Extracción de datos de Artículos de BBC usando Github Actions con automatización

Nombre: Carlos Esteban Trujillo Paz

Fecha: 24 de mayo de 2025

Curso: Ingeniería de Software y Datos

Materia: Programación para Análisis de Datos

Profesor: Andrés 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 datos duplicados y asegurando la organización de los archivos generados utilizando también Github Actions para obtener datos actualizados en cada commit de la aplicación.

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

Se desarrolló un proyecto de extracción de datos que permite obtener automáticamente información actualizada de la sección de tecnología del portal BBC News. Este ejercicio tiene como propósito aplicar técnicas básicas de extracción de datos para capturar títulos y enlaces de los artículos más recientes publicados en dicho sitio web, y almacenarlos de forma estructurada

Además, se integró GitHub Actions para automatizar el proceso, de modo que cada vez que se realice un commit en el repositorio, se ejecuta automáticamente el script de extracción de datos y se obtienen los datos más recientes en ese momento. Esto permite mantener actualizada la información sin intervención manual, implementando un flujo básico de integración y despliegue continuo.

Objetivos

- Extraer los datos del 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 y almacenamiento de datos.

- Utilizar Git y GitHub como herramientas de control de versiones para el desarrollo del proyecto.

- Configurar un pipeline de integración y despliegue continuo (CI/CD) con **GitHub Actions**, que permita ejecutar la aplicación de extracción de datos de forma automática en cada commit.

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.

Configuración del Workflow en GitHub Actions

El archivo de configuración del flujo (.github/workflows/accionables.yml) realizamos los siguientes pasos:

Trigger del workflow: se activa automáticamente con cada push a la rama main.

Checkout del repositorio: descarga el contenido del repositorio para trabajar con él.

Instalación de Python: configuramos la versión 3.12 para el entorno de ejecución.

Creación y activación de un entorno virtual: para aislar las dependencias.

Instalación de dependencias: a partir del archivo requirements.txt del repositorio.

Ejecución del script principal (main.py): que realiza la aplicación de extracción de datos, guarda los datos en SQLite y genera archivos CSV.

Resultados y conclusiones

Como resultado de esta segunda fase del proyecto, se logró automatizar completamente el proceso de extracción de datos desde la sección de tecnología de BBC News. Los artículos fueron capturados correctamente, almacenados en una base de datos SQLite y exportados a archivos CSV de manera estructurada y sin duplicaciones.

Uno de los principales logros fue la integración exitosa de GitHub Actions como herramienta de automatización. Gracias a su configuración, el sistema ejecuta el script automáticamente en cada commit, generando los archivos actualizados y versionándolos directamente en el repositorio. Esto permitió establecer un flujo de trabajo continuo, en el que los datos siempre se mantienen sincronizados y disponibles sin necesidad de intervención manual.

Durante este proceso, se consolidaron conocimientos clave sobre CI/CD, el manejo de entornos virtuales, la gestión de paquetes en Python, y la automatización de tareas dentro de un entorno DevOps. Además, se comprobó la efectividad de la aplicación de extracción de datos automatizado como herramienta para la recolección periódica y organizada de información útil, aplicable en contextos reales de análisis de datos.

Bibliografía

- 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