Extracción de datos de Artículos de Tecnología - BBC

Nombre: Carlos Esteban Trujillo Paz

Fecha: 12 de mayo de 2025

Curso: Ingeniería de Software y Datos

Materia: Programación para Análisis de Datos

Profesor: Andres Felipe Callejas

Repositorio de Github: EstebanAdso/BBC-scraper-Python

Introducción

Este proyecto realiza la extracción de datos de artículos de la sección de tecnología de la BBC, almacenando los resultados tanto en una base de datos SQLite como en archivos CSV. El objetivo principal es mantener un registro histórico de noticias tecnológicas, evitando la inserción de duplicados y asegurando la organización de los archivos generados.

Descripción de la página y artículo a analizar

La página seleccionada para este análisis es la sección de tecnología del sitio BBC (https://www.bbc.com/news/technology), debido a su relevancia y constante actualización de noticias en el ámbito tecnológico. Esta elección permite obtener un listado de artículos de manera sencilla y rápida para su posterior análisis y almacenamiento.

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

El tema de interés está enfocado en automatizar el proceso de recopilación de noticias tecnológicas actuales, con el propósito de construir un repositorio histórico de eventos relevantes en tecnología. Esto puede ser útil para análisis de tendencias, estudio de evolución tecnológica o desarrollo de proyectos de análisis de datos.

Objetivos

Formular los objetivos a partir de la siguiente pregunta: ¿por qué deseas analizar este artículo y la empresa de comercio?

- Analizar el sitio de noticias de la BBC por su confiabilidad y cobertura global en temas tecnológicos.

- Obtener artículos actualizados de forma automatizada y organizada.

- Desarrollar habilidades prácticas en extracción de datos, almacenamiento de datos y procesamiento de archivos.

Metodología empleada de extracción de datos

Para realizar la extracción de datos se utilizó el lenguaje Python en el entorno de desarrollo Visual Studio Code. Se utilizaron las siguientes bibliotecas y recursos:
- requests
- BeautifulSoup
- sqlite3
- csv

Se creó una clase para manejar los artículos (modelo Article), una clase para gestionar la base de datos (DatabaseManager), y otra para manejar los archivos (FileManager). Los datos fueron almacenados en una base SQLite y exportados a archivos CSV.

Resultados y conclusiones

Como resultado se logró obtener múltiples artículos de la sección de tecnología, los cuales fueron correctamente almacenados en la base de datos y exportados a archivos CSV sin duplicación. Se evidenció que la extracción de datos es una herramienta útil para la automatización de recolección de datos, y se resaltó la importancia de organizar adecuadamente los datos obtenidos. Además, aprendí a utilizar GitHub Actions, una herramienta poderosa de integración continua, que implementé tanto para gestionar artifacts como para realizar acciones automáticas con commits, mejorando así el flujo de trabajo del proyecto.

Bibliografía (Normas APA)

- BBC. (2025). Technology. BBC News. https://www.bbc.com/news/technology GitHub Docs. (n.d.). *Understanding GitHub Actions*. GitHub.
- https://docs.github.com/en/actions/learn-github-actions/understanding-github-actions
- Tech With Tim. (2021, August 12). *BeautifulSoup Web Scraping Tutorial with Python* [Video]. YouTube. https://www.youtube.com/watch?v=87Gx3U0BDlo