginas y todos los enlaces, hasta la última línea del sitio web

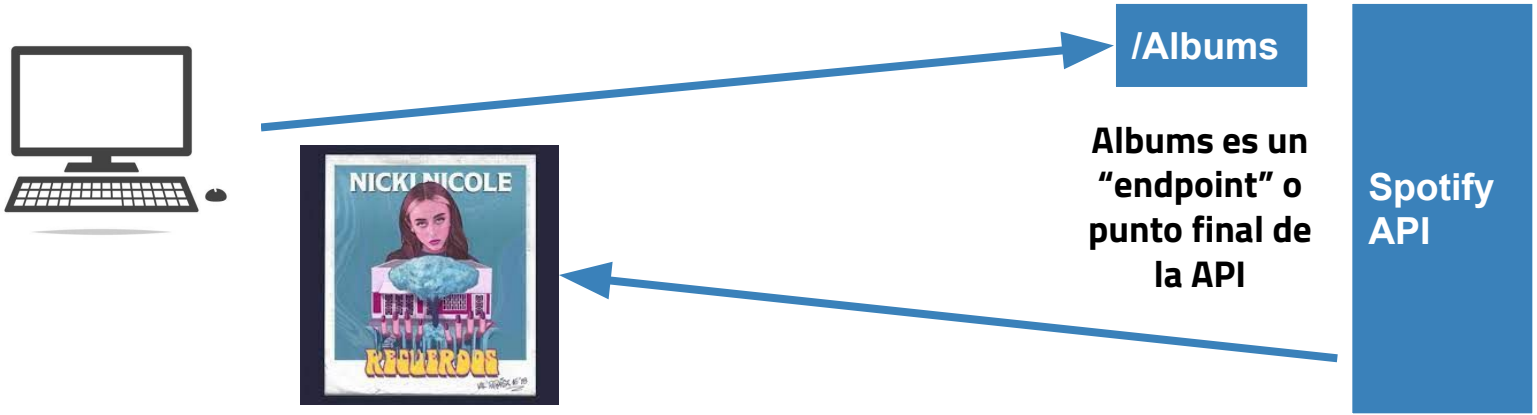
Antes de empezar ⚠️

1. ¿Lo que estoy haciendo es legal en mi país? ¿Puedo estar afectando al servidor?

2. ¿Es necesario utilizar web scraping o se podría haber resuelto de manera más sencilla?

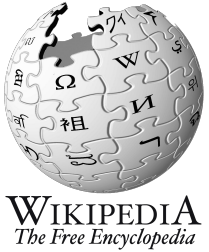
API (Interfaz de programación de aplicaciones)

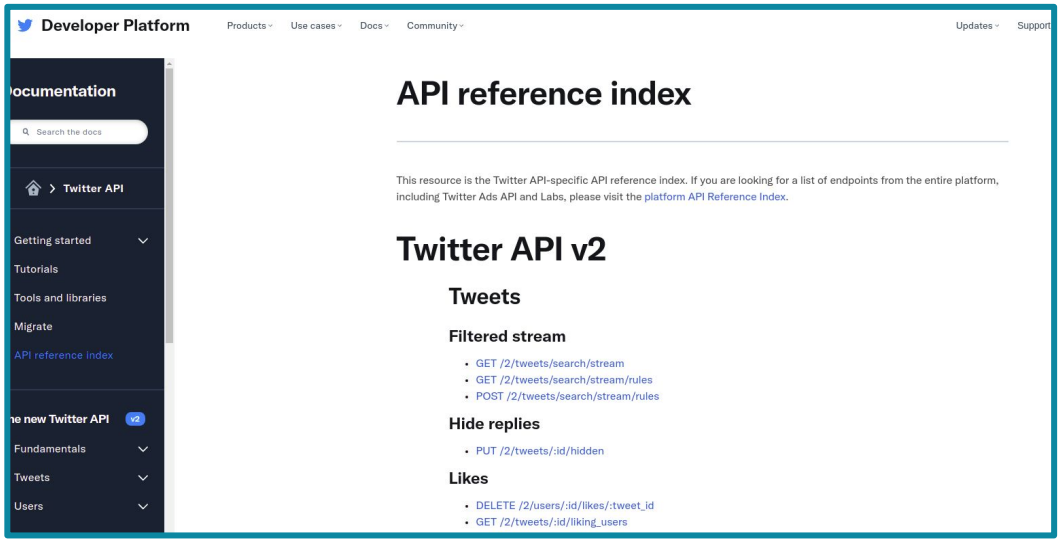
Conjunto de funciones, métodos y protocolos de comunicación que ofrece un programa para ser utilizado por otro.

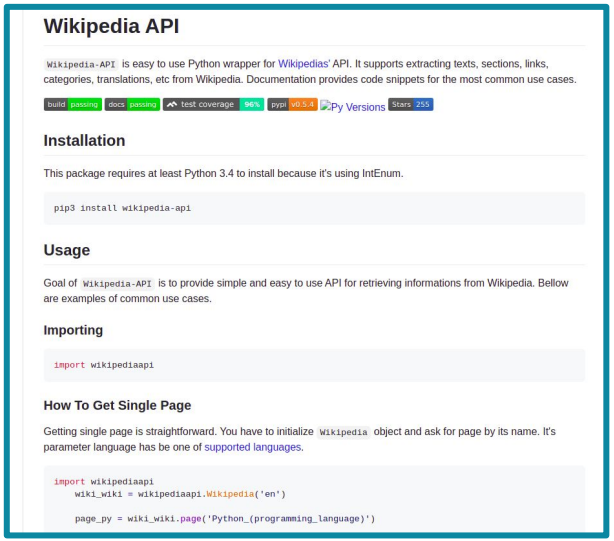


Nos provee el contenido con un determinado formato y especifica las limitaciones (ej: cantidad de pedidos al servidor).

Ejemplos de APIs





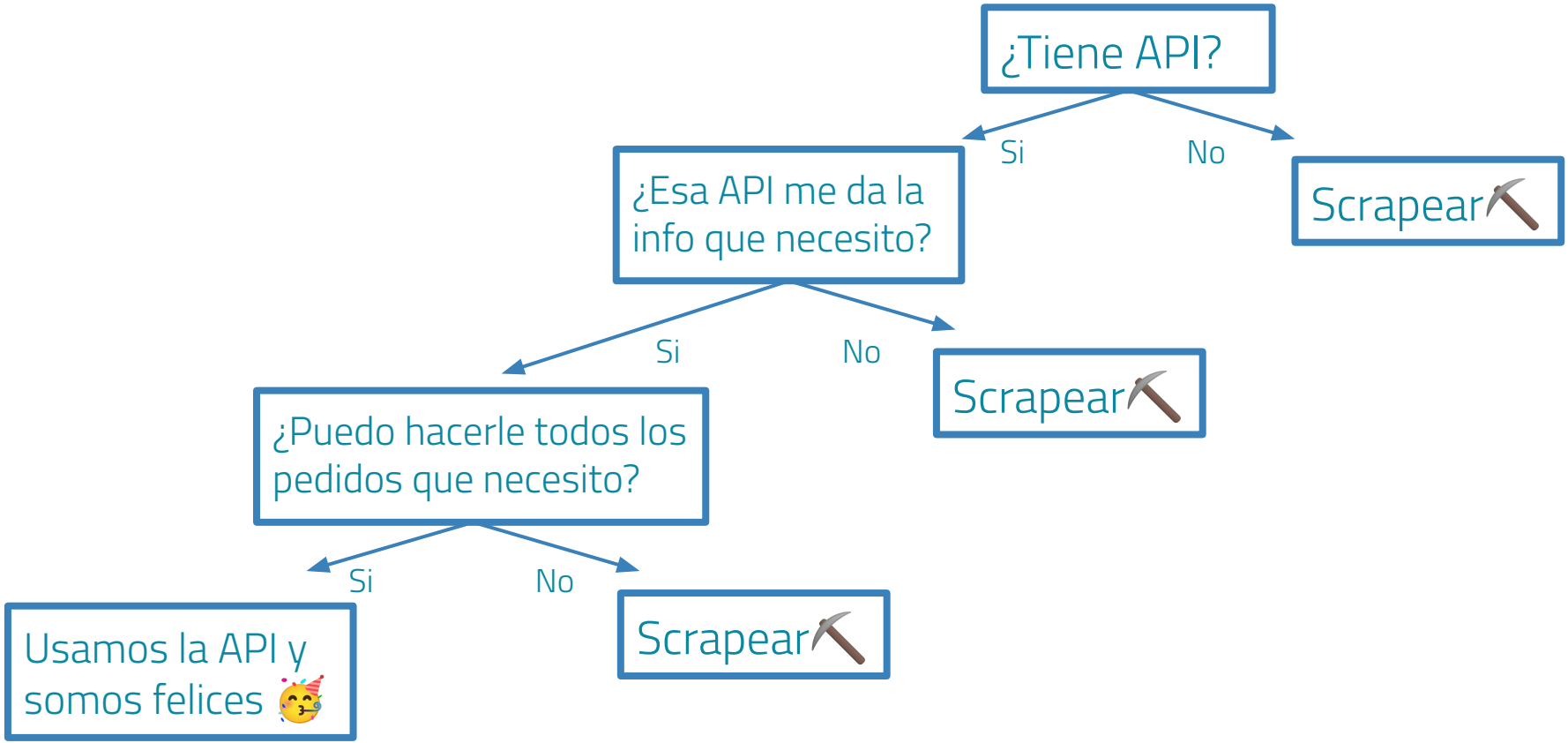


API vs Web scraping (I)

Web scraping	API
Extraer información de un sitio web usando un programa informático.	Proveer acceso a los datos de una aplicación, sistema operativo u otro servicio.

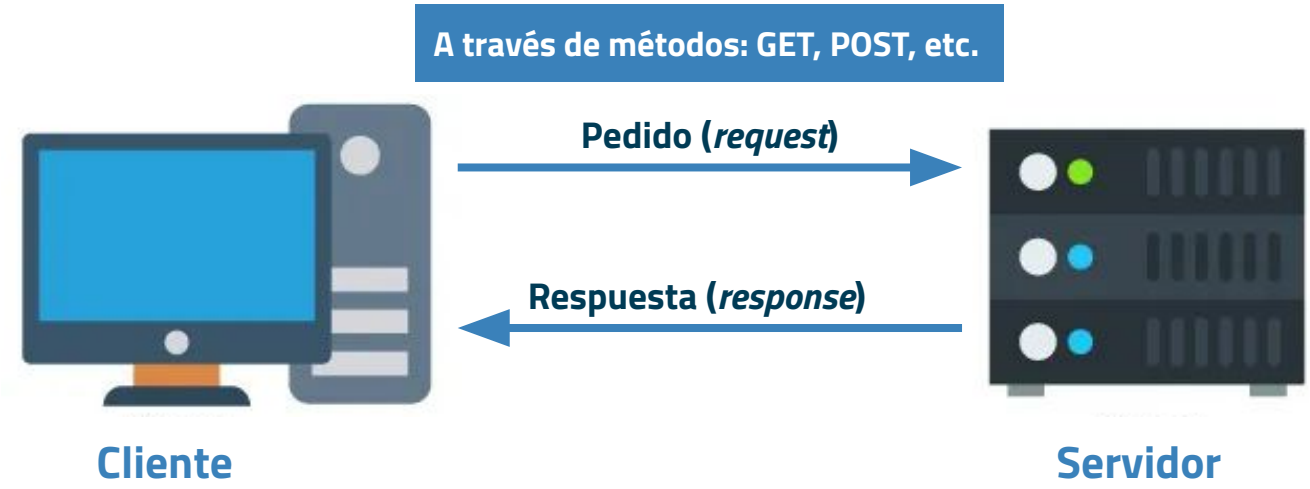
Mismo objetivo: Acceder a los datos del sitio web

API vs Web scraping (II)

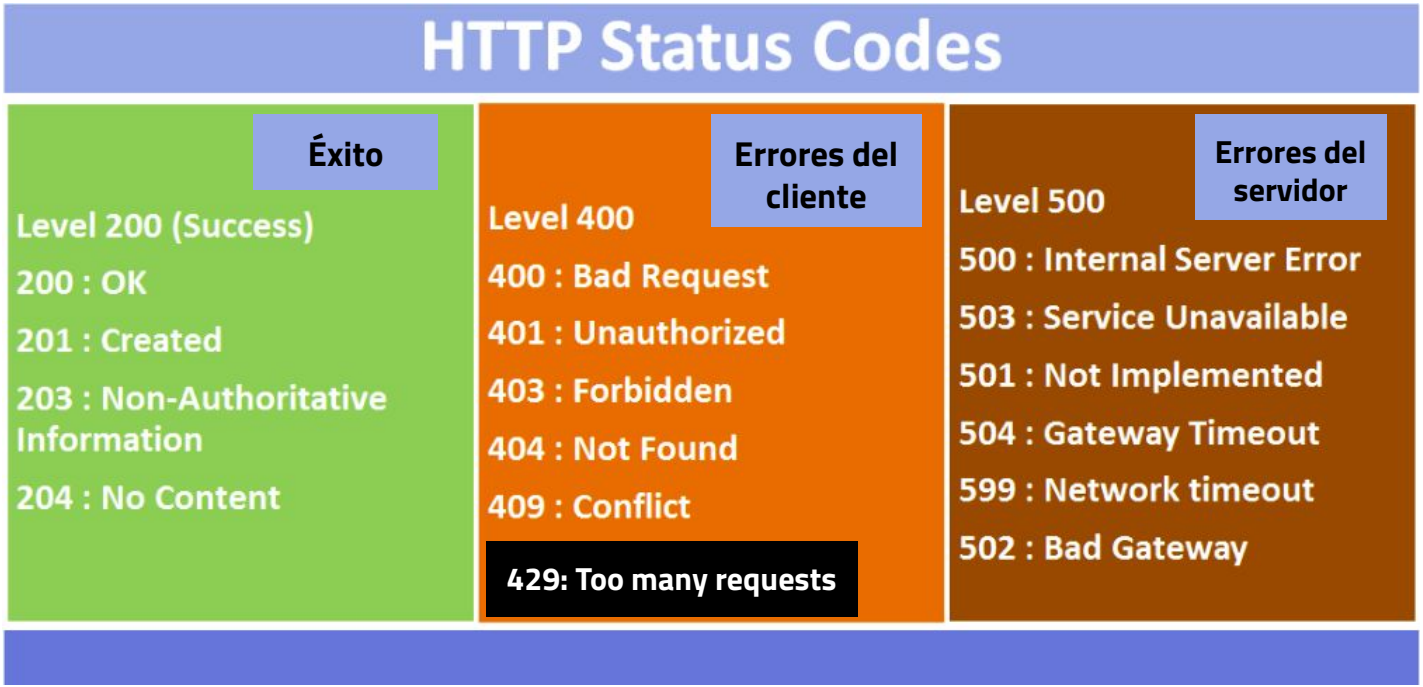


Conceptos básicos sobre la web (I)

- **HTTPS** (Protocolo seguro de transferencia de hipertexto): Protocolo de mensajería que permite a los navegadores web comunicarse con los servidores web (donde se almacenan los sitios web).



Conceptos básicos sobre la web (II)

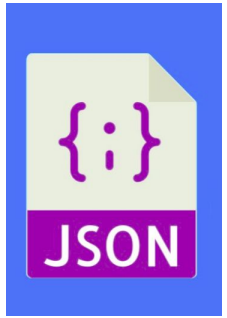


Fuente: https://gabicuesta.blogspot.com/2019/01/http-status-codes.html

Conceptos básicos sobre la web (III)

comma-separated values

```
album, año, ranking
The White Stripes, 1999, 2
De Stijl, 2000, 3
Lo mejor del amor, 1996, 1
```



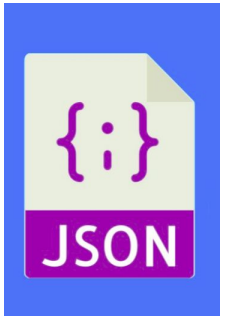
Javascript Object Notation

```
[
  {
    "album": "The White Stripes",
    "año": 1999,
    "ranking": 2
  },
  {
    "album": "De Stijl",
    "año": 2000,
    "ranking": 3
  },
  {
    "album": "Lo mejor del amor",
    "año": 1996,
    "ranking": 1
  }
]
```

https://csvjson.com/csv2json

Conceptos básicos sobre la web (III)

comma-separat
values



Javascript
Object
Notation

```
album, año, ranking
The White Stripes,
De Stijl, 2000, 3
Lo mejor del amor,
```

```
<?xml version="1.0" encoding="UTF-8"?>
<root>
  <element>
    <album>The White Stripes</album>
    <year number="true">1999</year>
    <ranking number="true">2</ranking>
  </element>
  <element>
    <album>De Stijl</album>
    <year number="true">2000</year>
    <ranking number="true">3</ranking>
  </element>
  <element>
    <album>Lo mejor del amor</album>
    <year number="true">1996</year>
    <ranking number="true">1</ranking>
  </element>
</root>
```

```
"The White Stripes",
999,
": 2

"De Stijl",
000,
": 3

"Lo mejor del amor",
996,
": 1
```

https://csvjson.com/csv2jso

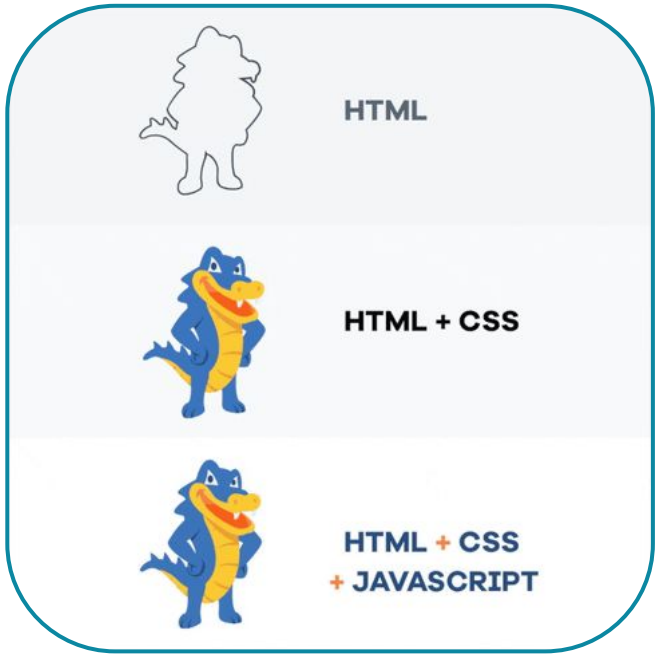
Conceptos básicos sobre la web (IV)

- **HTML, CSS y JavaScript son los tres lenguajes principales con los que está hecho la parte de la web que vemos.**

Conceptos básicos sobre la web (IV)

- **HTML, CSS y JavaScript son los 3 lenguajes principales con los que está hecho la parte de la web que vemos.**

Estructura Funcionalidad Estilo



Fuente: https://geekflare.com/es/css-formatting-optimization-tools/

HTML: Lenguaje de marcado de hipertexto
CSS: Hojas de estilo en cascada

```
<!DOCTYPE html>
<html>
  <head>
      <title>Mi primer página</title>
  </head>
  <body>

      <h1>Título</h1>
      <h2 style='color:red;'> Subtitulo en rojo</h2>
      <p>Primer párrafo.</p>
      <hr>

        <h3>Gatito<h3>
        <img style='width: 100px;'
src='https://i.pinimg.com/originals/d4/8d/07/d48d074f8c
9ec84486128222956867 54.jpg' >


  </body>
</html>
```

HTML: Lenguaje de marcado de hipertexto
CSS: Hojas de estilo en cascada

```
<!DOCTYPE html>
<html>
  <head>
       <title>Mi primer página</title>
  </head>
  <body>

       <h1>Título</h1>
       <h2 style='color:red;'> Subtitulo en rojo</h2>
       <p>Primer párrafo.</p>
       <hr>

         <h3>Gatito<h3>
         <img style='width: 100px;'
src='https://i.pinimg.com/originals/d4/8d/07/d48d074f8c
9ec84486128222956867 54.jpg' >


  </body>
</html>
```

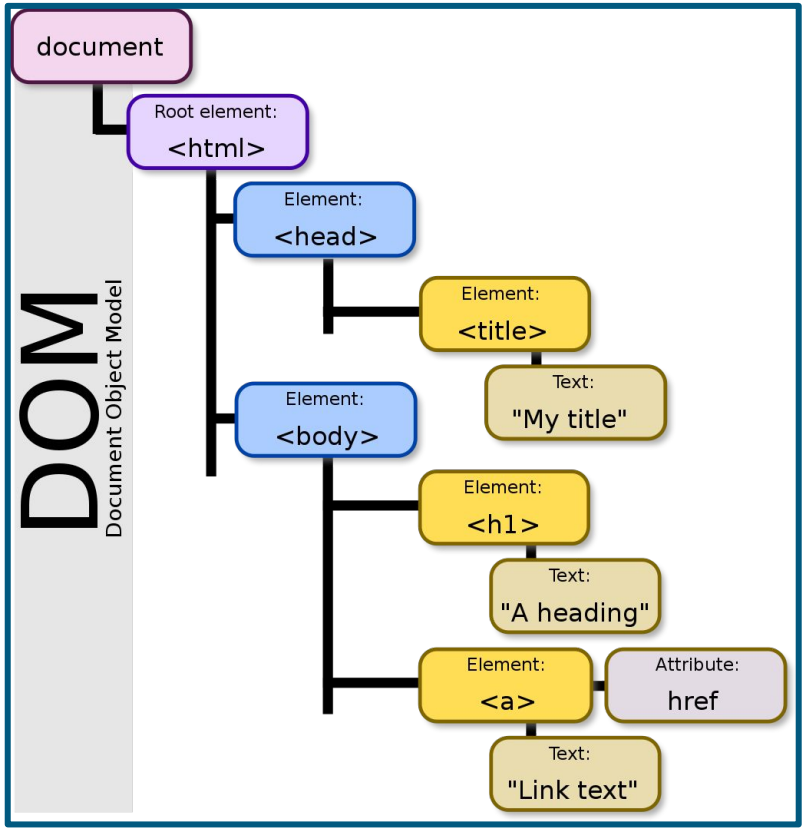
HTML TAG

CSS

HTML ATTRIBUTE

DOM (Document Object Model)

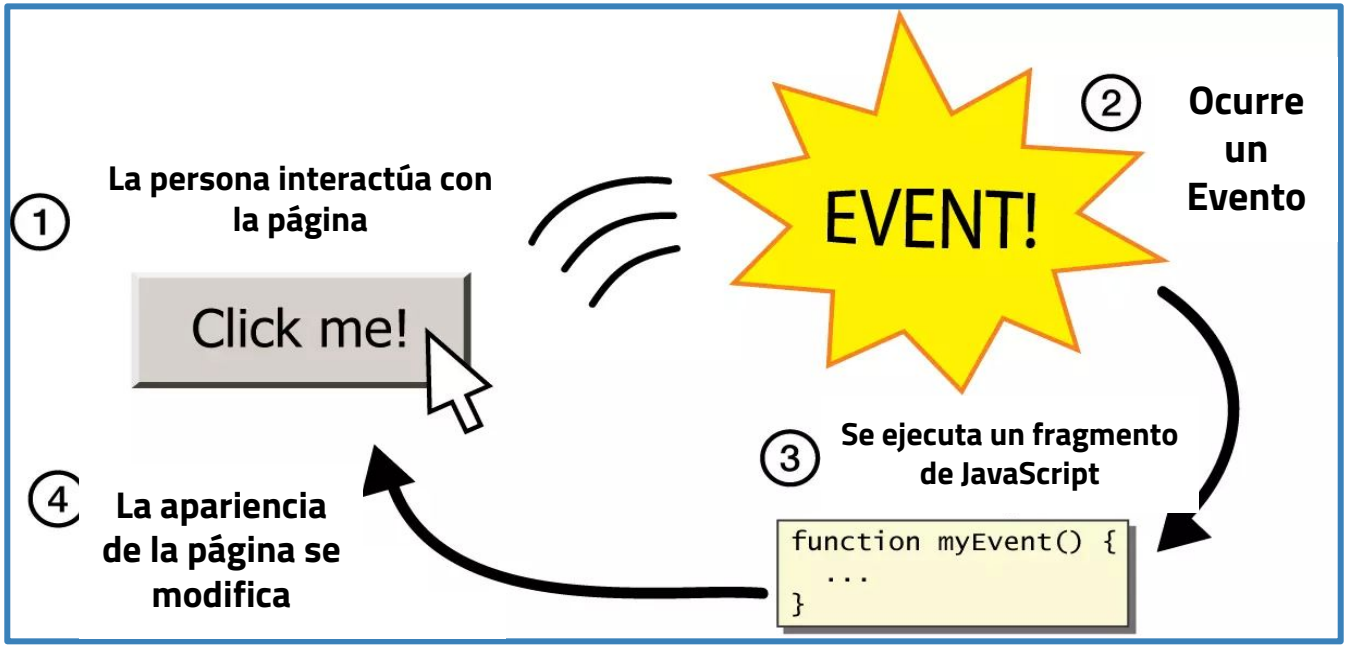
Interfaz independiente del lenguaje que trata un documento XML o HTML como una estructura de árbol



Fuente:
https://en.wikipedia.org/wiki/Document_Object_Model
Autor: Birger Eriksson

JavaScript

- Permite usar un paradigma de programación orientado a eventos (entre otros).
- Más del 97% de las páginas web lo usan para generar comportamientos interactivos.



Modificado de:https://itzone.com.vn/en/article/event-in-javascript/

Uso de herramientas de desarrollador en el navegador

[Practiquemos ...]

¿¿¿Y Python???

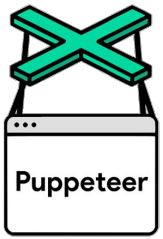
🐍 Web scraping con Python

- Python no es la única tecnología con la cual podemos hacer scraping (ej: R, JavaScript, Java, C++, etc)

R

Java

JavaScript



- Python nos puede ayudar en muchas cosas:
  - Pedidos HTTP (urllib, requests)
  - Parseo de la información (Beautiful Soup)
  - Automatización
  - Control de excepciones
  - etc

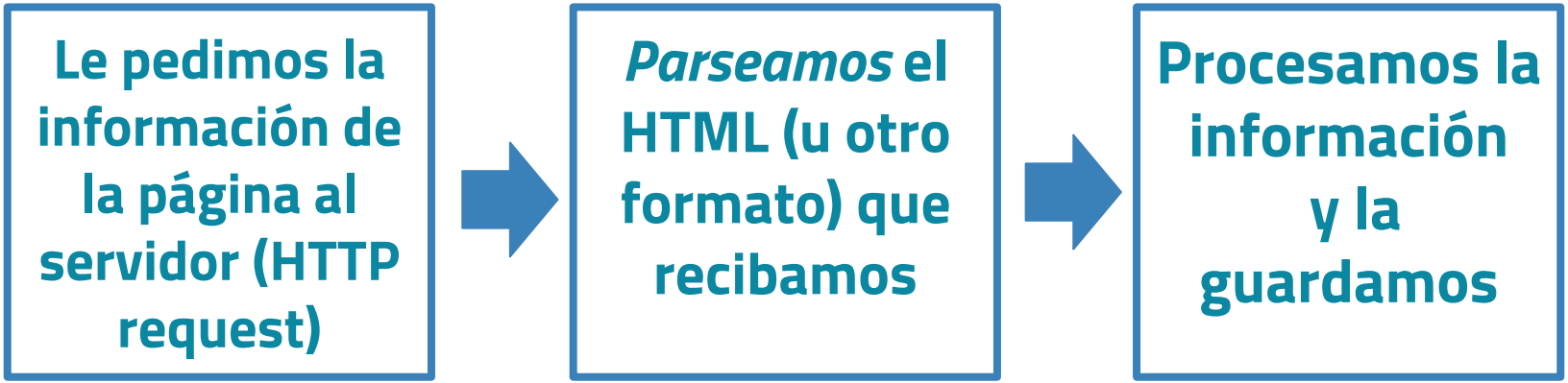
”Parsear” la información

- Dividir un texto en sus componentes y describir sus roles sintácticos.

- El “parseo” de un documento HTML es básicamente tomar código HTML y extraer información relevante como el título de la página, párrafos, encabezados, enlaces, texto en negrita, etc.

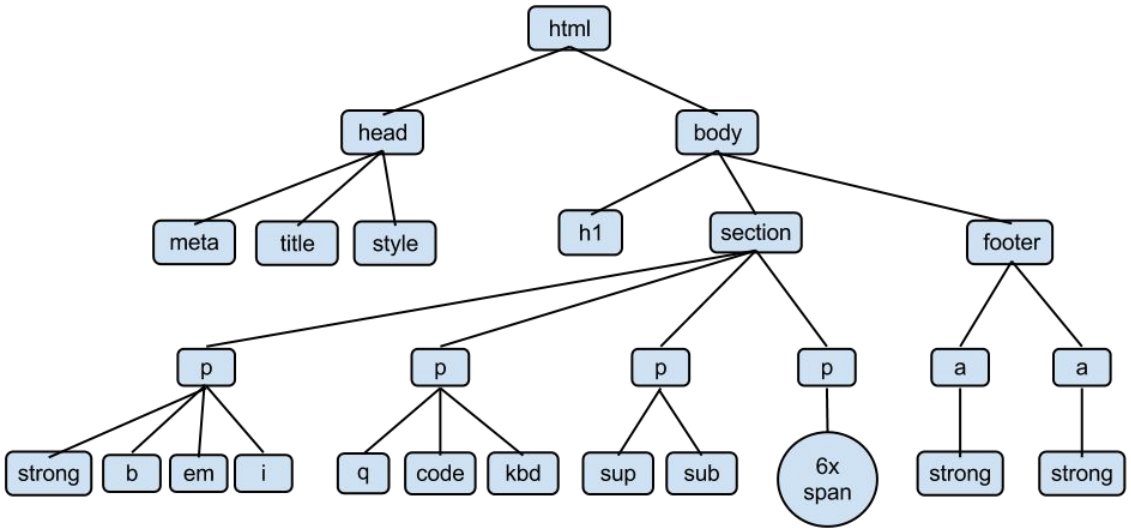
Fuente: https://stackoverflow.com/a/20421508

Flujo de trabajo en *web scraping*



Parseando con Python: Beautifulsoup

- Beautiful Soup es una librería de Python para web scraping
- Se usa para extraer los datos de archivos HTML y XML.
- Crea un árbol de análisis a partir del código fuente de la página



Fuente: https://medium.com/milooproject/python-simple-crawling-using-beautifulsoup-8247657c2de5

Ahora si: ¡Vamos a Colab!