Web Scraping con ML

I.S.P.C

TSIT4.0

Proyecto Tecnologico Integrador

Autores

William Bernardo Leyton Segovia

Tecnicatura Superior en Innovacion con Tecnlogias 4.0

Cohorte 2 0 2 2

ISPC

Introducción

El web scraping, o raspado web, es una técnica utilizada para extraer información de sitios web de manera automatizada. Esta práctica se ha vuelto esencial en el ámbito de la recopilación de datos, investigación y análisis, permitiendo a los desarrolladores y analistas acceder a datos específicos de páginas web de manera eficiente.

Definición: el web scraping implica la extracción automática de datos de páginas web, generalmente mediante el envío de solicitudes HTTP a las URL correspondientes y el análisis del HTML resultante para extraer información estructurada.

¿Quiénes la utilizan?

El scraping es utilizado por una variedad de profesionales y entidades, incluyendo:

- **Desarrolladores de Software:** Para recopilar datos relevantes para aplicaciones y servicios.
- **Analistas de Datos:** Para obtener conjuntos de datos utilizables en análisis y visualizaciones.
- **Empresas:** Para monitorear competidores, recopilar datos de mercado y realizar investigaciones.
- **Investigadores:** Para obtener información específica en investigaciones académicas o científicas.

Principios Básicos:

- **Solicitudes HTTP:** Se utilizan para acceder al contenido de las páginas web.
- **Análisis HTML:** Las librerías como BeautifulSoup son esenciales para analizar la estructura HTML y extraer datos específicos.

Aplicaciones Comunes:

- **Recopilación de Datos:** Obtención de información para análisis y toma de decisiones.
- **Monitorización:** Seguimiento de cambios en sitios web específicos.
- **Automatización:** Realización automática de tareas como completar formularios en línea.

¿Es legal? ¿Existe legislación al respecto?

La legalidad del scraping depende del contexto y de la conformidad con los términos de servicio del sitio web. Algunos sitios prohíben explícitamente el scraping en sus políticas. En términos legales, la legislación varía, pero en muchos lugares, realizar scraping sin permiso puede infringir derechos de autor o leyes de acceso no autorizado.

Es crucial realizar el web scraping de manera ética y respetando los términos de servicio de los sitios web. Algunos sitios pueden tener restricciones o políticas contra el scraping, por lo que es esencial obtener el permiso necesario antes de realizar esta técnica.

Herramientas:

- **Requests:** Librería para realizar solicitudes HTTP en Python.
- **Beautiful Soup (BeautifulSoup4):** Utilizada para analizar y extraer datos de documentos HTML.
- **Pandas:** Librería para manipulación y análisis de datos en Python.
-

Objetivos del Web Scraping:

- Recopilación eficiente de datos específicos.
- Automatización de procesos de recolección de información.
- Facilitar el análisis y la toma de decisiones.

Elección de Librerías

Durante la fase de investigación, se llevó a cabo una cuidadosa evaluación de diversas librerías disponibles para Python con el objetivo de seleccionar las herramientas más apropiadas para el proyecto de web scraping. Las librerías finalmente elegidas son Requests, BeautifulSoup4 y Pandas, y su selección se basó en las siguientes razones:

1. **Requests:**
    - **Justificación:** Requests es una librería de Python diseñada para hacer solicitudes HTTP de manera sencilla y eficiente. Se seleccionó por su facilidad de uso y su capacidad para gestionar de manera efectiva las solicitudes y respuestas HTTP.
    - **Características Principales:**
        1. Interfaz simple para realizar solicitudes HTTP.
        2. Soporte para diversos métodos HTTP (GET, POST, etc.).
        3. Manejo fácil de parámetros y encabezados en las solicitudes.
2. **Beautiful Soup (BeautifulSoup4):**
    - **Justificación:** BeautifulSoup4 es una poderosa librería de Python diseñada para extraer información de documentos HTML y XML. Se eligió por su capacidad para analizar y buscar datos en el HTML resultante de las solicitudes HTTP.
    - **Características Principales:**
        1. Análisis de documentos HTML y XML de manera intuitiva.
        2. Navegación sencilla a través de la estructura del documento.
        3. Soporte para encontrar y filtrar elementos HTML de manera eficiente.

3. **Pandas:**

- **Justificación:** Pandas es una librería esencial para manipulación y análisis de datos en Python. Se seleccionó por su capacidad para organizar y estructurar los datos extraídos en un formato tabular, facilitando su análisis y visualización.
- **Características Principales:**
    1. Estructuras de datos flexibles (DataFrames) para manejar datos tabulares.
    2. Herramientas para realizar operaciones de limpieza y transformación de datos.
    3. Integración con otras librerías para análisis de datos.