Actividad Extracurricular 12

Anthony Contreras

Tabla de Contenidos

¿Qué es Web Scraping?

El **web scraping** se refiere al proceso de extracción de contenidos y datos de sitios web mediante software automatizado. Esta técnica se utiliza para recolectar información de manera eficiente, especialmente cuando se necesita extraer grandes volúmenes de datos.

El web scraping abarca una serie de prácticas que permiten obtener datos de la web de manera automática. Las aplicaciones del web scraping son diversas e incluyen áreas como la **investigación de mercado**, la **comparación de precios**, la **supervisión de contenidos** y la recopilación de cualquier tipo de información disponible en sitios web.

¿Cómo Funciona?

El proceso de scraping puede realizarse a través de diferentes herramientas y librerías en varios lenguajes de programación, siendo **Python** uno de los más populares gracias a sus librerías como **BeautifulSoup** y **Scrapy**.

El web scraping puede extraer desde motores de búsqueda, **feeds RSS**, hasta información de sitios gubernamentales. La mayoría de los sitios web permiten el acceso automatizado a través de scrapers y crawlers. Sin embargo, no todos los datos están siempre disponibles para

ser scrapados de manera sencilla. Dependiendo del sitio web y de las restricciones impuestas, puede que sea necesario usar trucos y herramientas específicas para obtener la información.

Por ejemplo, los scrapers web tienen dificultades para extraer datos significativos de **contenido visual**, como imágenes, o de elementos que requieren interacción con JavaScript.

Consideraciones Legales y Éticas

Es importante tener en cuenta las consideraciones legales y éticas cuando se realiza web scraping. Muchos sitios web tienen restricciones que limitan el acceso automatizado a sus contenidos, las cuales se encuentran en el archivo `robots.txt` de la página. Ignorar estas restricciones puede resultar en consecuencias legales o en la prohibición del acceso al sitio.

Pruebas con Python

```python
# Usando BeatifulSoup
import requests
from bs4 import BeautifulSoup

# URL del sitio web
url = 'https://www.epn.edu.ec'

# Realizamos la solicitud HTTP
response = requests.get(url)

# Si la solicitud es exitosa (status 200), continuamos
if response.status_code == 200:
    soup = BeautifulSoup(response.text, 'html.parser')

    # Extraemos los títulos del sitio
    titles = soup.find_all('h1')  # Busca todos los <h1> en la página

    for title in titles:
        print(title.text)
```

```
SÍGUENOS
LLÁMANOS
UBICACIÓN
```

```python
# Usando Scrapy
from requests_html import HTMLSession

# Crear una sesión de requests HTML
session = HTMLSession()

# URL del sitio web
url = 'https://www.epn.edu.ec'

# Realizar la solicitud HTTP al sitio web
response = session.get(url)

# Extraer todos los títulos de los encabezados <h1>
for title in response.html.find('h1'):
    print({'Title': title.text})

# Cerrar la sesión cuando termine
session.close()
```

```
{'Title': ''}
{'Title': 'SÍGUENOS'}
{'Title': 'LLÁMANOS'}
{'Title': 'UBICACIÓN'}
```

He creado un script basico en python el cual nos permite realizar varios análisis aún la sigo desarrollando pero aquí está el link https://github.com/7heAnsw3r/FireScrap

Referencias

https://kinsta.com/es/base-de-conocimiento/que-es-web-scraping/