Evidencia de aprendizaje 1. Análisis y herramientas de extracción de datos

Estudiante: Andrés Camilo Arango Calle

Docente: Diego León Zapata Ruiz

Curso: Programación para Análisis de Datos

Grupo: PREICA2402B020100

Programa: Ingeniería de Software y Datos

Facultad: Ingenierías y Ciencias Agropecuarias

Institución Universitaria Digital de Antioquia

2024

Introducción

En el competitivo mundo del comercio electrónico, la información es poder. Conocer los precios de la competencia, la disponibilidad de productos y las tendencias del mercado puede marcar la diferencia para empresas y consumidores. Esta actividad se centra en la adquisición de dicha información a través del web scraping, utilizando las herramientas BeautifulSoup, Selenium y Scrapy para extraer datos de un artículo específico en una plataforma de comercio electrónico como Mercado Libre.

Se explorarán tres enfoques distintos para el web scraping. En primer lugar, se utilizará BeautifulSoup para analizar el código HTML del artículo, extrayendo información clave como nombre, precio y disponibilidad. A continuación, se empleará Selenium para interactuar con la página web de forma automatizada, simulando la navegación de un usuario real y capturando datos dinámicos que podrían no estar disponibles en el código fuente estático. Finalmente, se implementará Scrapy, un framework de web scraping que permite construir arañas web robustas y escalables para extraer datos de manera eficiente.

Cada técnica presenta ventajas y desafíos únicos. BeautifulSoup destaca por su simplicidad para analizar HTML, mientras que Selenium sobresale en el manejo de páginas web dinámicas. Scrapy, por su parte, ofrece un enfoque más completo y robusto para proyectos de web scraping a gran escala. Al finalizar la actividad, los participantes comprenderán las diferencias entre estas herramientas y podrán elegir la más adecuada según las necesidades de cada proyecto de web scraping.

Descripción de la página y artículo a analizar.

La página web es Mercado Libre, Mercado Libre es una plataforma de comercio electrónico líder en Latinoamérica, que permite a los usuarios comprar y vender una amplia variedad de productos, desde artículos deportivos como las camisetas del Atlético Nacional hasta electrónica, moda y hogar. Ofrece diversas funcionalidades como pagos en cuotas, envíos gratuitos y un programa de protección al comprador. Los usuarios pueden acceder a información detallada sobre los vendedores, incluyendo sus calificaciones y otros productos que ofrecen. Además, la plataforma utiliza algoritmos para sugerir productos relacionados que podrían interesar al comprador, como otras camisetas de fútbol o artículos deportivos.

El artículo que se analiza es una camiseta del Atlético Nacional, un equipo de fútbol colombiano. La camiseta está disponible en color verde musgo y viene en tallas M, L y XL. El precio es de $75,000 pesos colombianos y se ofrece envío gratuito y una política de devolución de 30 días. El vendedor es RICE5506661, quien tiene una buena calificación y un buen historial en la plataforma.

Descripción del tema de interés que deseas desarrollar en la primera práctica.

Este proyecto se enfoca en la extracción de información de un artículo específico en Mercado Libre, una camiseta del Atlético Nacional, utilizando técnicas de web scraping. Se busca aplicar y comparar tres herramientas populares en este campo: BeautifulSoup, Selenium y Scrapy.

Con BeautifulSoup, se analizará el código HTML de la página del producto para obtener datos como el nombre, precio y disponibilidad. Selenium permitirá interactuar con la página de forma automatizada, simulando la navegación de un usuario, para acceder a información dinámica. Por último, Scrapy, un framework de web scraping, se utilizará para construir una araña web que extraiga los datos de manera eficiente.

El interés en este proyecto surge de la necesidad de aprender y aplicar técnicas de web scraping, una habilidad cada vez más relevante en el análisis de datos y la inteligencia de mercado. Además, como hincha del Atlético Nacional, existe un interés personal en obtener información sobre la camiseta y los productos disponibles para el equipo.

Objetivos

- Adquirir conocimiento práctico sobre las diferentes herramientas y métodos de web scraping, incluyendo BeautifulSoup, Selenium y Scrapy, para extraer información de sitios web de forma automatizada.
- Comparar la eficiencia y las limitaciones de BeautifulSoup, Selenium y Scrapy al extraer datos de un sitio web de comercio electrónico, identificando las ventajas y desventajas de cada una en diferentes escenarios.
- Obtener información detallada sobre un producto específico (camiseta del Atlético Nacional) en Mercado Libre, incluyendo su precio, disponibilidad, características y vendedor, para comprender cómo se presenta la información en la plataforma.
- Investigar las prácticas de Mercado Libre en la presentación de productos, la gestión de vendedores y la interacción con los usuarios, utilizando el web scraping para obtener información sobre la estructura y el funcionamiento de la plataforma.

Metodología empleada de scraping.

El proyecto de web scraping se centró en la extracción de información de una camiseta del Atlético Nacional en Mercado Libre. Se implementaron tres bibliotecas de Python, cada una con un proceso específico:

BeautifulSoup:

1. Se utilizó la biblioteca requests para enviar una petición HTTP a la URL del producto y obtener el código HTML de la página.
2. Se creó un objeto BeautifulSoup para analizar el código HTML recibido.
3. Se utilizaron métodos para buscar elementos HTML específicos (título, precio, disponibilidad, etc.) basándose en sus etiquetas y atributos. Se extrajo el texto y los atributos relevantes de estos elementos.

Selenium:

1. Se configuró un navegador web (Chrome) en modo "headless" (sin interfaz gráfica) mediante webdriver.ChromeOptions().
2. Se utilizó el método driver.get(url) para cargar la página web del producto en el navegador.
3. Se emplearon métodos como find_element() para localizar elementos HTML específicos en la página, utilizando selectores CSS o XPath.
4. Se obtuvo la información deseada (texto, atributos) de los elementos encontrados.

Scrapy:

1. Se creó un nuevo proyecto de Scrapy con el comando scrapy startproject.

2. Se definió un Item para especificar los campos de datos a extraer (nombre, precio, vendedor, etc.).

3. Se creó una clase Python que define cómo rastrear la página web y extraer la información. Se especificaron las start_urls, el método parse() para procesar la respuesta, y los selectores CSS o XPath para identificar los elementos HTML relevantes.

4. Se extrajeron los datos de los elementos HTML seleccionados y se almacenaron en el Item.

5. Se ejecutó la araña con el comando scrapy crawl, que rastreó la página web, extrajo los datos y los guardó en un archivo (CSV, JSON, etc.).

Resultados y conclusiones.

- Se logró comprender y aplicar con éxito tres bibliotecas de Python para web scraping: BeautifulSoup, Selenium y Scrapy. Se aprendió a extraer información de sitios web, tanto estática como dinámica, utilizando diferentes técnicas y enfoques.
- Se observó que BeautifulSoup es eficiente para extraer información de páginas web con contenido estático y estructura HTML bien definida. Selenium, por otro lado, demostró ser útil para manejar páginas web dinámicas con JavaScript, mientras que Scrapy se destacó por su capacidad para construir arañas web robustas y escalables.
- Se extrajo información relevante de la camiseta del Atlético Nacional, incluyendo su nombre, precio, disponibilidad, vendedor y otros detalles relevantes. Esto demuestra la capacidad del web scraping para recopilar datos específicos de productos en sitios de comercio electrónico.
- Se pudo observar cómo Mercado Libre estructura la información de sus productos y cómo se gestionan las interacciones con los usuarios. El web scraping permitió obtener una visión interna del funcionamiento de la plataforma.
- Se comprobó la utilidad del web scraping como herramienta para la recopilación de datos en el análisis de mercado, la inteligencia empresarial y otras áreas donde se requiere información de la web.

Bibliografía

Palakollu, S. M. (3 de julio de 2019). *Scrapy Vs Selenium Vs Beautiful Soup for Web Scraping.* Obtenido de https://medium.com/analytics-vidhya/scrapy-vs-selenium-vs-beautiful-soup-for-web-scraping-24008b6c87b8

Tobella, P. (5 de agosto de 2021). *Introducción a las Técnicas y Herramientas de Web Scraping.* Obtenido de https://www.octoparse.es/blog/introduccion-a-las-tecnicas-y-herramientas-de-web-scraping