Web Scraping

Narly Sanchez Caviedes
Email: *narlysanchez@javerianacali.edu.co*

Jeisson Rodriguez Rodriguez
Email: *jeissonrodriguez@javerianacali.edu.co*

I. Introducción

los datos se han convertido en el pilar fundamental sobre el cual se construyen decisiones estratégicas en todos los sectores de la sociedad. Desde grandes corporaciones hasta pequeñas empresas, y desde instituciones gubernamentales hasta organizaciones sin fines de lucro, todos dependen de los datos para entender su entorno, anticipar tendencias y tomar decisiones informadas. Carretero & Velthuis citan a Westerman destacando que "los datos son el tipo de activo prominente en el mundo digital y requieren una atención muy significativa por parte de los ejecutivos. Los datos se han convertido en los activos digitales más valiosos para las empresas".[[1] , p3]. La transformación digital ha acelerado la generación de datos a niveles sin precedentes. Cada interacción en línea, transacción comercial, y actividad en redes sociales genera datos que, cuando son analizados correctamente, pueden revelar patrones, comportamientos y oportunidades. La capacidad de recopilar, procesar y analizar estos datos otorga a las organizaciones una ventaja competitiva, permitiéndoles optimizar sus operaciones, personalizar experiencias para los clientes, y descubrir nuevas oportunidades de negocio.

El **Web Scraping** se presenta como una técnica poderosa que permite la extracción automática de grandes volúmenes de datos desde sitios web. Este proceso implica la recolección de información estructurada de páginas web, que de otra manera estaría disponible solo a través de la navegación manual. A través del uso de herramientas y tecnologías específicas, como bibliotecas de programación y frameworks, el Web Scraping facilita la obtención de datos que pueden ser utilizados para una variedad de propósitos, como análisis de mercado, monitoreo de precios, agregación de contenido, entre otros. En [2] señalan que el Web Scraping se ha convertido en la columna vertebral de muchos procesos basados en datos, desde el seguimiento de las marcas y las comparaciones de precios actualizadas hasta la realización de valiosos estudios de mercado.

II. DESARROLLO

Los datos al ser un valioso bien natural dentro de cualquier proceso de una organización, estas deben ser bien tratadas y estructuradas con el fin de extraer el mayor beneficio posible; Por esto los datos que se obtienen deben ser de buena calidad teniendo en cuenta el origen o la fuente de los mismo. Una de las primeras fases dentro de los proyectos de investigación es la recolección de los datos, para ello una de las técnicas existentes más usadas dentro de este contexto es **Web Scraping** o conocida comúnmente como Extracción Web; es un conjunto de prácticas para extraer automáticamente datos de internet, es una herramienta extremadamente útil para la recopilación de datos online, donde estos son guardados en una base de datos para posteriormente ser tratados y analizados. Hernandez et al. [3] lo definen como el proceso de rastreo y descarga de sitios web de información y la extracción de datos no estructurados o poco estructurados a un formato estructurado.

¿Qué herramientas o tecnologías se necesitan para realizar Web Scraping?

En el uso de esta técnica existen diversos medios, como lenguajes de programación, bibliotecas y herramientas, que se comunican para llegar a tal fin. Entre los lenguajes de programación más usados está Python y JavaScript, empleando bibliotecas como BeautifulSoup, Urllib2, Scrapy, Requests en Python, Selenium y Puppeteer en JavaScript. Cada uno de estos tiene una función especifica, por ejemplo:

BeautifulSoup (Python) facilita la extracción de datos de HTML y XML, es una herramienta fácil de usar, flexible, su desventaja se encuentra en que cuando tiene que procesar grandes cantidades de datos se torna lento.

Scrapy (Python) es un Framework que permite crear crawlers que pueden recorrer páginas web y extraer datos de manera rápida y eficiente, su desventaja está en que requiere una curva de aprendizaje alta

Requests se usa para hacer peticiones HTTP de manera simple y **Selenium** es utilizado para interactuar con páginas web dinámicas, permite automatizar los navegadores para acceder a este tipo de contenidos dinámicos, una desventaja consiste en su lentitud y el alto consumo de recursos, Zhao lo describe de la siguiente manera: "Hay dos módulos esenciales en un programa de web scraping: un módulo para componer una solicitud HTTP, como Urllib2 o selenium, y otro para analizar y extraer información del código HTML en bruto, como Beautiful Soup o Pyquery". [[4], p1]. La elección de cuál utilizar depende del tipo de tarea, la complejidad del sitio web y las necesidades específicas del proyecto.

Ahora, para obtener datos de un sitio web mediante Web Scraping, un Scraper debe seguir una serie de pasos básicos:

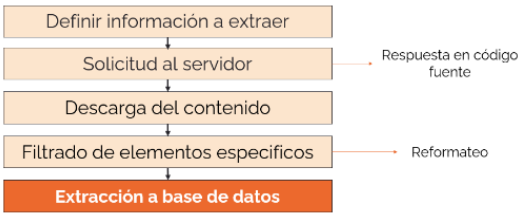

Fig. 1. Pasos del Web Scraping [5]

En la figura 1 se puede observar que, primero se identifica el sitio web de destino, se recopila las URL de las páginas de las que desea extraer, después se realiza una solicitud a estas URL para obtener el HTML de la página, luego Utiliza localizadores para extraer la información deseada del HTML, seguidamente se encuentra la etapa de limpieza y formateo de los datos para finalmente guardarlos en un archivo JSON o CSV o en algún otro formato estructurado. En la figura 2 se muestra el proceso general de la técnica Web Scraping, acá se puede visualizar tres fases, la primera es la extracción de los datos, la segunda es la clasificación y finalmente el almacenamiento en formatos estructurados.

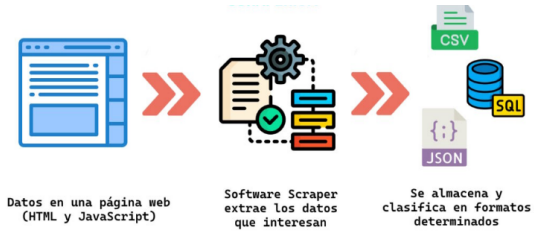

Fig. 2. Proceso de Web Scraping de una página web [6]

¿Creen que el Web Scraping es una práctica ilegal? ¿Por qué?

El web scraping no es por si misma ilegal, pero su legalidad depende del contexto y de cómo se utilice. Aquí algunos factores clave a considerar:

- **Términos de Servicio (TOS):** Muchas páginas web tienen términos de servicio que prohíben explícitamente el scraping sin permiso. Si se viola este acuerdo, podría haber consecuencias legales, aunque no siempre esto se lleva a juicio.

- **Propiedad intelectual:** Algunos sitios web tienen contenido protegido por derechos de autor o restricciones de uso. Copiar grandes cantidades de información sin permiso podría infringir leyes de propiedad intelectual.

- **Protección de Datos:** En algunos países, como los que están bajo la jurisdicción del Reglamento General de Protección de Datos (GDPR) en Europa, el scraping de información personal sin el consentimiento adecuado puede ser ilegal.

- **Computer Fraud and Abuse Act (CFAA):** En EE.UU., el CFAA ha sido utilizado en ciertos casos de scraping para argumentar que acceder a sistemas informáticos sin autorización es ilegal. El caso más notable fue el de HiQ Labs v. LinkedIn, donde la corte determinó que scraping de datos públicamente accesibles no violaba el CFAA, pero las circunstancias pueden variar.

Uso de bots y automatización: Algunas páginas prohíben explícitamente el uso de bots o automatización. Realizar scraping mediante estos métodos podría llevar a una demanda, especialmente si interfiere con el rendimiento del sitio.

¿En qué situaciones usarían Web Scraping?

El web scraping se utiliza en multiples escenarios cuando se necesita recopilar datos de sitios web de manera automática. Por ejemplo:

1. Monitoreo de precios

• **Comercio electrónico:** Empresas de e-commerce usan web scraping para rastrear los precios de sus competidores y ajustar los propios para mantener competitividad.
• **Consumidores:** Los consumidores o aplicaciones de comparación de precios lo utilizan para encontrar las mejores ofertas en productos y servicios.

2. Investigación de mercado

• Empresas recopilan información sobre tendencias de productos, opiniones de los consumidores y datos de ventas de sitios web públicos para analizar el comportamiento del mercado y ajustar sus estrategias de negocio.

3. Recopilación de datos financieros

• **Análisis bursátil:** Los inversores usan web scraping para obtener datos de acciones, criptomonedas y otras métricas financieras de sitios web como Yahoo Finance o CoinMarketCap, para realizar análisis de tendencias.
• **Noticias:** Los analistas financieros también extraen noticias relevantes sobre empresas o sectores que puedan afectar el mercado.

4. Análisis de contenido y medios sociales

• **Opiniones y tendencias:** Investigadores y empresas analizan las menciones de productos, opiniones de usuarios y temas populares en redes sociales o foros.
• **Estudios académicos:** Para realizar análisis de contenido o sentiment analysis (análisis de sentimiento), los investigadores extraen comentarios, artículos, o publicaciones en redes sociales.

5. Monitoreo de empleo

• **Agencias de empleo:** Herramientas de scraping se utilizan para rastrear listados de empleo en varios portales,

lo que permite a las agencias y plataformas de búsqueda de trabajo crear bases de datos actualizadas con ofertas.
• **Investigación de salarios:** El scraping de sitios web de empleo puede ayudar a analizar las tendencias salariales y las demandas de diferentes profesiones.

6. Recopilación de datos académicos o científicos

• Investigadores pueden usar scraping para recopilar datos de publicaciones científicas, repositorios de datos abiertos, o incluso bases de datos sobre patentes e innovaciones tecnológicas.

7. Automatización de tareas repetitivas

• Cuando una tarea requiere extraer información regularmente de un sitio web y no hay una API disponible, el scraping puede ser una solución para evitar hacerlo manualmente.

8. Seguimiento de cambios en sitios web

• **Cambios legislativos o normativos:** Organizaciones monitorean sitios web del gobierno para estar al tanto de actualizaciones en leyes o regulaciones.
• **Inventarios**: Algunas empresas usan scraping para monitorear inventarios y cambios en la disponibilidad de productos en varios sitios.

9. Extracción de datos inmobiliarios

• **Precios de propiedades:** Empresas del sector inmobiliario pueden extraer datos de sitios web para analizar tendencias de precios y comparar el valor de las propiedades en diferentes ubicaciones.

10. Enriquecimiento de bases de datos

• Las empresas pueden utilizar scraping para complementar sus bases de datos con información pública, como perfiles de usuarios o empresas disponibles en la web.

Es importante que, en cualquiera de estas situaciones, el web scraping se realice respetando las leyes de protección de datos y los términos de servicio de los sitios web.

III. Conclusión

El web scraping es una herramienta extremadamente útil y versátil para obtener datos de diversas fuentes en línea, permitiendo a las organizaciones acceder a información de manera eficiente, automatizada y en grandes cantidades. Esto es especialmente valioso en sectores como el comercio electrónico, la investigación de mercado, las finanzas y el análisis de medios sociales.

Las herramientas de web scraping ofrecen una amplia gama de soluciones, desde opciones sencillas para pequeños proyectos hasta frameworks avanzados que permiten scraping de sitios complejos y dinámicos. La elección de la herramienta adecuada depende de las necesidades del proyecto, el tipo de datos a extraer y el nivel de interacción que se requiere con el sitio web.

La legalidad del **web scraping** depende principalmente del uso que se haga de la información, las restricciones impuestas por el sitio web, y las leyes de protección de datos. Aunque el scraping en sí no es inherentemente ilegal, es importante estar al tanto de las regulaciones locales y respetar los términos de servicio del sitio.

References

[1] A. I. G. Carretero and M. P. Velthuis, “Importancia de la calidad de los datos en la transformación digital,” *RUIDERAe: Revista de Unidades de Información.(ISSN 2254-7177)*, no. 13, 2018.
[2] Kinsta, “¿qué es el web scraping? cómo extraer legalmente el contenido de la web,” Available at https://kinsta.com/es/base-de-conocimiento/que-es-web-scraping/.
[3] A. T. Hernández, E. G. Vázquez, C. A. B. Rincón, J. M. García, A. C. Maldonado, and R. Ibarra-Orozco, “Metodologías para análisis político utilizando web scraping.” *Res. Comput. Sci.*, vol. 95, pp. 113–121, 2015.
[4] B. Zhao, *Web Scraping*. Cham: Springer International Publishing, 2017, pp. 1–3. [Online]. Available: https://doi.org/10.1007/978-3-319-32001-4_483-1
[5] Datademia, “¿qué es web scraping?” Available at https://datademia.es/blog/que-es-web-scraping.
[6] Scraperium, “¿qué es el web scraping y para qué sirve?” Available at https://scraperium.com/que-es-y-para-que-sirve-web-scraping/.