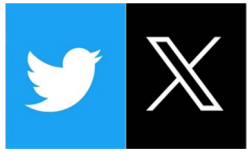
MCD | MAESTRÍA EN CIENCIA DE LOS DATOS



amazon

Web Scraping

SHEIN

mercado libre

Gustavo Mendoza Navarro
Roberto Carlos Ramírez Ramos
Reyna Viviana Jaramillo Flores

¿Qué es el Web Scraping?

Es un método empleado para extraer datos de páginas web. El scraping significa literalmente “raspado”. Aunque en el contexto al que nos referimos esté relacionado con la filtración de datos y limpieza. Una vez que encuentra datos ocultos hace que resulten útiles para que les puedas sacar provecho.

El mismo se dedica a leer información, extrayendo, almacenando y usando para su beneficio. Esta técnica de scraping cuenta con 3 fases:

1. Descarga de datos.
2. Análisis de la información.
3. Almacenamiento,

Casos de aplicación del web scraping

El *web scraping* puede tener aplicaciones muy diversas. Además de para la **indexación de buscadores**, el *web scraping* también puede usarse con los siguientes fines, entre muchos otros:

- Crear bases de datos de contactos
- Controlar y comparar ofertas *online*
- Reunir datos de diversas fuentes *online*
- Observar la evolución de la presencia y la reputación *online*
- Reunir datos financieros, meteorológicos o de otro tipo
- Observar cambios en el contenido de páginas web
- Reunir datos con fines de investigación
- Realizar exploraciones de datos o data mining

Ventajas del Web Scraping

- Extracción automatizada de datos
- Rapidez
- Información precisa
- Contar con mayor información referente a la competencia o temas de interés.

Limitaciones del Web Scraping

Para los operadores de las páginas web suele ser **ventajoso limitar las posibilidades de *scraping* automático en su contenido *online.*** Por un lado, porque el acceso masivo a la web que realizan los *scrapers* puede perjudicar el rendimiento del sitio y, por otro, porque suele haber secciones internas de la web que no deberían mostrarse en los resultados de búsqueda.

Para limitar el acceso a los *scrapers*, se ha extendido el uso del estándar robots.txt. Se trata de un archivo de texto que los operadores web ubican en el directorio principal de la página web. En él hay entradas especiales que establecen **qué *scrapers* o *bots* están autorizados a acceder a qué áreas de la web**. Las entradas del archivo *robots.txt* siempre se aplican a un dominio entero.

Antes de hacer Web Scraping

Es importante responder las siguientes preguntas antes de comenzar

¿Esto es legal en México o en el país donde me encuentro?

¿Es necesaria esta técnica o se puede resolver de alguna otra forma?

¿Puedo afectar al servidor?

API vs Web Scraping

MCD | MAESTRÍA EN CIENCIA DE LOS DATOS

API

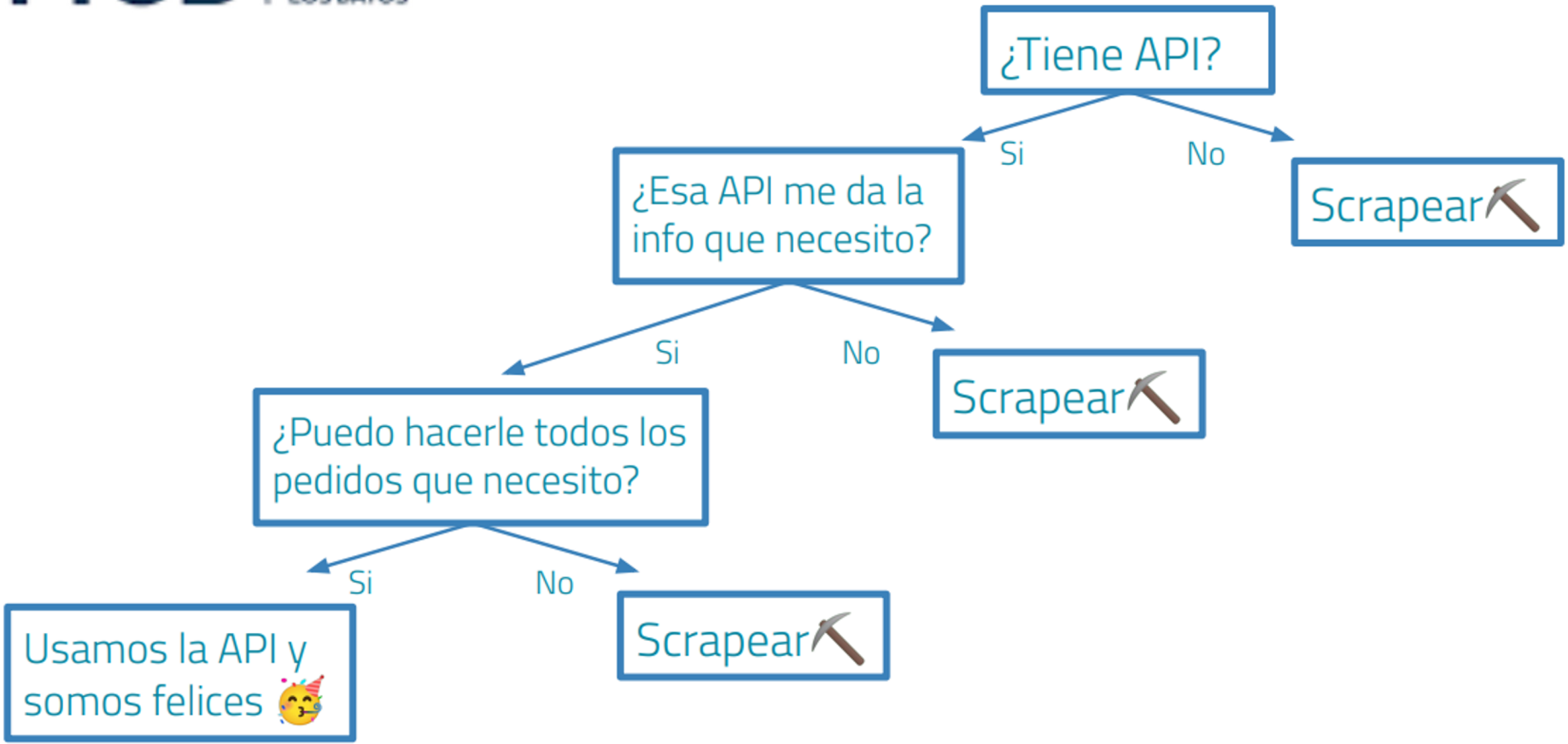
Proveer acceso a los datos de una aplicación, sistema operativo u otro servicio.

Web Scraping

Extraer información de un sitio web usando un programa informático.

Acceder a los datos del sitio web

MCD | MAESTRÍA EN CIENCIA DE LOS DATOS



¿Cómo puedo hacer Web Scraping?

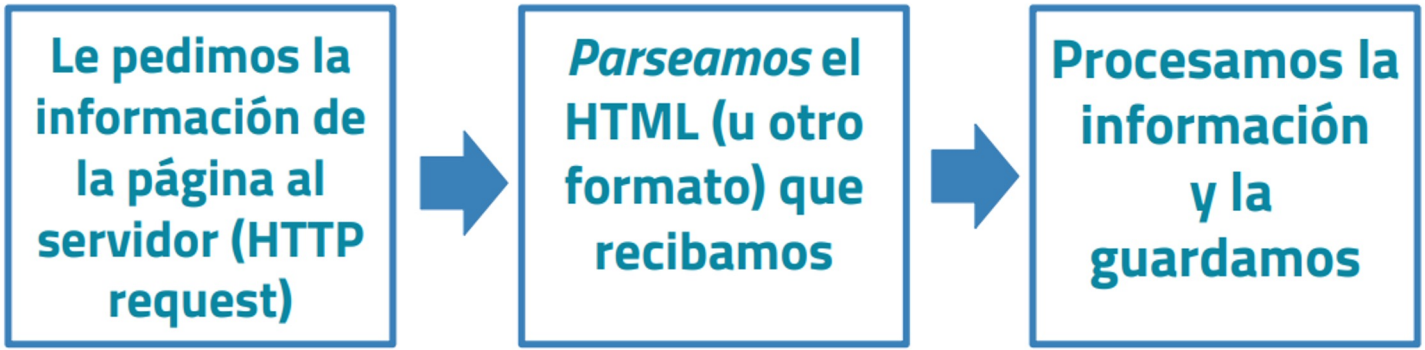
Python nos puede ayudar en:

- Pedidos HTTP (urllib, requests)
- Parseo de la información (Beautiful Soup)
- Automatización
- Control de excepciones

¿Qué es “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.

Flujo de trabajo del Web Scraping



Herramientas de Scraping para Python

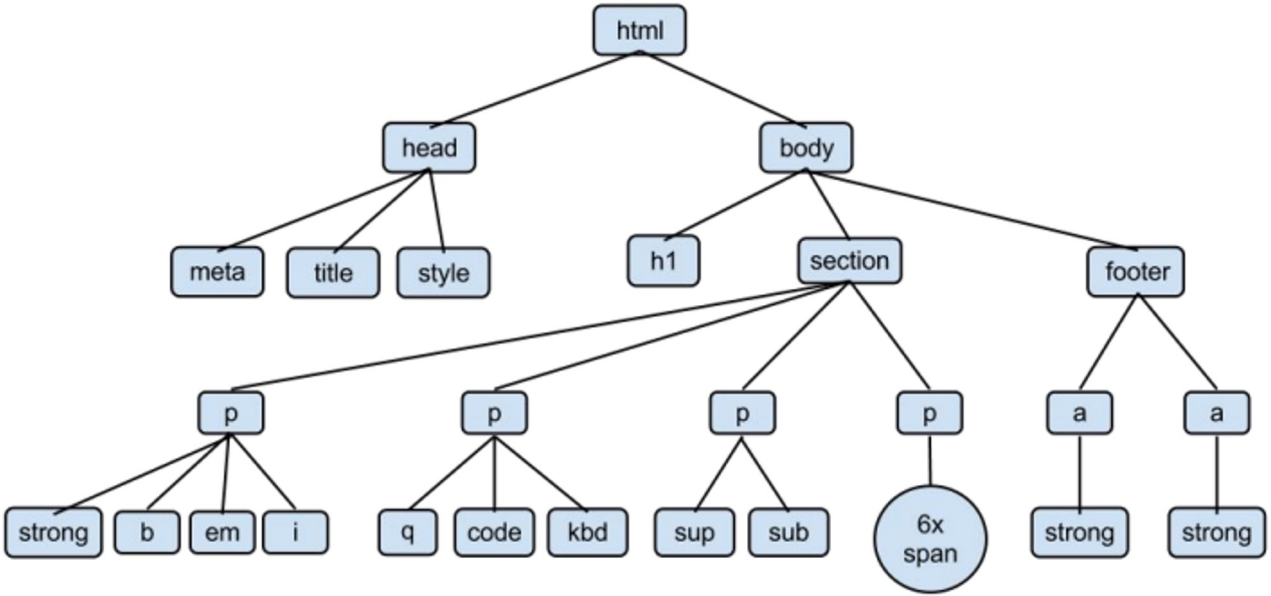
MCD | MAESTRÍA EN CIENCIA DE LOS DATOS

- **Scrapy**
- **Selenium**
- **BeautifulSoup**

	Scrapy	Selenium	BeautifulSoup
Facilidad de aprendizaje y manejo	++	+	+++
Lectura de contenidos dinámicos	++	+++	+
Realización de aplicaciones complejas	+++	+	++
Robustez frente a fallos HTML	++	+	+++
Optimización del rendimiento del scraping	+++	+	+
Calidad del ecosistema	+++	+	++

¿Cómo funciona Beautifulsoup?

- Se usa para extraer los datos de archivos HTML y XML.
  - Crea un árbol de análisis a partir del código fuente de la página



Bibliografía

https://www.xenonfactory.es/blog/que-es-el-scraping-y-como-utilizarlo-en-mi-web/

file:///C:/Users/focus/Downloads/web_scraping_freeCodeCamp.pdf

https://www.ionos.es/digitalguide/paginas-web/desarrollo-web/web-scraping-con-python/