WEB SCRAPING

Juliana Maritza Zarate Jimenez

Pontificia Universidad Javeriana sede Cali

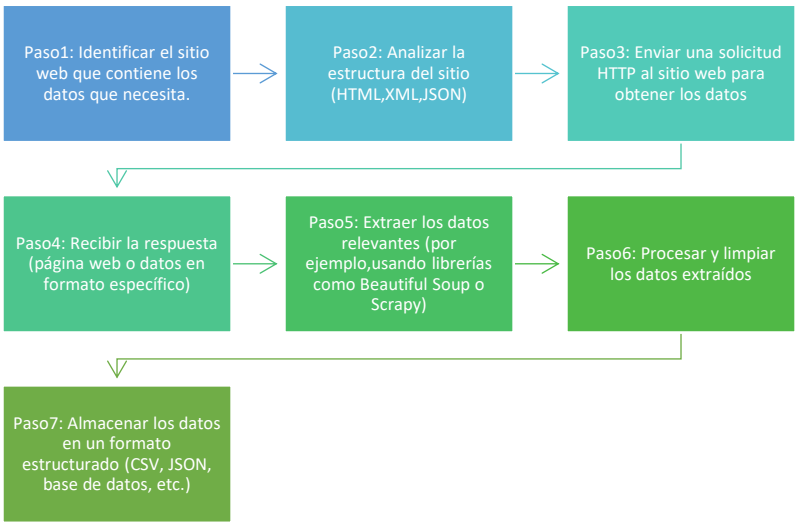
El **Web Scraping** es una técnica utilizada para extraer de manera automatizada grandes volúmenes de datos de sitios web. Esta práctica implica la utilización de herramientas que recorren sitios web (bots o crawlers) y descargan el contenido en varios formatos, como HTML, XML, JSON o archivos multimedia. Los datos obtenidos se procesan para convertirlos en información estructurada y utilizable, que puede ser almacenada en bases de datos o sistemas de archivos. El objetivo del web scraping es analizar estos datos posteriormente, ya sea para estudios de investigación, análisis de mercado, o como apoyo en la toma de decisiones empresariales.

¿Qué herramientas o tecnologías se necesitan para realizar Web Scraping?

Existen múltiples herramientas y tecnologías utilizadas en la práctica del web scraping, entre las que destacan:

- **Librerías de programación**:
  - Beautiful Soup: Una biblioteca de Python que facilita la extracción de datos desde documentos HTML o XML.
  - Scrapy: Un framework de Python diseñado para proyectos de scraping a gran escala.
  - Selenium: Herramienta que permite automatizar la interacción con un navegador web, emulando el comportamiento de un usuario.
  - Urllib2: Un módulo en Python que maneja solicitudes HTTP.
- **Herramientas con interfaz gráfica**:
  - Import.io: Permite realizar web scraping sin necesidad de escribir código, ofreciendo una interfaz gráfica para seleccionar y extraer datos de sitios web.

Esquema de los pasos que debe seguir un Web Scraper para obtener los datos de un sitio:



¿Creen que el Web Scraping es una práctica ilegal? ¿Por qué?

El Web Scraping no es inherentemente ilegal, pero existen ciertas situaciones en las que su uso puede estar en conflicto con la ley o con normas éticas. La legalidad depende principalmente de las condiciones establecidas en los términos de uso de los sitios web, la propiedad de los datos y el propósito del scraping.

- Legalidad: No existe una legislación específica que prohíba el web scraping, pero puede haber implicaciones legales si el scraping infringe los derechos de autor del contenido, viola los términos de uso del sitio web o accede a datos protegidos de manera no autorizada. Algunas leyes como la Ley de Fraude y Abuso Informático (CFAA) en EE.UU. podrían ser aplicables si el scraping causa daños a los servidores o a los sistemas.
- Ética: Es importante considerar la ética de esta práctica, ya que puede comprometer la privacidad de los usuarios o causar perjuicios a las empresas si se abusa de la extracción de datos.

¿En qué situaciones usarían Web Scraping?

El web scraping puede ser útil en diversos escenarios, como:

- **Investigación académica**: Para recopilar grandes volúmenes de datos públicos en línea, como publicaciones en redes sociales, blogs o sitios de noticias.
- **Análisis de mercado**: Para monitorear los precios de productos en varias tiendas en línea y analizar las tendencias del mercado en tiempo real.
- **Monitorización de datos meteorológicos**: Obtener datos de cambios climáticos de diversos portales y sistemas.
- **Recopilación de reseñas de productos**: Para analizar la opinión de los consumidores sobre productos específicos en sitios de reseñas o tiendas en línea.