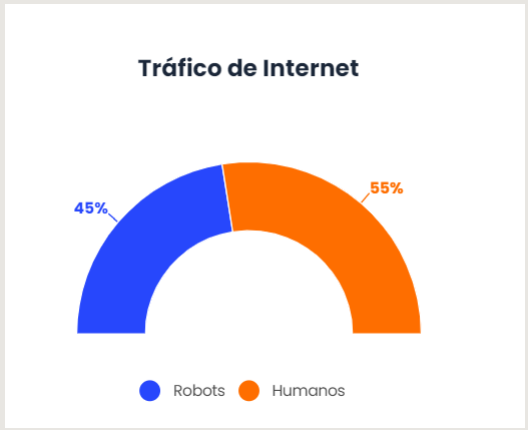
Scraping

Julio 2021

Que es Web Scraping



- El web scraping es una técnica que consiste en extraer datos de páginas de internet de forma automatizada. Es decir, **convertimos en una base de datos estructurada la información que podemos encontrar publicada en una web.**

- El uso del scraping, a pesar de ser una técnica desconocida para muchas empresas, es mucho más habitual de lo que se pueda pensar. Algunos autores hablan de que más del 45% del tráfico de la red está realizado por robots y no por humanos.

Data Scraping vs Web Crawling

- El scraping y el crawling no son lo mismo. Aun así solemos utilizar estos términos de forma indiscriminada porque la mayoría de los usuarios conocen la técnica por el término scraping, aunque lo que realmente necesitan es web crawling.
- Un crawler, o araña, se arrastra por las diferentes páginas webs imitando el comportamiento humano.

Usos del Scraping

- **Marketing y ventas**: el raspado web puede ayudar a conseguir clientes potenciales, analizar los intereses de las personas y monitorear el sentimiento del consumidor al extraer regularmente las calificaciones de los clientes de diferentes plataformas.
- **Comparación de precios**: una de las mejores formas de utilizar la tecnología de raspado de datos es recopilar información sobre precios. Por un lado, se puede recopilar datos para uno mismo para ayudarse a posicionar un producto frente a la competencia, y, por otro lado, para extraer los precios de los competidores siguiendo todos sus movimientos.
- **Gestión de la reputación y de la marca**: el raspado es una buena manera de hacer un seguimiento de lo que la gente dice sobre una empresa. Se puede administrar múltiples canales de reputación de manera eficiente. Además, también ayuda a extraer información sobre la frecuencia con la que se mencionó a la empresa en Internet. De esa manera, la empresa podría identificar cualquier desarrollo negativo desde el principio y evitar que la marca se dañe.
- **Análisis de clientes**: el Scraping puede ayudar a recopilar información demográfica útil sobre los clientes, se pueden crear estrategias de anuncios más efectivas usando esa información, y además también se pueden recopilar datos de comportamiento de los clientes para conocer el tipo de audiencia y la elección de anuncios que sean ver.
- **Generación de leads**: el raspado de datos es una herramienta muy buena para identificar posibles clientes potenciales. Puede ayudar a crear listas propias basadas en lo que se sabe sobre los posibles clientes, mirando datos como la ubicación, industria comercial, compras anteriores y más.
- **Preocupaciones estratégicas**: con esta tecnología se puede encontrar información para ayudar a las empresas con casi cualquier consideración estratégica posible. La clave es tener el conjunto de herramientas adecuado para ayudar a hacer el trabajo de manera organizada y constructiva. Esta parte también es muy útil, por ejemplo, en banca en cuanto a decisiones de inversión, ya que el scraping puede ayudar a detectar riesgos y oportunidades de inversión.
- **Mejora de las actividades de SEO**: el raspado web resuelve el problema de la búsqueda de las palabras clave correctas al rastrear las palabras clave comunes que ya se han utilizado. También puede raspar la información de la competencia para descubrir las palabras clave utilizadas por ellos. De esta manera, uno puede usar palabras clave diferentes y únicas para crear un impacto positivo en la estrategia de SEO.

Crawler

- tenemos un hotel y queremos saber el precio de la competencia en booking. Para ello programaremos un crawler que:
- 1.-Entrará en la página principal de booking
- 2.-Realizará una búsqueda por ciudad, fechas, número de personas a alojarse, etc.
- 3.-Obtendrá como resultado una lista de hoteles
- 4.-Copiará la URL de cada página de hotel en la plataforma
- 5.-Entrará en cada página y descargará los datos que necesitemos (precio, rating, disponibilidad, etc)
- 6.-Repetirá todo el proceso con todas las búsquedas que necesitemos realizar para los diferentes escenarios
- 7.-Nos devolverá una base de datos estructurada con los resultados

De todo el proceso que ha realizado nuestro crawler, solo la parte referida a la descarga de la información se consideraría data scraping. El resto se denomina web crawling.

Ventajas

- Con web crawling y data scraping los procesos de encontrar y recabar información se automatizan, con ello conseguimos:
- Disminuir carga de trabajo.
- Abaratar costes de personal.
- Aumentar la velocidad de los procesos.
- Eliminar el error humano.
- Manejar grandes cantidades de datos.
- Conseguir los datos en formatos procesables.

Desafíos Técnicos

- **Webs menos complejas**: cuanto más compleja sea la web que se desea raspar, más difícil será el raspado. Las razones son porque configurar el raspador se vuelve más difícil, y los costes de mantenimiento pueden aumentar, porque es más probable que el experto tenga errores y problemas.
- **Página de inicio estable**: el Web Scraping automatizado solo tiene sentido si la página de inicio de destino no cambia su estructura con frecuencia. Cada cambio de estructura implica costes adicionales, porque el Scraping necesitará ser ajustado.
- **Datos estructurados**: el raspado web no funcionará si se quiere raspar datos de 1000 webs diferentes y cada web tiene una estructura completamente diferente. Será necesario que exista alguna estructura básica que difiera solo en ciertas situaciones.
- **Protección baja**: si los datos en la web están protegidos, el raspado también puede convertirse en un desafío y aumentar los costes.

Como Mitigar el Scraping?

- **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; un ser humano nunca navegará a 100 páginas por segundo, por ejemplo. Por su parte, los ordenadores pueden realizar solicitudes a una magnitud más rápida que un humano.
- **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.
- **Uso de CAPTCHAs**: CAPTCHA es el acrónimo de "Completely Automated Public Turing test to tell Computers and Humans Apart". En español, "Prueba de Turing pública y automática para diferenciar máquinas y humanos". Un captcha es una prueba de tipo desafío-respuesta que se utiliza para determinar cuándo el usuario de un sistema informático es o no humano.

Referencias

- https://www.vectoritcgroup.com/tech-magazine/data-ecosystem/es-data-scraping-una-de-las-habilidades-mas-demandadas-por-las-empresas/
- http://www.noeliaespinosa.com/que-es-un-scraper/
- https://www.cloudflare.com/es-es/learning/bots/what-is-data-scraping/