


WEB SCRAPING

PAULO CÉSAR TUYA

ÍNDICE

- **Módulo 1: Introducción a Web Scraping**
 - Conociéndonos un poco
 - ¿Qué es Web Scraping? 
 - Conociendo las API's
 - Conceptos relacionados a Web Scraping
 - Introducción al HTML
- Módulo 2: Herramientas para el Análisis de una Página Web
- Módulo 3: Web Scraping con Scrapy
- Módulo 4: Scraping en páginas dinámicas
- Módulo 5: Despliegue de un Spider

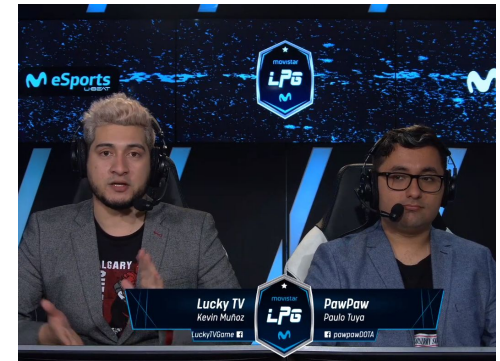


CONOCIÉNDONOS UN POCO

EL PROFESOR



- PUCP - Ingeniería Informática
- Data Science
- Predocente EE.GG.CC. y FCI
- Banco de Crédito del Perú
- Movistar Esports





LOS ALUMNOS



- Nombre
- Background
 - Estudios
 - Experiencia o actualidad laboral
- Interés por el curso



¿QUÉ ES WEB SCRAPING?



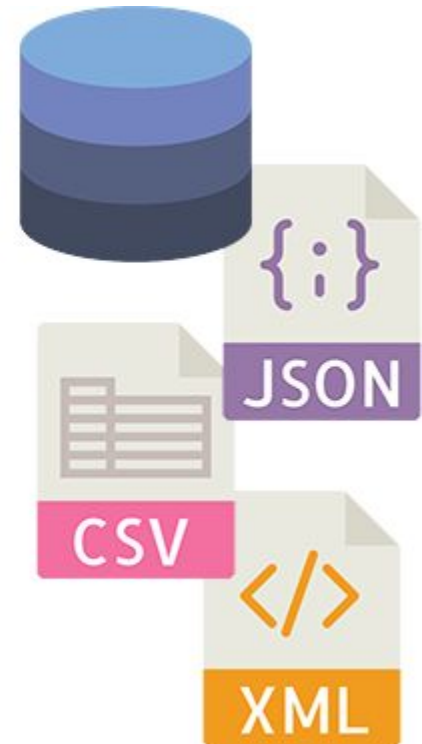
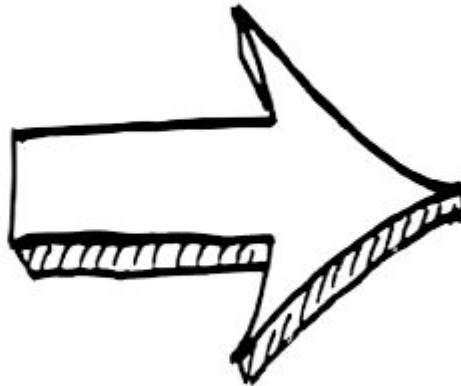
SCRAPING

- **Técnica** de extracción de datos
- Su peculiaridad es que opera sobre **contenido final**
- Técnica *ad hoc* y poco estable 
- Ejemplos de uso:
 - Interfaz con un sistema no compatible 
 - Creación de una API no oficial
 - **Extracción y minería de datos**

SCRAPING VS PARSING



WEB SCRAPING





WEB SCRAPING

- **Técnica** que **extrae** la información de un sitio web y la **transforma en data estructurada**
- Características:

Automatizado

Gran escala

vs

Manual


Página por página



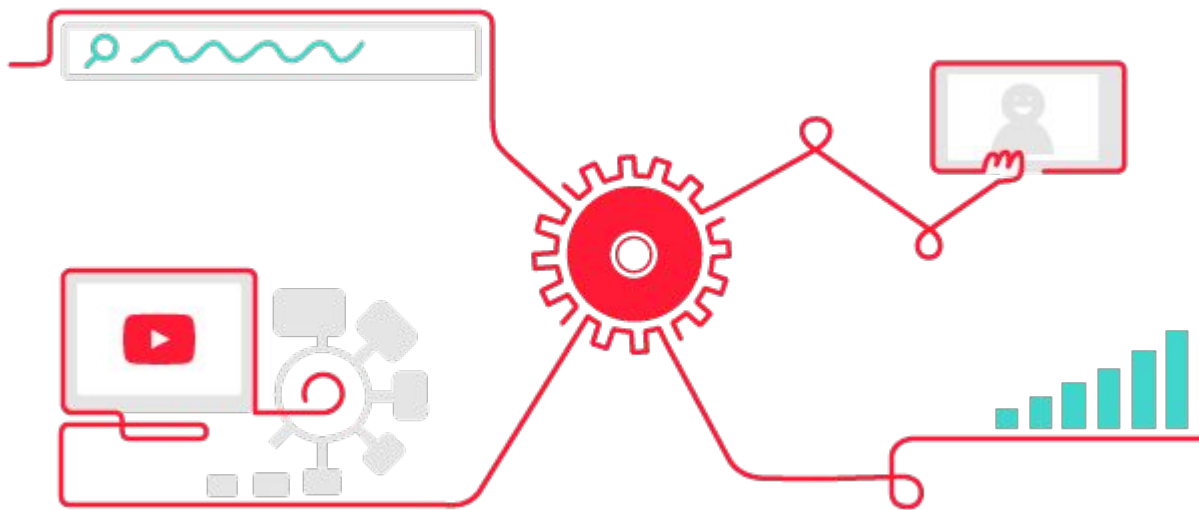
CONOCIENDO LAS APIS



API

- Application Programming Interface 
- Interfaz - Protocolo de comunicación entre un cliente y un servidor
- Permite interactuar con una aplicación y **realizar solicitudes de información**
- Muchos sitios web cuentan con una API abierta que ofrece data estructurada

API





ALGUNAS APIS CONOCIDAS

- Facebook - <https://developers.facebook.com/docs/graph-api/>
- NASA - <https://api.nasa.gov/>
- PokémonGO - <https://pokeapi.co/>
- Instagram - <https://www.instagram.com/developer/>
- YouTube - <https://developers.google.com/youtube/v3>
- Spotify - <https://developer.spotify.com/documentation/web-api/>



¿WEB SCRAPING O USAR APIS?


- Utilizar API's ofrece **múltiples ventajas**:
 - Soporte oficial
 - Documentación de uso
 - Output estructurado
- Sin embargo, **no todas las páginas** ofrecen un API público



CONCEPTOS RELACIONADOS A WEB SCRAPING



TÓPICOS DE WEB SCRAPING

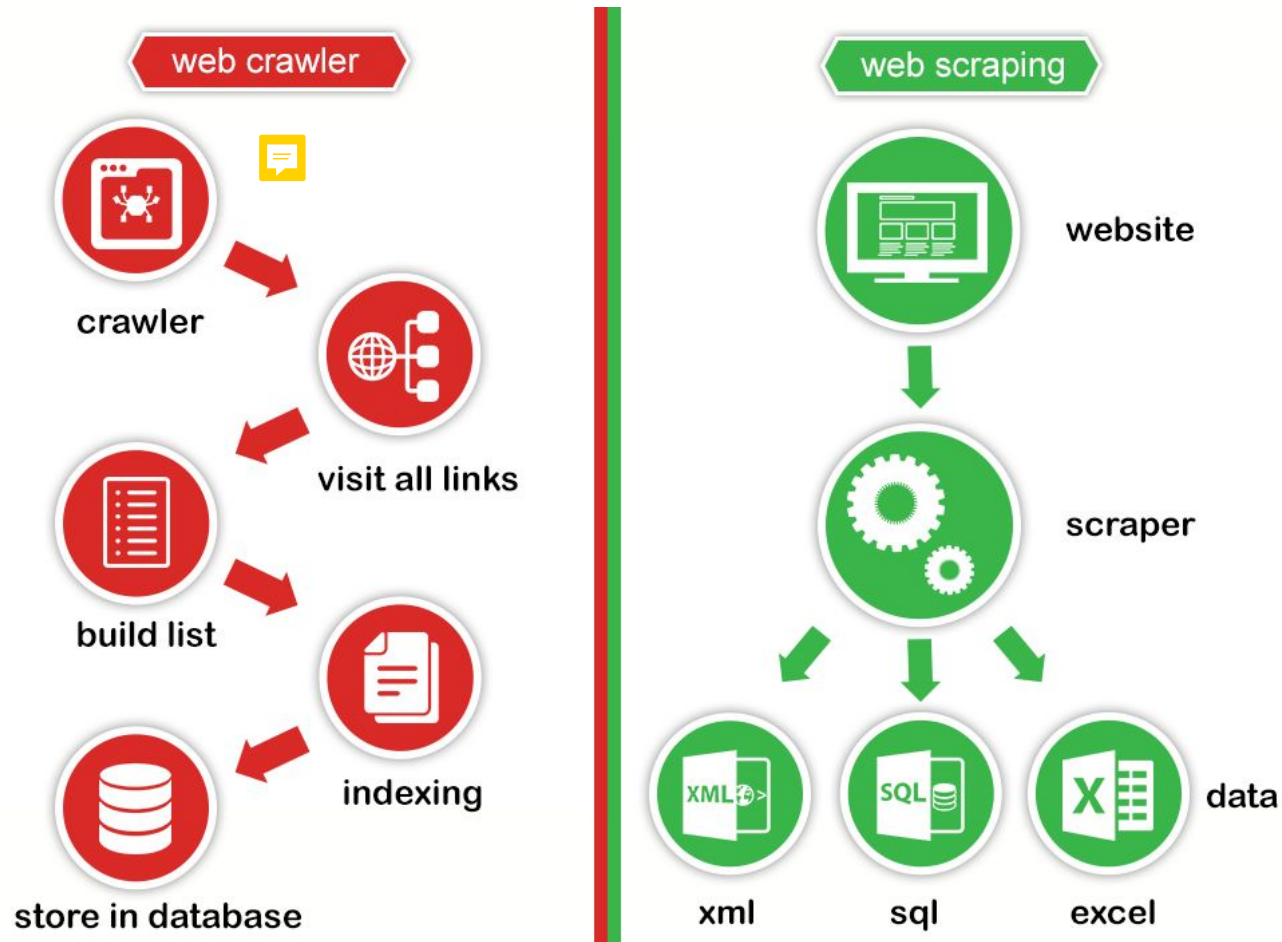
- Web Crawling 
- Hipertexto
- HTTP
- Peticiones y Respuestas HTTP
- Markup
- HTML



WEB CRAWLING

- Técnica que consiste en navegar exhaustivamente **sitios web** para extraer **hipervínculos**
- Dichos hipervínculos son almacenados y utilizados para generar un **índice** del sitio web
- Técnica fundamental para los **motores de búsqueda**
- **No** almacena **información** que se encuentre dentro del sitio web

WEB CRAWLING VS WEB SCRAPING

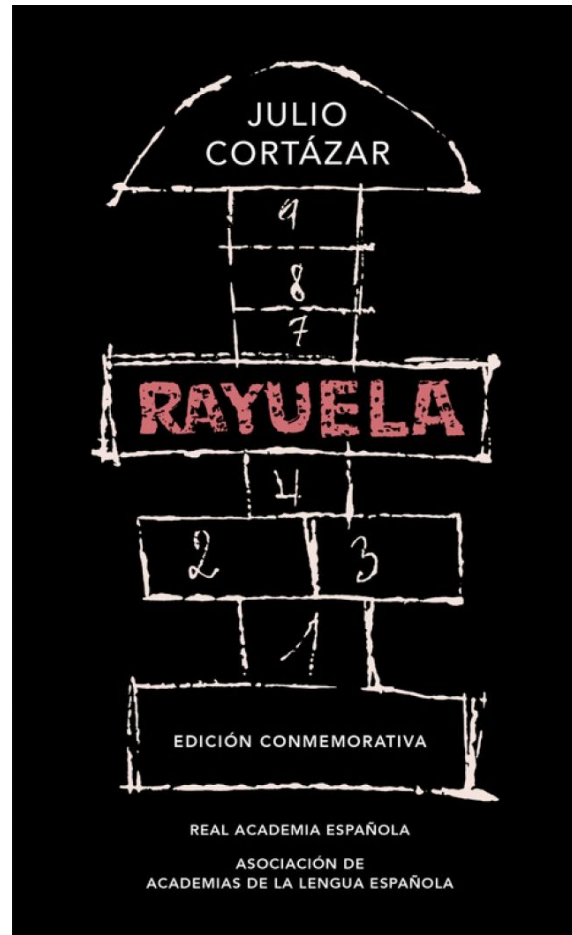




OTROS TOPICOS DE WEB SCRAPING

- Hipertexto
- HTTP
- HTTP REQUESTS
- MARKUP
- Notebooks de Python
-

HIPERTEXTO



¿PENSAMIENTO HUMANO LINEAL O HIPERTEXTUAL?



HIPERTEXTO




- Estructura de texto no secuencial
- Pensamiento humano hipertextual
- Enlaces no secuenciales entre textos
- Implementación en Informática: **hipervínculos**

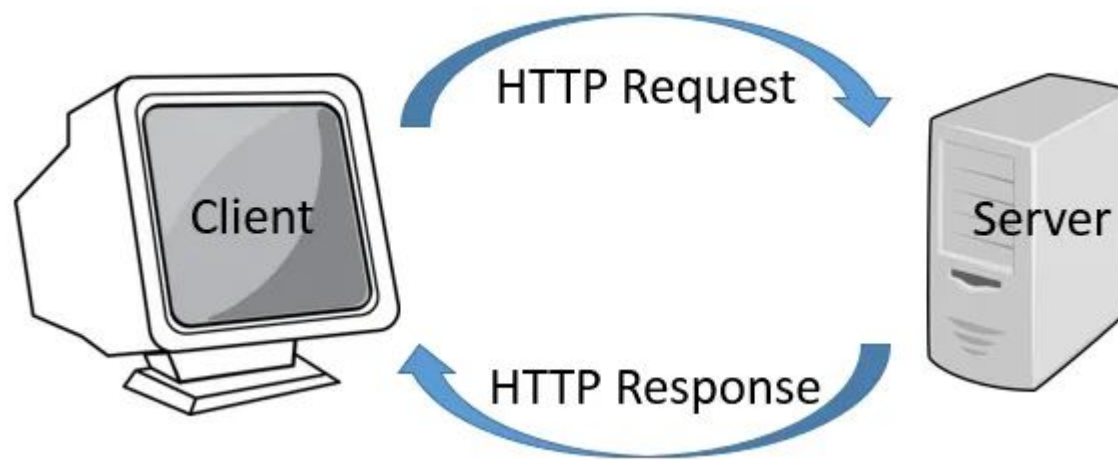




PROTOCOLO HTTP

- **H**ypertext **T**ransfer **P**rotocol 
- Protocolo de transferencia que permite la comunicación en la WWW
- Esquema petición-respuesta:
 - Posee métodos para realizar peticiones
 - Envía respuestas con un código de respuesta adjunto

PROTOCOLO HTTP





HISTORIA DEL PROTOCOLO HTTP

- 1991 - HTTP/0.9 - Solo soporta el comando GET
- 1996 - HTTP/1.0 - Permite métodos GET, HEAD, POST
- 1999 - HTTP/1.1 - Versión más usada. Pipelining
- 2000 - HTTP/1.2 - Extensión experimental de HTTP
- 2015 - HTTP/2.0 - Mejor empaquetado y transferencia
- 2019 - HTTP/2.4.39 - Última versión



MÉTODOS HTTP

- POST
- GET 
- PUT
- PATCH
- DELETE
- ...

PETICIÓN HTTP

GET /doc/test.html HTTP/1.1

Host: www.test101.com

Accept: image/gif, image/jpeg, */*

Accept-Language: en-us

Accept-Encoding: gzip, deflate

User-Agent: Mozilla/4.0

Content-Length: 35

bookId=12345&author=Tan+Ah+Teck

Request Line

Request Headers

Request
Message
Header

A blank line separates header & body

Request Message Body

RESPUESTA HTTP

HTTP/1.1 200 OK

Date: Sun, 08 Feb xxxx 01:11:12 GMT

Server: Apache/1.3.29 (Win32)

Last-Modified: Sat, 07 Feb xxxx

ETag: "0-23-4024c3a5"

Accept-Ranges: bytes

Content-Length: 35

Connection: close

Content-Type: text/html

<h1>My Home page</h1>

Status Line

Response Headers

Response
Message
Header

A blank line separates header & body

Response Message Body



MARKUP



- **Sistema** de anotación de texto
- Distingue **anotaciones** de **contenido**
- Incorpora **etiquetas** para lograr esta diferenciación
- Ejemplos:
 - XML
 - XHTML
 - SGML
 - BBC
 - **HTML**

LENGUAJE MARKUP

```
1 <receta>
2   <titulo>Arroz con Leche</titulo>
3   <ingredientes>
4     <ingrediente>Arroz</ingrediente>
5     <ingrediente>Leche</ingrediente>
6   </ingredientes>
7
8   <preparacion>
9     Sumerja el arroz crudo en la
10    leche. Deje reposar por diez
11    minutos. Sirva frio, con canela
12    y clavo de olor al gusto
13  </preparacion>
14 </receta>
```



INTRODUCCIÓN A HTML



HTML

- **H**ypertext **M**arkup **L**anguage
- Lenguaje markup usado para la construcción de **páginas web**
- Separación texto-estructura-diseño
- HTML se centra en texto y estructura
- Estructura de **árbol**

Código HTML vs Output

```
1 <!-- Esto es un comentario -->
2 <h3>Bienvenidos al curso <b>Web Scraping en
  Python</b></h3>
3 <p><strong>Este curso les permitirá adquirir los
  conocimientos necesarios para usar las herramientas de
  Web Scraping disponibles en Python</strong></p>
4 <p>El profesor les guiará a lo largo de cada una de las
  sesiones con ejemplos, actividades, conceptualizaciones
  y explicaciones. Recuerden estar muy atentos a la clase
  y preguntar en caso tengan alguna duda.</p>
5 
```

Bienvenidos al curso **Web Scraping en Python**

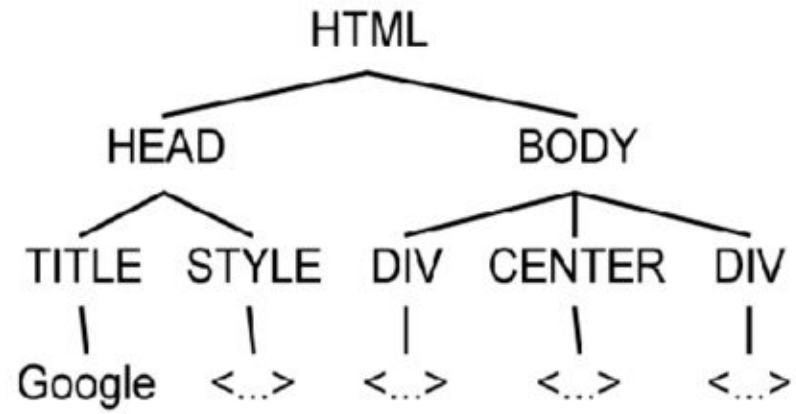
Este curso les permitirá adquirir los conocimientos necesarios para usar las herramientas de Web Scraping disponibles en Python

El profesor les guiará a lo largo de cada una de las sesiones con ejemplos, actividades, conceptualizaciones y explicaciones. Recuerden estar muy atentos a la clase y preguntar en caso tengan alguna duda.



Estructura de árbol

```
<html>
  <head>
    <title>Google</title>
    <style>...</style>
  </head>
  <body>
    <div>...</div>
    <center>...</center>
    <div>...</div>
  </body>
</html>
```

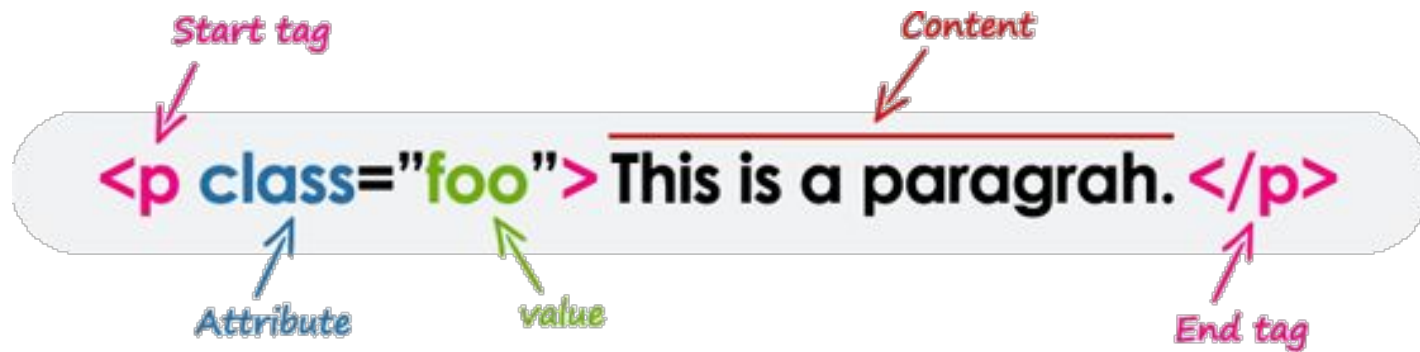




ETIQUETAS HTML

- Etiqueta que abarca un pedazo de **texto**
- Etiqueta HTML + texto interior = Elemento HTML
- En un árbol: nodo HTML = elemento HTML

ETIQUETAS HTML



NUESTRO AMBIENTE DE TRABAJO

- Lenguaje de programación **Python**
- Entorno de desarrollo **Google Colab**





ACTIVIDAD MÓDULO 1

**GRACIAS
POR SU ATENCIÓN**