Informática y comunicaciones
ies azarquiel

CURSO ESPECIALIZACIÓN INTELIGENCIA ARTIFICIAL Y BIG DATA

Tema 4: WEBSCRAPING

Sebastián Rubio

SEPTIEMBR 2025

Tiempo ()

1 Introducción

Actualmente en internet hay alrededor 1.4 millones de terabytes de información. Lo interesante es extraer datos de la web, lo que supone un reto ya que hacerlo manualmente es practicamente imposible. La opción que nos queda es hacerlo de forma automatizada mediante programación. No todas las páginas web nos proporcionan métodos para la descarga de datos (API), por lo que tendremos que hacer uso de webscrping. La programación webscrping esta fuertemente ligada a la estructura de la página, por lo que se hace necesario tener conocimientos de html y css.

Un inconveniente del webscrping viene dado por los cambios que una página web sufre durante su vida útil, es decir no siempre va a tener la misma estructura, por lo que nuestro programa para extraer información de la página no funcionará correctamente, teniendo que reprogrmarlo para obtener los datos que queremos. Además, algunas páginas web están pendientes de si hay software haciendo scraping produciendose de esta forma baneos temporales.

Ejemplo

Vamos a inspeccionar la página web de los idiomas de wikipedia

1.1 Tipos de webscrping

Podemos distinguire tres tipos diferentes de webscrping:

- De una sóla página webscrping o estático. Utilizaremos las librerias *requests*,*lxml* y *scrapy*

- Estático de varias páginas del mismo dominio, usaremos *scrapy*
- WebScraping dinámico, utilizaremos la herramienta de *Selenium*
- WebScraping utilizando API'S

Para hacer webscraping tenemos que seguir una serie de pasos:

1. Tenemos que definir una semilla, hace referencia al html desde el que iniciamos el webscraping
2. Hay que hacer una petición al correspondiente servidor.
3. Como resultado de la solicitud anterior, vamos a tener la página web correspondiente.
4. Extraer localmente la información de mi interés.
5. Volver ha realizar el paso dos con otras URL's del mismo dominio

1.2 XPATH

Me permite extraer datos de una página web mediante la creación de expresiones que recorren y procesan documentos XML. Hay que aclarar que un documento HTML es un tipo de documento XML.
Para empezar vamos a utilizar las siguiente página web *https://xpather.com/*, que nos va a permitir comprobar nuestras expresiones.
La búsqueda se puede realizar partiendo de la raiz "/" o bien directamente en cualquier parte del documento "//" Veamos como ejemplo algunas expresiones xpath:

- **.//div[@id="hero"]**:Localiza en el raiz del documento, etiquetas de la forma .
  De la misma forma podemos utilizar no es igual $!=, <, >$.
- En la búsqueda podemos utilizar expresiones lógica **//div[@id='lista-cursos' and @class='container']**

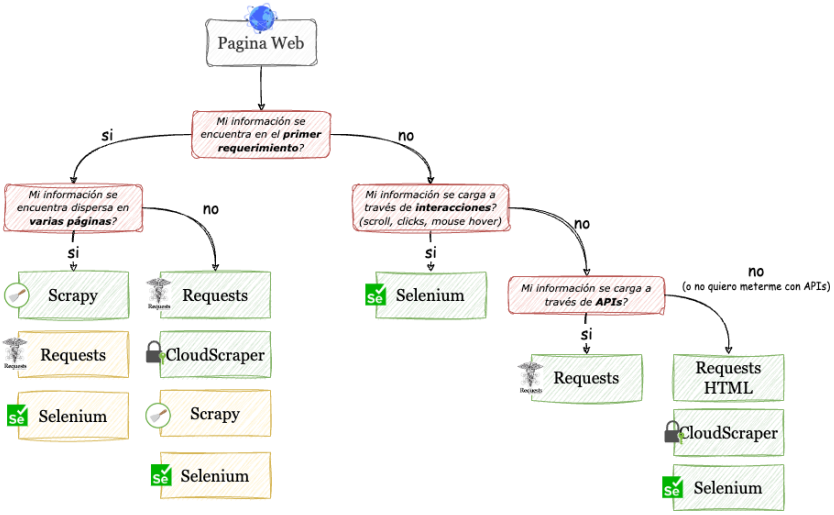
- **//div[@id='lista-cursos' and @class='container']//div[@class='row'][2]** , busca por la posición
- **//div[@id='lista-cursos' and @class='container']//div[@class='row'][last()]**, el último
- **//div[@id='lista-cursos' and @class='container']//div[@class='row'][position()=2]**
- **//div[@id='lista-cursos' and @class='container']//div[contains(@class,"ro")]**, subcadena
- **//div[@id='lista-cursos' and @class='container']//div[starts-with(@class,"r")]**,comienz por
- **//div[@id='lista-cursos' and @class='container']//div[ends-with(@class,"w")]**, finaliza por
- **//div[@id='lista-cursos' and @class='container']//div[not(ends-with(@class,"w"))]**, negación
- **//div[@id='lista-cursos' and @class='container']//div[@class='row'][1][text()]**,devuelv texto
- **//div[@id='lista-cursos' and @class='container']//text()**, obtengo sólo el texto
- **//div[@id='lista-cursos' and @class='container']//@class**, valor del atributo class

Por útlimo podemos utilizar la consola del navegador para realizar búsquedas con expresiones xpath.

Ejercicio 1

El apartado correspondiented del aula virtual hay un fichero index.html, del que hay que extraer la siguiente información:
El valor del gasto realizado
El nombre del gasto
Obtener sólo el segundo gasto (el nombre)
Obtener sólo el precio del gasto "Compra"
Obtener el valor del atributo del ejercicio anterior

En la siguiente página web tienes una sintaxis completa de xpath:https://devhints.io/xpath



2 Librerias para Python

Para hacer WebScraping con el lenguaje Python vamos a necesitar algunas librerías. Lo primero que hacer es crearnos un nuevo entorno virtual y en el vamos a instalar las siguientes librerias.
Para instalar webScraping vamos a usar conda, si lo vamos a instalar con el gestor de paquetes de python(pip), hay que ver las instrucciones de instalación https://docs.scrapy.org/en/latest/intro/install.html

- requests= pip install requests
- lxml= pip install lxml
- BeatifullSoup= pip install bs4
- Scrapy = conda install -c conda-forge scrapy, la web https://www.scrapy.org/
- itemloaders
- Selenium
- webdriver-manager