Scrapper Distribuido

Web scraping es una técnica utilizada mediante programas de software para extraer información de sitios web. Usualmente, estos programas simulan la navegación de un humano en la World Wide Web ya sea utilizando el protocolo HTTP manualmente, o incrustando un navegador en una aplicación.

El web scraping está muy relacionado con la indexación de la web, la cual indexa la información de la web utilizando un robot y es una técnica universal adoptada por la mayoría de los motores de búsqueda. Sin embargo, el web scraping se enfoca más en la transformación de datos sin estructura en la web (como el formato HTML) en datos estructurados que pueden ser almacenados y analizados en una base de datos central, en una hoja de cálculo o en alguna otra fuente de almacenamiento. Alguno de los usos del web scraping son la comparación de precios en tiendas, la monitorización de datos relacionados con el clima de cierta región, la detección de cambios en sitios webs y la integración de datos en sitios webs. También es utilizado para obtener información relevante de un sitio a través de los rich snippets.

En los últimos años el web scraping se ha convertido en una técnica muy utilizada dentro del sector del posicionamiento web gracias a su capacidad de generar grandes cantidades de datos para crear contenidos de calidad.

Su tarea consiste en a partir de un conjunto inicial de urls, devolver el html correspondiente a cada una de ellas

Descripción del sistema distribuido

El sistema estará constituido por nodos los cuales tendrán una responsabilidad específica (role), que puede ser almacenar datos, enrutar pedidos, consultar el sistema, etc. La existencia de un único role para cada nodo no impide que dos nodos no puedan estar en un mismo host. El sistema debe estar disponible siempre que halla algún nodo disponible por cada role (eso no implica que deba dar una respuesta rápida ante los pedidos). El sistema no puede perder datos en caso de que falle un nodo de almacenamiento. En caso de que el sistema sufra una partición (se divida en dos o más subsistemas producto de la pérdida de nodos o enlaces entre nodos), debe ser capaz de reconectarse y funcionar como un solo sistema cuando exista la posibilidad de comunicación entre los nodos de los subsistemas.

Nota

Cualquier enriquecimiento del proyecto es válido y se tendrá en cuenta en la evaluación del mismo. En caso de modificar la orden del proyecto debe consultarse a los profesores con anterioridad.