TEC | Tecnológico de Costa Rica

Introducción al Desarrollo de Páginas Web

Tarea Investigación 1

Estudiante:
Marco Soto Morera
2014046514

Profesor:
Efrén Jiménez Delgado

Grupo 51

I Semestre - 2022

Resume:

Web scraping es una forma de minería de datos no estructurada, que permite extraer información de páginas web, escanear su código HTML y generar patrones de extracción de datos. El documento muestra las técnicas que existen, situaciones legales de ciertas páginas, herramientas existentes y con que lenguajes de programación se puede utilizar.

Además, se encuentra una metodología de extracción de una página web, que se contempla la obtención de datos, el parseo de los datos, y el almacenaje de los datos. Se incluye el lenguaje de programación a utilizar, la base de datos y el diagrama conceptual de esa metodología. La página web contiene una lista de elementos.

Tabla de contenidos

Introducción:

El Web scraping es una técnica mediante la cual es posible obtener información estructurada a partir de un sitio web, para ello se realiza una simulación de la navegación humana con el objetivo de realizar un análisis de datos (Big Data) o realizar procesos Web automatizados. El objetivo puede ser transformar contenidos, almacenar datos de la web, reconocer estructuras de código HTML único, recopilar información, hacer minería y análisis de datos, automatizar la creación de enlaces, price mapping, caza de tendencias, monitorización de la competencia, optimización de precios.

El Web scraping se utiliza en aquellos casos en los que no se dispone de una API pública para conseguir el mismo objetivo. La extracción de datos web es utilizada por personas y empresas que desean hacer uso de la gran cantidad de datos web disponibles públicamente para tomar decisiones más inteligentes.

Desarrollo:

El web scraping se utiliza para una gran variedad de tareas, entre ellas, para recopilar datos de contacto o información especial con gran rapidez. En el ámbito profesional, se utiliza a menudo para obtener ventajas respecto a la competencia. De esta forma, por medio del harvesting de datos, una empresa puede examinar todos los productos de un competidor y compararlos con los propios. El web scraping también resulta valioso en relación con los datos financieros: es posible leer datos desde un sitio web externo, organizarlos en forma de tabla y después analizarlos y procesarlos.

Existen diferentes modos de funcionamiento, el web scraping automático y el manual. El web scraping manual define el copiado y pegado manual de información y datos, como recortar y guardar artículos de periódico y solo se lleva a cabo si se desea encontrar y almacenar alguna información en específico. Es un proceso muy laborioso que raras veces se aplica a grandes cantidades de datos.

El web scraping automático, se realiza mediante un software o un algoritmo que analiza diferentes páginas web para extraer información. Se utiliza software especializado según el tipo de página web y el contenido. Dentro del web scraping automático, se diferencian varios modos de realizarse:

Analizador sintáctico: se pueden denominar parsers, se utilizan para convertir un texto en una nueva estructura. En los análisis de HTML, el software lee un documento HTML y almacena la información. Un analizador DOM utiliza la representación de contenidos del lado del cliente en el navegador para extraer datos.

Bots: son software dedicados a realizar determinadas tareas y automatizarlas. Se utilizan para examinar páginas web automáticamente y recopilar datos.

Texto: aquellos que tienen experiencia con la línea de comandos pueden aprovechar la función grep de Unix para buscar en la web determinados términos en Python o Perl. Este es un método muy sencillo para extraer datos, aunque requiere más trabajo que la utilización de un software.

Para hacer Web scraping es necesario analizar aspectos cómo:

- Accesibilidad de los datos de origen
- Análisis de patrones de los datos
- Frecuencia de extracción de los datos con el objetivo de buscar la vía óptima para obtener los datos.

El web scraping no siempre es legal. Siempre se debe tener en cuenta los derechos de propiedad intelectual de los sitios web. El web scraping tiene consecuencias muy negativas para algunas tiendas online y proveedores.

El web scraping es legal, siempre y cuando los datos recabados estén disponibles libremente para terceros en la web. Para garantizar la legalidad del web scraping, hay que tener en consideración lo siguiente:

- Observar y cumplir con los derechos de propiedad intelectual. Si los datos están protegidos por estos derechos, no se pueden publicar en ninguna otra parte.
- Los operadores de las páginas tienen derecho a recurrir a procesos técnicos para evitar el web scraping que no pueden ser evadidos.

- Si, para la utilización de datos, se requiere el registro de usuarios o un contrato de utilización, estos datos no podrán ser aprovechados mediante web scraping.
- No se permite la ocultación de publicidad, de términos y condiciones, ni de descargos de responsabilidad mediante tecnologías de web scraping.

Algunas herramientas son:

- 80legs.com: un plan gratuito para web scraping
- Apifier.com: el web scraper para los que dominan JavaScript
- Dexi.io: herramienta de web scraping para usuarios avanzados
- Diffbot.com: inteligencia artificial para la extracción de datos
- Driffbot es una herramienta de web scrapping diseñada para hacerlo todo muy fácil.
- Hunter.io: una herramienta de web scraping para capturar correos electrónicos
- Import.io: extrae datos casi de cualquier web
- Mozenda.com: el binomio de web scraping y data as a service más completo
- Parsehub.com: una herramienta de web scraping especializada en páginas dinámicas Salestools.io: un scraper para equipos comerciales
- Webhose.io: transforman los datos desestructurados de una web en dato estructurados

El web scraping se puede utilizar con los siguientes lenguajes:

- Python
- Node.js
- C++
- PHP

Conclusiones y recomendaciones:

El Web Scraping permite obtener gran cantidad de información de una manera automática de diferentes sitios web; gracias a esto, se puede crear bases de datos aprovechables, las cuales tienen una infinidad de maneras de ser aplicadas.

Hoy la información es sumamente valiosa para optimizar los diferentes procesos de una organización, conocer mejor a un cliente y para tomar decisiones acertadas.

Bibliografía:

Llorens, J. (2021). *Aplicación web para obtener de manera automática un estado del arte de un tema de investigación*. Recuperado 12 de marzo de 2022, de https://rua.ua.es/dspace/bitstream/10045/115976/1/Aplicacion_web_para_obtener_de_manera_automatica__Llorens_Vera_Jeronimo_Jose.pdf

Murillo, D. (2018). *Vista de Web Scraping de los Perfiles y Publicaciones de una Afiliación en Google Scholar utilizando Aplicaciones Web e implementando un Algoritmo en R | Memorias de Congresos UTP*. revistas.utp.ac.pa. Recuperado 13 de marzo de 2022, de https://revistas.utp.ac.pa/index.php/memoutp/article/view/1465/html