# Web Scraping

## Trabajando con Beautiful Soup

Beautiful Soup (BS) es una librería de Python que permite hacer todo tipo de escrapeos

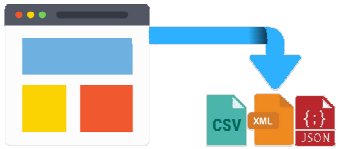
BS permite **“parsear”** las etiquetas HTML y cargarlas en Python como un objeto vivo, capaz de ser recorrido, indagado y hasta modificado.

Se llama DOM al objeto HTML con todas sus estructuras de etiquetas y propiedades


```
>> pip install beautifulsoup4

Luego en el código: import bs4
```

# Práctica
## Web Scraping con Beautiful Soup




Beautiful Soup

-Instalación de Beautiful Soup
-Descargando y parseando HTML
-Navegando el DOM de un documento HTML
-Métodos de búsqueda de etiquetas
-Práctica de escraping

# Protocolos

## XML

**Extensible Markup Language** (Lenguaje de Marcado Extensible) y es una especificación de W3C como lenguaje de marcado de propósito general, muy similar a HTML.

A diferencia de otros lenguajes de marcado, XML no está predefinido, por lo que debes definir tus propias etiquetas. El propósito principal del lenguaje es compartir datos a través de diferentes sistemas, como Internet.


```
<?xml version="1.0" encoding="UTF-8"?>
<biblioteca>
    <libro>
        <titulo>La vida está en otra parte</titulo>
        <autor>Milan Kundera</autor>
        <fechaPublicacion año="1973"/>
    </libro>
    <libro>
        <titulo>Pantaleón y las visitadoras</titulo>
        <autor fechaNacimiento="28/03/1936">Mario Vargas Llosa</autor>
        <fechaPublicacion año="1973"/>
    </libro>
    <libro>
        <titulo>Conversación en la catedral</titulo>
        <autor fechaNacimiento="28/03/1936">Mario Vargas Llosa</autor>
        <fechaPublicacion año="1969"/>
    </libro>
</biblioteca>
```

# Práctica
## Parseando XML con Beautiful Soup


Beautiful Soup

-Identificación del documento XML
-Parseo de un archivo XML
-Traspasando datos XML a un archivo de tablas