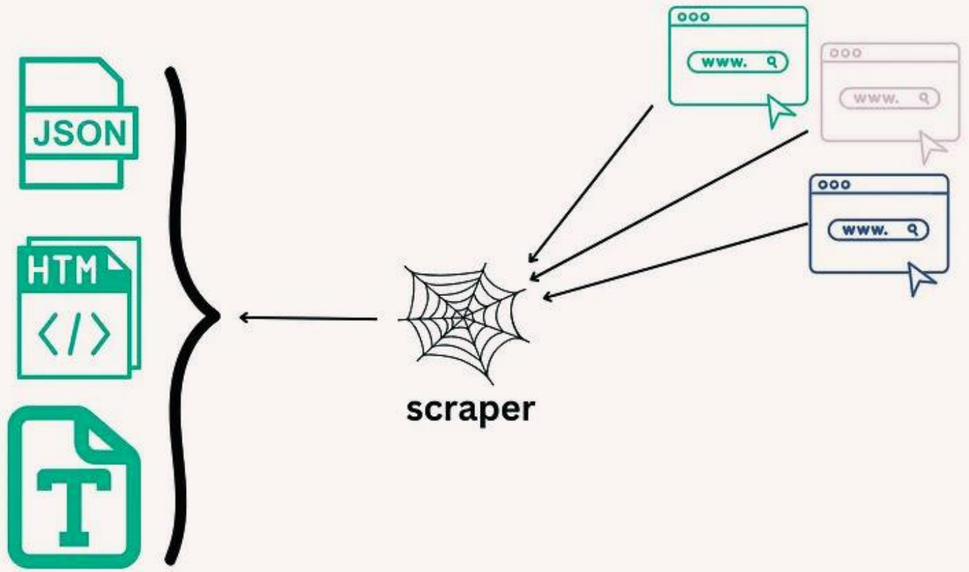
Web Scraping

Programación de Aplicaciones Web Orientadas a Objetos

HERNÁNDEZ SÁNCHEZ SARA ABIGAIL

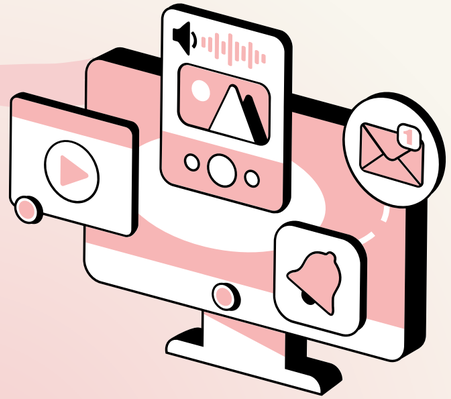
Definición

Web Scraping es el proceso automatizado de extraer información de sitios web. Utiliza programas o scripts que simulan la navegación de un usuario para recolectar datos estructurados desde páginas web.

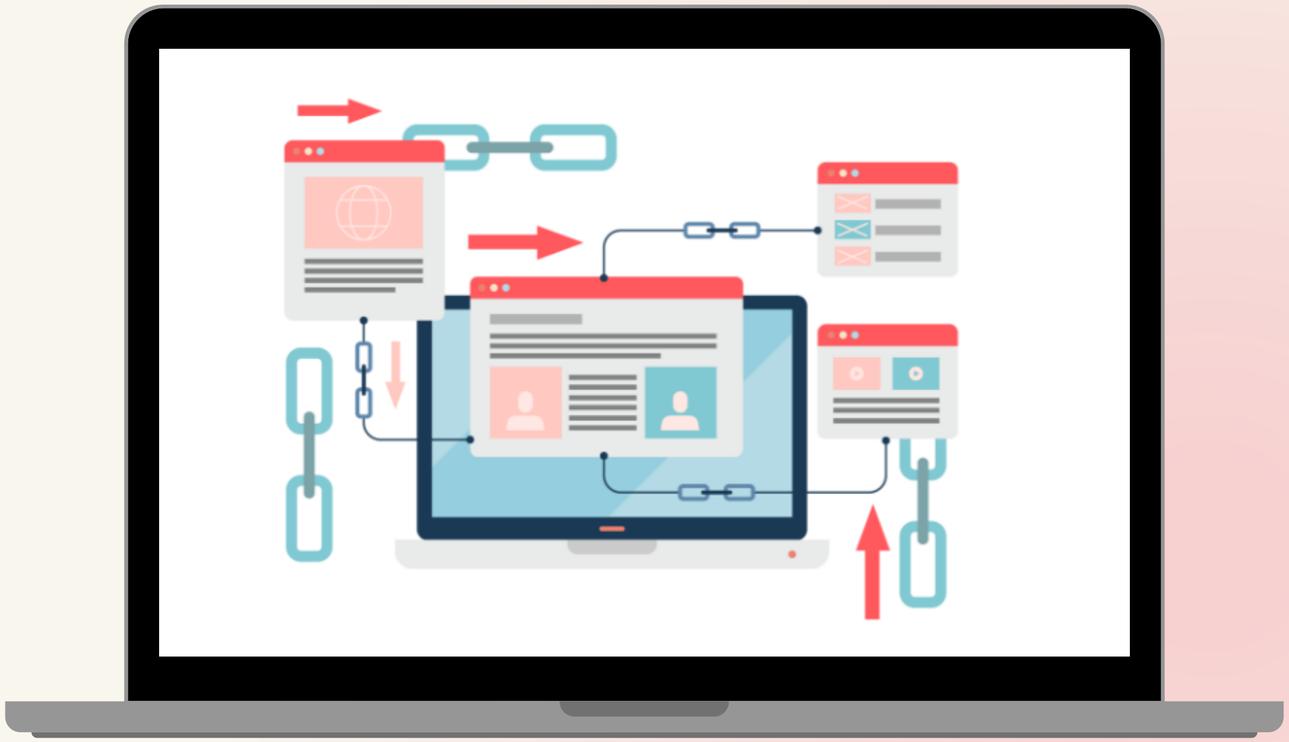


Ejemplos DE USO

- Análisis de mercado
- Monitorización de precios
- Análisis de tendencias en redes sociales
- Estudios académicos sobre información publicada
- Estrategias de comercio electrónico



Fases del proceso



1. Enviar solicitud HTTP
2. Descargar HTML
3. Analizar HTML (parseo)
4. Extraer y guardar datos

Scrapers y crawlers

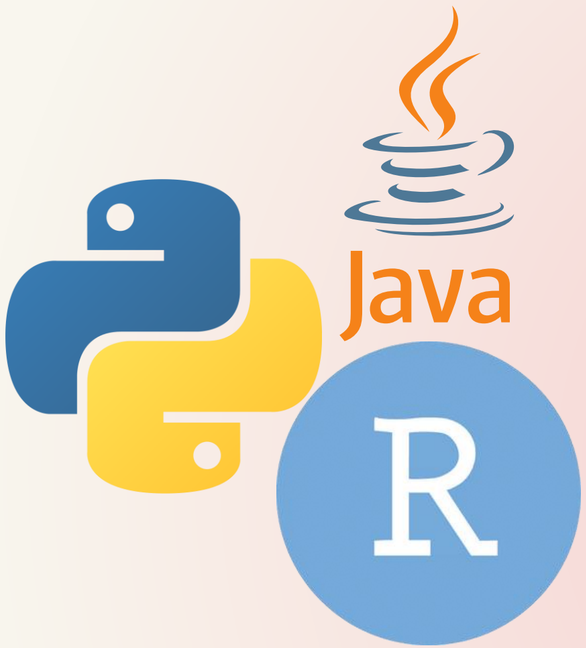
Crawlers (arañas) son programas básicos que navegan por la web buscando e indexando contenidos.

Los rastreadores suelen estar disponibles como herramientas preconstruidas que permiten especificar un determinado sitio web o término de búsqueda.

Los **scrapers** hacen el trabajo sucio de extraer rápidamente la información relevante de los sitios web.

Por ejemplo, puedes dar a tu web scraper una expresión regular que especifique el nombre de una marca o una palabra clave.

Herramientas



Python:

- requests (descarga de HTML)
- BeautifulSoup (parseo)
- Selenium (navegación automática)
- Scrapy, Puppeteer, Octoparse (visual)

JavaScript:

- Puppeteer
- Cheerio

R

- rvest
- httr
- xml2
- polite

Ventajas

- Ahorro de tiempo:
- Aumenta la precisión de los datos
- Más datos a escala
- Rentable
- Flexible
- Frescura de datos
- Diversas fuentes de datos

Desventajas

- Cuestiones legales
- Limitaciones técnicas
- Problemas de rendimiento
- Mantenimiento y actualización
- Costo

Apectos legales

- ¿Es legal hacer web scraping?

Depende del uso, el contenido y los términos del sitio

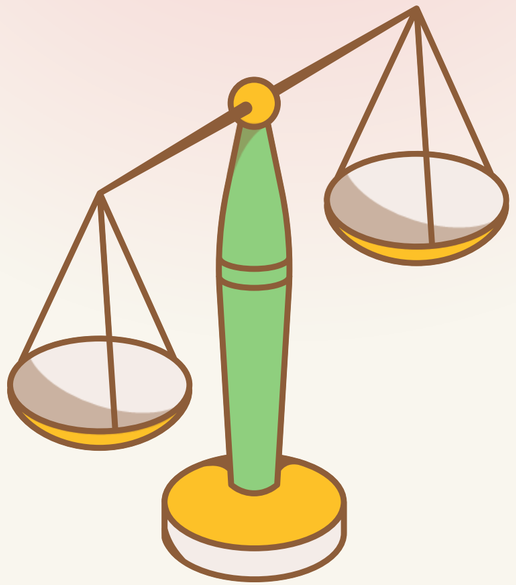
- Importancia de:

- Revisar robots.txt
-Respetar términos de uso
-No recolectar datos personales sin consentimiento

Caso

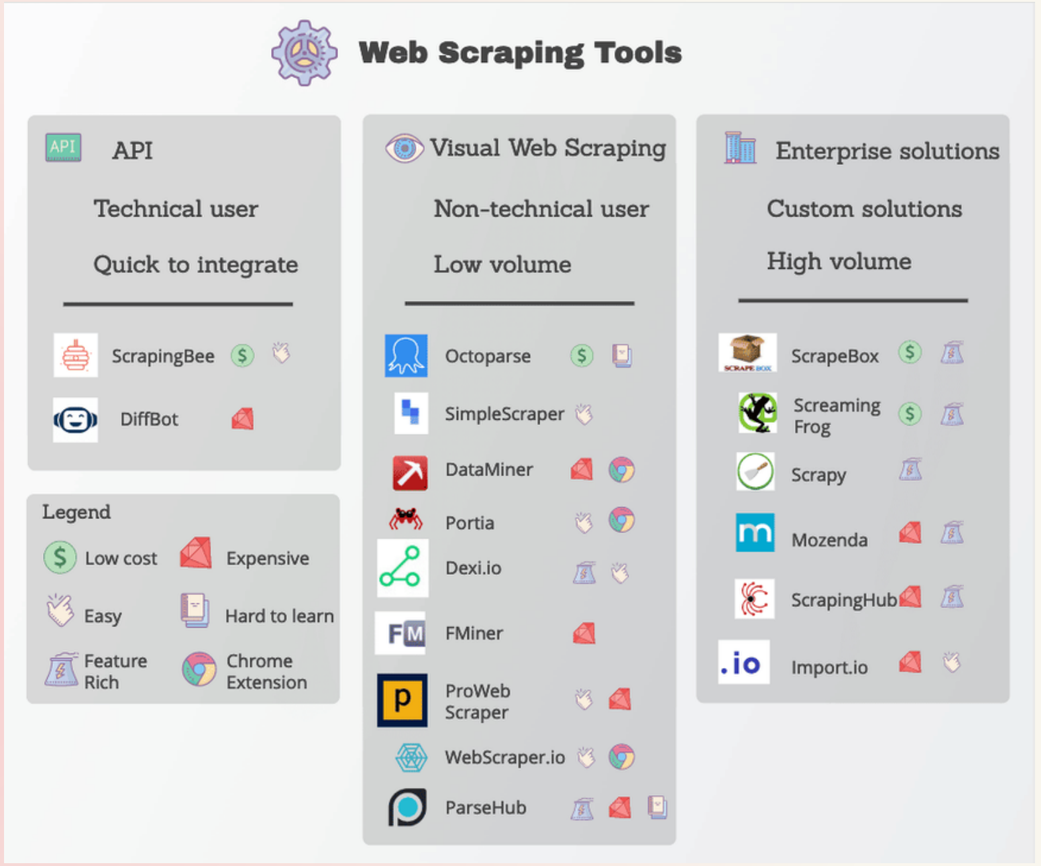
LinkedIn vs hiQ Labs

hiQ es una pequeña empresa de análisis de datos que utilizaba bots automatizados, web scraping, para obtener información de perfiles públicos de LinkedIn.



Buenas prácticas

- Hacer peticiones respetuosas (delays, headers)
- Usar APIs oficiales si están disponibles
- Evitar scraping masivo o malicioso
- No redistribuir datos protegidos



Ejemplo Python

Intro to Web Scraping: Build Your First Scraper in 5 Minutes

A quick guide to help you build a simple web scraper.

Joe Osborne · Follow
5 min read · Apr 7, 2024

Example Domain

This domain is for use in illustrative examples in documents. You may use this domain in literature without prior coordination or asking for permission.

More information...

Conclusión

- WEB SCRAPING ES UNA HERRAMIENTA PODEROSA
- ES ÚTIL, PERO CONLLEVA RESPONSABILIDADES LEGALES Y ÉTICAS
- SIEMPRE QUE SE USE DE MANERA RESPETUOSA, ES UNA GRAN ALIADA PARA LA CIENCIA DE DATOS, LA INVESTIGACIÓN Y LA AUTOMATIZACIÓN

Gracias