Que es Scrapy?

Scrapy es una plataforma colaborativa de código libre que corre en Python para extraer datos de páginas web usado para una serie de aplicaciones como minería de datos, procesamiento de información o registro histórico. Scrapy no es una librería como pudiera parecer, sino un completo Framework de web scraping para Python.

A grandes rasgos, el funcionamiento de Scrapy es el siguiente:

• En primer lugar, el Engine recibe las peticiones (Requests) iniciales que le envía la araña (Spider), las programa en el scheduler y solicita las siguientes peticiones a rastrear.

• El Scheduler va enviando las peticiones a procesar al Engine y este a su vez las envía al componente Downloader.

• La página es descargada y se crea una respuesta (Response) de esta página que el Engine se encarga de enviar a la araña para que sea procesada.

• Es entonces cuando la araña o rastreador devuelve al Engine los ítems (Items) con la información extraída de la página.

• Después, estos ítems son enviados a los Pipelines para procesar y almacenar la información.

• El proceso se repite hasta que el Scheduler se quede sin peticiones.

Dado que es un framework, Scrapy tiene una serie de herramientas poderosas para hacer el "scraping" o extraer información de webs de manera fácil y eficiente.

<u>Ventajas:</u>

Una de las principales ventajas de este framework es que las peticiones se procesan y programan de manera asíncrona. Debido a ello, Scrapy no espera a que termine una petición para enviar otra y se pueden ejecutar a la vez de forma concurrente, acelerando el proceso en gran medida. Además, si una petición falla, el resto de las peticiones seguirán ejecutándose.

1. Ventajas de la programación orientada a objetos (POO).
2. Patrón de diseño Modelo Vista Controlador (MVC).
3. Trabajo con bases de datos, mapeo relacional de objetos(ORM), constructores de consultas, etc.
4. Trabajo con formularios y validación facilitado.
5. Enrutamiento y URLs amigables.
6. Seguridad.
7. Librerías y funcionalidades ya hechas.

Desventajas

Una de las mayores desventajas en los framework es la curva de Aprendizaje que en si puede ser muy exhustiva y cansadora a la hora de aprender todas su funciones.

1) Desconocimiento del funcionamiento del core del framework.
2) A mayor abstracción, menor rendimiento.
3) A veces, sufren muchos cambios de código entre versión y versión. sufren muchos cambios de código entre versión y versión.

Conclusion

En resumen, Scrapy es una herramienta valiosa para aquellos que necesitan extraer datos de la web de manera sistemática y eficiente, al tener todo un set de herramientas te habe muchisimas posibilidades.

Referencia:

Victor. (2018). Pros y contras de los frameworks de desarrollo web. *Victor Robles*. https://victorroblesweb.es/2018/10/31/pros-y-contras-de-los-frameworks-de-desarrollo-web/

De Los Datos, E. M. (2021, 19 febrero). *Extracción de datos de sitios web con Scrapy (I): recopilando información de productos de Zara*. El mundo de los datos. https://elmundodelosdatos.com/extraccion-datos-sitios-web-productos-zara/

Montoya, S. (2018). Extrae información de páginas web con Scrapy — gidahatari. *gidahatari*. https://gidahatari.com/ih-es/extrae-informacion-de-paginas-web-con-scrapy/

Edix, R. (2022b, julio 26). *Framework: qué es, para qué sirve y algunos ejemplos*. Edix España. https://www.edix.com/es/instituto/framework/