Creado por:

Isabel Maniega

Qué es Web Scraping

- Tambien Llamado Scrapeo Web, raspado Web
- Permite extraer datos de la web de forma automatizada
- Se puede usar para obtener mucha información o cosas muy concretas
- Web crawling proceso de extraer los enlaces, y web scraping sirve para extraer los datos
- Lo usan los buscadores, por ejemplo google tienen Bots que hace todo esto.

Campos de uso:

- Marketing Digital, SEO, etc
- Análisis de precios
- Ofertas
- Nuevos productos
- Comprobar la opinión de clientes en Twitter

Requisitos:

- HTML entre otros (no es como tal lenguaje de programación
- entre otras cosas

Legalidad:

- Problemáticas, a veces en caso en los que hay competencia desleal
- Condiciones legales del sitio web
- En algunos casos se contacta a la empresa para ver si permiten screarles

Encontradas en:

- Condiciones de uso
- Aviso legal
- Si hay muchas peticiones, podría colapsar la web de la cual queremos obtener datos
- En estadísticas de la web, puede haber falseo de datos.
- Estos bots se les suele llamar arañas

Muchas empresas se protegen contra el scraping, (la mayoría se protegen):

- Captcha
- Tratan de evitar comportamiento no humano
- Pueden limitar el número de peticiones, por eso en algunos casos no nos dejan hacerlo tras unas cuantas peticiones.
- Es por eso que las peticiones que se hagan deben no ser muy repetitivas.

(simulando comportamiento humano)

Recomendaciones:

- Usar API siempre que se pueda
- El problema que las APIs no siempre ofrecen toda la información

Ejemplos:

API de Twitter: https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api

Twitter- con WebScraping https://github.com/twintproject/twint

Yahoo Finance API https://www.yahoofinanceapi.com/

Python el scrapeo web:

- Dispone de muchas librerías
- Podemos hacer todo el proyecto en Python no solo recopilación de información

---

Gracias por la atención

Isabel Maniega