

Big Data

Tema 6

Data scraping



Big Data



Data scraping



Big Data

- Índice:
 - ¿Qué es el scraping de datos?
 - ¿Cómo funciona el scraping de datos?
 - Aplicaciones del scraping de datos.
 - Cómo mitigar el scraping de datos.

Big Data



¿Qué es el
scrapping de datos?



Raspado de datos. Data Scrapping

- ¿Qué es el scraping de datos?
 - El **scraping de datos, data scraping o extracción de datos**, es una técnica automatizada mediante la cual un programa informático extrae datos de un sitio web o una aplicación.

La función básica del scraping es la de convertir datos que son semi-estructurados en estructurados para hacerlos fáciles de leer y entender.

- ¿Qué es el web scraping?
 - El **web scraping** es la forma más común de scraping de datos, y consiste en la extracción de contenido de páginas web, que luego puede ser utilizado para distintos fines como análisis, seguimiento de precios, investigaciones de mercado, entre otros.

En términos sencillos, es como si el programa navegara por la web como lo haría un usuario, pero con la capacidad de recopilar y almacenar la información de manera masiva y estructurada.

<https://utips.cl/que-es-el-scraping-de-datos-y-como-se-utiliza/>

Raspado de datos. Data Scrapping

El **web scraping o raspado web** es una técnica que utiliza programas de software para extraer información de sitios web.

- ¿Cómo funciona el scraping de datos?
 - El scraping web funciona enviando una solicitud a un sitio web en particular, y luego el programa encargado (conocido como scraper) recopila el código fuente de esa página web. A partir de ese código, el scraper localiza la información relevante, la extrae y la guarda en un formato que es fácil de analizar, como un archivo CSV o una base de datos.
 - Los scrapers generalmente siguen los siguientes pasos:
 - **Solicitar acceso a la página web:** El programa envía una solicitud HTTP al servidor del sitio web, como lo haría un navegador.
 - **Analizar el código HTML:** Una vez recibida la respuesta, el scraper analiza el código HTML de la página para encontrar los elementos específicos que necesita extraer.
 - **Extraer la información:** El scraper ubica y extrae los datos, que pueden ser texto, imágenes, enlaces, precios, entre otros.
 - **Almacenar los datos:** Finalmente, los datos extraídos se almacenan en un formato estructurado, como hojas de cálculo o bases de datos, para facilitar su uso posterior.

<https://utips.cl/que-es-el-scraping-de-datos-y-como-se-utiliza/>

Raspado de datos. Data Scrapping

- Los bots de scraper se pueden diseñar para múltiples propósitos, como:
 - **Extracción de contenido:** el contenido se puede extraer del sitio web para imitar las ventajas exclusivas de un producto o servicio específico basado en el contenido. Por ejemplo, un competidor podría extraer todo el contenido de las reseñas de un sitio web y reproducirlo en su propio sitio, como si fuera original.
 - **Extracción de precios:** al extraer los datos de los precios, los competidores pueden añadir información sobre la competencia, lo que les permite obtener una ventaja competitiva.
 - **Extracción de contactos:** muchos sitios web contienen direcciones de correo electrónico y números de teléfono en texto no cifrado. Al extraer información de un directorio de empleados en línea, un scraper puede reunir datos de contacto para listas de correo electrónico masivo o intentos maliciosos de ingeniería social. Este es uno de los principales métodos que utilizan tanto los spammers como los estafadores para encontrar nuevos objetivos.

<https://www.cloudflare.com/es-es/learning/bots/what-is-data-scraping/>

Raspado de datos. Data Scrapping

- Aplicaciones del scraping de datos
 - El scraping de datos es ampliamente utilizado en una variedad de industrias. Aquí algunas de las aplicaciones más comunes:
 - **Seguimiento de precios:** Empresas de comercio electrónico utilizan scrapers para monitorear los precios de sus competidores y ajustar los suyos en consecuencia.
 - **Análisis de mercados:** El scraping permite obtener información actualizada sobre las tendencias de la industria, el comportamiento del consumidor o el rendimiento de productos.
 - **Marketing digital:** Para construir bases de datos de correos electrónicos, información de contacto o análisis de competencia.
 - **Inmobiliarias:** Recopilar datos de propiedades, precios y tendencias en sitios como Portal Inmobiliario o Zillow.
 - **Investigación académica:** En algunos casos, los investigadores recurren al scraping para recopilar datos de fuentes públicas para sus estudios.

<https://utips.cl/que-es-el-scraping-de-datos-y-como-se-utiliza/>

Raspado de datos. Data Scrapping

- ¿Cómo se mitiga el scraping web?
- Algunos métodos para limitar la exposición al scraping web son los siguientes:
 - **Solicitudes de limitación de velocidad:** la velocidad de interacción de un visitante humano que va haciendo clic en diversas páginas de un sitio web es bastante predecible. Por su parte, los ordenadores pueden realizar solicitudes a una velocidad más rápida que un humano. Al limitar la velocidad del número máximo de solicitudes que puede realizar una dirección IP determinada en un lapso de tiempo dado, los sitios web pueden protegerse de las solicitudes que buscan aprovecharse de las vulnerabilidades y limitar la cantidad de extracción de datos que puede suceder en un lapso concreto.
 - **Modificar el formato HTML regularmente:** los bots de extracción de datos necesitan un formato consistente para poder recorrer eficazmente el contenido del sitio web, analizar datos útiles y guardarlos. Un método para interrumpir este flujo de trabajo es cambiar con regularidad los elementos del formato HTML para que la extracción consistente sea más complicada.
 - **Usar CAPTCHA para solicitantes de gran volumen:** además de usar una solución de limitación de velocidad, otro paso útil para frenar a los scrapers de contenido es el requisito de que los visitantes del sitio web respondan a una pregunta que sea difícil de resolver para un ordenador. No obstante, las preguntas constantes de CAPTCHA pueden afectar de forma negativa la experiencia del usuario.

<https://www.cloudflare.com/es-es/learning/bots/what-is-data-scraping/>

Big Data

- Enlaces:
 - <https://utips.cl/que-es-el-scraping-de-datos-y-como-se-utiliza>
 - <https://www.cloudflare.com/es-es/learning/bots/what-is-data-scraping/>
 - [Inteligencia artificial y datos abiertos | datos.gob.es.](https://datos.gob.es/)
 - Extraer datos de una web a excel:
 - <https://www.octoparse.es/blog/scraping-de-datos-del-sitio-web-a-excel>
 - <https://excelparatodos.com/importar-datos-de-una-web-a-excel/>
 - https://medium.com/@ines_holaOP/4-formas-de-extraer-datos-de-una-web-a-excel-49a656ae1a47
 - Extraer datos de una web a Google Sheets:
 - <https://newsinitiative.withgoogle.com/es-es/resources/trainings/google-sheets-scraping-data-from-the-internet/>
 - Extraer datos de una web con extensiones de chrome.

Big Data

Fin

Muchas gracias por su atención