Universidad Autónoma de Coahuila

Facultad de Sistemas

Materia: Diseño y Arquitectura de Software

Maestro: Ángel Santiago Jaime Zavala

Alumno: Héctor Daniel Bucio Montes

Parsing o Web Scraping

JUSTIA - Parsing o Web Scraping

El joven que venía de parte de Justia comenzó definiendo lo qué es un blog, y en pocas palabras dijo que es una página web en la que se puede poner diversos tipos de contenidos, luego pasó con la siguiente pregunta…

* Qué es Web Scraping?

- Y él lo defino como una variedad de métodos para recolectar información de internet, aunque también mencionó que lo bueno de esto es que generando scripts puedes obtener de manera automatizada esta información y aún más sorprendente es que haciendo web scraping basado en inteligencia artificial podemos obtener exactamente la información que deseamos, ignorando aquellas partes que en realidad no te interesan y que solo te van a ocupar un espacio en tu archivo, o sea es como si realmente una persona estuviera recopilando esta información.

Luego el joven habló de otros usos del web scraping y mencionó como ejemplo algo muy conocido debido a que nos lo encontramos muy fácilmente en la televisión o en internet, y esto es la comparación de precios y su monitoreo, e igualmente mencionó como ejemplo que las empresas que hacían esto son Trivago y Kayak.

* Pero ¿Cómo se realiza?

- Por medio de Scripts (Expresiones regulares, librerías)
- Frameworks

*¿Cuáles son los aspectos que se deben considerar?

- Frecuencia de extracción.
- Monto de datos a extraer y recursos disponibles.
- Accesibilidad al origen de datos.

* ¿Cuáles son los obstáculos?

- Las sesiones de navegación
- Que requiere usuario y contraseña.

* Posibles soluciones

- Crear un usuario para acceder a la primer página
- Usar un framework o herramienta basada en JAVASCRIPT
- Pool de IP's, servicios de resolución de CAPTCHA, Computer Vision

*¿ Por qué es importante en general?

- Para obtener listados de inmuebles en venta.
- Reunir direcciones de correo electrónico o teléfonos.
- Reviews de productos de la competencia.
- Para extraer información de diferentes redes sociales.
- Para obtener cantidades masivas de datos para fines de investigación
- Para generar listas de perfiles de personas [Reclutamiento]

* ¿Para qué me sirve?

- Para tener datos cuando inicie un sitio desde cero.
- Para monitorear el momento en que se abre la venta de boletos en alguna página.
- Para monitorear el precio de un producto.
- Para conseguir trabajo.

* Definición de conceptos usados en la conferencia

- Scraper – Es un programa o script que busca y extrae información de páginas web de manera automatizada.

- Crawler/rastreador – Es un programa o script que busca páginas de internet existentes, las scrapea y las indexa en sus motores de búsqueda/caché.

- Expresión Regular – Es un conjunto de caracteres que definen un patrón.

- Código Fuente – Es un conjunto de líneas de código que definen la estructura, diseño y funcionamiento de una página de internet.