Proyecto Tecnológico Integrador

Informe sobre la Técnica de Scraping

El termino “Scraping” es muy utilizado en el ámbito de la informática, pero que significa en si? y porque se utiliza?

Se trata una palabra de origen Ingles, cuyo significado literal es “raspar”. En Programación hacemos referencia, habitualmente a “web scraping”, y es para describir un proceso importante que puede ser de mucha ayuda, en el mundo de los negocios.

“El web scraping es un conjunto de prácticas utilizadas para extraer automáticamente — o «scrapear» — datos de la web, y si es necesario, se reutiliza en una versión modificada de otra web”.

En el mejor de los casos, el web scraping sirve para muchos propósitos útiles en muchas industrias. En 2021, casi la mitad del web scraping se utiliza para reforzar las estrategias de comercio electrónico.

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.

Es posible scrapear todo tipo de datos de la web. Desde los motores de búsqueda y los feeds RSS hasta la información gubernamental, la mayoría de los sitios web ponen sus datos a disposición.

Sin embargo, eso no significa que estos datos, estén *siempre* disponibles. Dependiendo del sitio web, puede que se deban emplear algunas herramientas y trucos para obtener exactamente lo necesario — suponiendo que los datos sean accesibles en primer lugar. Por ejemplo, muchos scrapers web no pueden extraer datos significativos del contenido visual.

La realización del web scraping se lleva a cabo mediante una variedad de técnicas:

- Manipulación HTTP: Se copia contenido de páginas web estáticas o dinámicas a través de solicitudes HTTP.
- Minería de Datos: Se identifican contenidos a través de plantillas y scripts, para luego ponerlos a disposición en otras páginas web.
- Herramientas de Scraping: Realizan tareas automatizadas y controladas manualmente, extrayendo desde contenido hasta funcionalidades.

- Analizadores HTML: Recuperan y convierten datos de otras páginas web.
- Copia Manual: Desde la copia de texto hasta fragmentos completos de código fuente.
- Escaneo de Microformatos: Componentes populares en la web semántica.

Es crucial entender la complejidad de la legalidad en el web scraping. Aunque la técnica en sí no es intrínsecamente ilegal, el acceso a datos que no están destinados al público puede presentar problemas legales, especialmente en casos de datos personales o propiedad intelectual. La legalidad varía según la jurisdicción y la naturaleza de los datos, haciendo que el cumplimiento de regulaciones y normativas sea esencial.

En Argentina, aunque no existe una regulación específica para el web scraping, se aplican varias leyes y normativas que pueden ser relevantes:

- Ley de Protección de Datos: Regula la recopilación y procesamiento de datos personales, haciendo necesario obtener el consentimiento expreso de los titulares antes de realizar web scraping que involucre datos personales.
- Ley de Propiedad Intelectual: Protege los derechos de autor y propiedad intelectual, lo que implica que el web scraping de contenido protegido puede infringir derechos a menos que se tenga autorización o se cumpla una excepción legal.
- Ley de Lealtad Comercial: Considera desleal el web scraping con fines competitivos si implica el robo de información a la competencia.
- Términos de Uso del Sitio: Antes de realizar web scraping en un sitio, es vital revisar y acatar los términos de uso, ya que algunos prohíben o limitan esta técnica.

El software mas utilizado para realizar el web scraping es Python y en su nivel más básico, el web scraping se reduce a unos simples pasos:

- Primero, el fragmento de código utilizado para extraer la información, que llamamos bot de extracción, envía una solicitud HTTP GET a un sitio web específico.
- Cuando el sitio web responde, el scraper analiza el documento HTML para buscar un patrón de datos específico.
- Una vez se hayan extraído los datos, se convierten a cualquier formato específico proyectado por el autor del bot de scraper.

Los bots de scraper se pueden diseñar para múltiples propósitos, como:

1. 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 producto como Yelp se basa en reseñas; un competidor podría extraer todo el contenido de las reseñas de Yelp y reproducirlo en su propio sitio, como si fuera original.
2. Extracción de precios: al extraer los datos de los precios, los competidores pueden añadir información sobre la competencia. Esto les permite formular una ventaja competitiva.
3. Extracción de contactos: muchos sitios web contienen direcciones de correo electrónico y números de teléfono en texto no cifrado. Al extraer ubicaciones como un directorio de empleados en línea, un scraper puede reunir datos de contacto para listas de correo electrónico masivo, llamadas automáticas o intentos maliciosos de ingeniería social. Es uno de los principales métodos que utilizan tanto los spammers como los estafadores para encontrar nuevos objetivos.

Estas son solo algunas de las principales librerías de Python para Web Scraping.

Cada una tiene sus propias fortalezas y se adapta a diferentes necesidades y casos de uso:

- Beautiful Soup: Facilita análisis y extracción de datos de HTML y XML. Con sintaxis amigable y funcionalidades poderosas, navega la estructura del código fuente para extraer elementos deseados de forma intuitiva.

- Scrapy: Solución completa para proyectos de Web Scraping a gran escala. Alta eficiencia y código abierto, ofrece herramientas para extraer datos estructurados. Incluye sistema de peticiones, programación de spiders personalizados y procesamiento de datos.

- Selenium: Diseñada para automatizar navegadores web. Ideal para sitios con contenido dinámico generado por JavaScript u otras tecnologías interactivas. Simula acciones de navegación, completado de formularios y extracción de datos programáticamente.

- Requests: Herramienta esencial para solicitudes HTTP en Python. Sintaxis simple para enviar y recibir datos a través de solicitudes GET y POST. Usada junto a librerías como BeautifulSoup para descargar y analizar contenido web.

- PyQuery: Similar a jQuery, permite usar selectores CSS y manipular documentos HTML. Facilita la extracción y manipulación de elementos específicos, adecuada para tareas más simples de Web Scraping.

Estos son solo algunos ejemplos de cómo se puede aplicar el Web Scraping en diferentes áreas:

1. Extracción de precios y datos de productos: Las empresas de comercio electrónico pueden utilizar el Web Scraping para extraer información de precios y detalles de productos de diferentes sitios web. Esto les permite realizar análisis comparativos, ajustar sus estrategias de precios y mantenerse competitivos en el mercado.
2. Monitorización de opiniones y reseñas: Las empresas pueden utilizar el Web Scraping para recopilar opiniones y reseñas de productos o servicios de múltiples sitios web y plataformas. Esto les permite obtener una visión general de la percepción del cliente, identificar patrones y tendencias, y ajustar sus estrategias de marketing y desarrollo de productos.
3. Recopilación de datos para investigación académica: Los investigadores académicos pueden utilizar el Web Scraping para recopilar datos relevantes de sitios web, como estudios científicos, noticias, información demográfica, entre otros. Esto les permite obtener grandes volúmenes de datos para su análisis y estudio.
4. Análisis de mercado y seguimiento de la competencia: Las empresas pueden utilizar el Web Scraping para monitorear y recopilar datos sobre la actividad de sus competidores, como precios de productos, promociones, estrategias de marketing y más. Esto les permite tomar decisiones informadas y ajustar su enfoque comercial.
5. Extracción de datos de redes sociales: El Web Scraping se utiliza para extraer datos de plataformas de redes sociales, como Twitter o Instagram, para analizar tendencias, recopilar datos demográficos, realizar estudios de opinión pública, entre otros fines.

Referencias:

- https://es.ryte.com/wiki/Scraping
- ¿Qué Es el Web Scraping? Cómo Extraer Legalmente el Contenido de la Web - : https://kinsta.com/es/base-de-conocimiento/que-es-web-scraping/
- Qué es el web scraping y para qué sirve - https://www.antevenio.com/blog/2019/03/que-es-el-web-scraping-y-para-que-sirve/
- ¿Qué es el scraping de datos? - https://www.cloudflare.com/es-es/learning/bots/what-is-data-scraping/

- ¿Qué es el web scraping? – Regulacion Argentina - https://leponte.com.ar/implicancias-legales-del-web-scraping/

Autores:

Culasso, Franco

Godoy, Guillermo

Luján, Daniel

Pedrozo, Analía Belén

Rodriguez Yampa, Carmen Valentina

Sanchez, Martín