PROYECTO TECNOLÓGICO INTEGRADOR

TSIT4.0

2023

Scraping Web

Docente:
Arce Kessler Kevin

Estudiantes:
Arce José Emiliano
Fonseca Luis Fernando
Garza Ribana
Guzmán Federico
Yañez Valentina

TSIT4.0 – 2023

PROYECTO TECNOLÓGICO INTEGRADOR

Apertura: lunes, 14 agosto 2023, 12:01 PM

Cierre: lunes, 21 agosto 2023, 11:59 PM

CONSIGNA DE TRABAJO:

Realizar un informe de no más de tres páginas sobre:

¿Qué es la técnica de Scraping?

¿Quiénes la utilizan?

¿Cuáles son sus principales usos?

¿Es legal? ¿Existe legislación al respecto? Si existe, citarla.

¿Qué lenguajes de programación pueden utilizarse para la misma?

¿Qué librerías se utilizan?

De tres ejemplos en los que podría utilizarse la técnica.

Citar bibliografía.

Es un INFORME, no son preguntas a responder, se debe elaborar un texto informativo, las preguntas son disparadoras de temas que se deben tratar, pero no se deben responder explícitamente.

Si bien el trabajo debe hacerse de forma GRUPAL, la entrega de la actividad debe ser INDIVIDUAL, por cuestiones de corrección en el aula.

EXPLORANDO EL SCRAPING WEB: UNA VISIÓN INTEGRAL DE LA TÉCNICA Y SUS APLICACIONES

Introducción

En la era digital actual, la obtención y análisis de datos se ha vuelto fundamental para la toma de decisiones en diversos ámbitos. Una técnica ampliamente utilizada para extraer información de sitios web de manera automatizada es el "scraping web". En este texto informativo, explicaremos en detalle qué es el scraping web, quiénes lo utilizan, sus principales usos, su legalidad y legislación, lenguajes de programación, librerías asociadas y ejemplos de aplicación.

¿Qué es el scraping web?

El "scraping" (también conocida como "web scraping" en inglés) es una técnica informática utilizada para extraer información de sitios web de manera automatizada. Consiste en el proceso de obtener datos de páginas web de forma programática, generalmente a través de la descarga y análisis del código HTML de una página, para luego extraer la información relevante, como texto, imágenes, enlaces u otros datos estructurados o no estructurados.

Para realizar el scraping, se utiliza software o scripts que simulan la navegación humana a través de la web.

Quiénes la utilizan y principales usos:

El scraping web es ampliamente utilizado y a continuación, listamos diferentes entidades que lo emplean, como así también, con qué finalidad:

- Empresas de análisis de datos: para recopilar información sobre su competencia, tendencias del mercado, precios de productos y otros datos relevantes para la toma de decisiones estratégicas.
- Investigadores académicos: para recopilar datos y realizar análisis en áreas como la sociología, la economía, la ciencia política y otros campos que requieren grandes conjuntos de datos.
- Periodistas y medios de comunicación: para investigar temas, recopilar datos para reportajes y analizar información de interés público.
- Profesionales de marketing: para obtener información sobre tendencias en redes sociales, comentarios de clientes y otros datos relevantes para estrategias de marketing.
- Desarrolladores de aplicaciones: para obtener datos de sitios web y luego mostrarlos en sus aplicaciones o servicios.
- Usuarios individuales: para recopilar información específica para sus proyectos personales, como recetas de cocina, listas de reproducción de música, datos meteorológicos, entre otros.
- Agencias gubernamentales: para recopilar datos relacionados con la salud pública, la economía o el medio ambiente.

Legalidad y Legislación

La legalidad del web scraping varía según la jurisdicción y las circunstancias específicas en las que se realiza. Si bien el web scraping en sí mismo no es ilegal, puede involucrar acciones que infrinjan términos de uso de sitios web, políticas de privacidad, derechos de autor y otras leyes relacionadas. Aquí hay algunos aspectos a considerar:

Términos de uso y políticas de privacidad: Muchos sitios web tienen términos de uso que establecen cómo se puede acceder y utilizar su contenido. Algunos sitios pueden prohibir explícitamente el scraping en sus términos de uso. Si un sitio web prohíbe el scraping en sus términos, hacerlo podría llevar a problemas legales.

Derechos de autor: El scraping de contenido protegido por derechos de autor puede infringir la ley si se utiliza sin el permiso adecuado del titular de los derechos. Aunque algunos tipos de datos, como hechos o cifras, pueden no estar protegidos por derechos de autor, la forma en que se presenta la información podría estar protegida.

Protección de datos: Si el scraping involucra la recopilación de información personal de los usuarios, es importante tener en cuenta las leyes de protección de datos, como el Reglamento General de Protección de Datos (RGPD) en la Unión Europea, que establecen reglas estrictas sobre cómo se puede recopilar y procesar información personal.

Fraude, abuso y uso indebido: En algunos casos, el scraping puede ser utilizado con intenciones maliciosas, como el robo de información sensible, la distribución de spam o el fraude. Estas actividades pueden ser ilegales y estar sujetas a acciones legales.
Exclusión de robots: Algunos sitios web especifican en su archivo robots.txt cómo se debe acceder a su contenido por parte de rastreadores y scrapers. Ignorar estas instrucciones puede llevar a problemas legales.

Lenguajes de Programación y Librerías

El scraping web puede implementarse en varios lenguajes de programación, pero los más comunes incluyen Python, Ruby y JavaScript. Para Python, las librerías más populares son Beautiful Soup y Scrapy, mientras que, para JavaScript, Puppeteer es ampliamente utilizado.

Ejemplos de Aplicación

- Seguimiento de Precios en Comercio Electrónico: Una empresa puede usar scraping web para monitorear los precios de productos en varias tiendas en línea y ajustar sus propios precios en consecuencia.
- Extracción de Noticias: Un periodista podría utilizar scraping web para extraer automáticamente titulares de noticias de diferentes sitios web y analizar tendencias.
- Recopilación de Datos de Propiedades Inmobiliarias: Una agencia de bienes raíces podría utilizar scraping web para recopilar información sobre propiedades disponibles en línea y ofrecer a los clientes opciones actualizadas.
- A continuación, en la imagen que anexamos, puede visualizarse cómo se utilizan las librerías en la técnica de scraping web en el lenguaje de Python:

```python
import requests
from bs4 import BeautifulSoup

# URL del sitio web a hacer scraping
url = 'https://www.ejemplonoticias.com/'

# Realizar una solicitud HTTP a la URL
response = requests.get(url)

# Parsear el contenido HTML con Beautiful Soup
soup = BeautifulSoup(response.content, 'html.parser')

# Encontrar todos los elementos de título de noticias
news_titles = soup.find_all('h2', class_='news-title')

# Imprimir los títulos de las noticias
for title in news_titles:
    print(title.text)
```

Bibliografía:

- Baumer, EP (2017). Un ejemplo de flujo de trabajo de ciencia de datos: extracción de datos de la web. The Journal of Open Source Education , 20(3), 5.
- Lawson, R. y Adolfo, S. (2015). Raspado web con Python. "O'Reilly Media, Inc."
- Basta, C. y Bozzon, A. (2018). Uso de web scraping y análisis de sentimientos para investigar la cobertura de los medios italianos sobre los refugiados. Revista italiana de sociología de la educación, 10(1), 117-146.