Extracción de datos de Artículos de BBC usando Docker Hub con automatización

Nombre: Carlos Esteban Trujillo Paz
Materia: Programación para Análisis de Datos
Carrera: Ingeniería de Software y Datos
Profesor: Andrés Felipe Callejas

Repositorio de GitHub: https://github.com/EstebanAdso/BBC-scraper-Python
Imagen en Docker Hub: https://hub.docker.com/r/estebanadso/bbc-scraper

08 de junio de 2025
Institución Universitaria Digital de Antioquia

Introducción

Este proyecto tiene como propósito automatizar la extracción de datos de artículos publicados en la sección de tecnología del portal BBC News, utilizando herramientas modernas como Docker y GitHub Actions. Los datos recolectados se almacenan tanto en una base de datos SQLite como en archivos CSV. La automatización busca evitar la duplicación de datos y facilitar el mantenimiento de un historial de artículos tecnológicos actualizado, haciendo uso de flujos de integración y despliegue continuo (CI/CD).

El nuevo enfoque basado en contenedores permite que el proceso sea portable, escalable y controlado, reduciendo errores por diferencias en entornos locales. Al encapsular la aplicación en una imagen Docker, es posible ejecutarla automáticamente mediante GitHub Actions en cada nuevo commit al repositorio.

Descripción de la página y artículo a analizar

El sitio seleccionado para extracción de datos es la sección de tecnología del portal de noticias BBC (https://www.bbc.com/news/technology), reconocido mundialmente por su cobertura actualizada y confiable de temas tecnológicos. Este apartado permite acceder fácilmente a títulos y enlaces de las noticias más recientes, lo que lo convierte en una fuente ideal para pruebas de extracción de datos automatizadas.

Descripción del tema de interés

El proyecto busca aplicar técnicas de extracción de datos para capturar títulos y enlaces de los artículos más recientes en la sección de tecnología de BBC News. La información recolectada se almacena de forma estructurada en archivos .csv y en una base de datos SQLite, permitiendo su análisis posterior.

Además, se implementó un flujo de integración continúa usando GitHub Actions, de tal forma que la recolección se ejecute automáticamente cada vez que se realice un commit. Esto elimina la necesidad de intervención manual, facilita el control de versiones y garantiza que los datos estén siempre actualizados.

Objetivos

- - Extraer artículos de la sección de tecnología de BBC News de forma automatizada.
- - Almacenar los datos en una base SQLite y en archivos CSV.
- - Desarrollar competencias prácticas en extracción de datos.
- - Automatizar el flujo de ejecución del proyecto mediante Docker y GitHub Actions.

- - Aplicar principios de CI/CD en proyectos de análisis de datos.

Herramientas y tecnologías utilizadas

- - Lenguaje de programación: Python 3.12
- - Librerías: requests, BeautifulSoup, sqlite3, csv
- - Contenedores: Docker
- - Automatización: GitHub Actions
- - Repositorio: GitHub
- - Almacenamiento de imagen: Docker Hub
- - Editor: Visual Studio Code

Metodología

Extracción de datos

La aplicación consta de tres componentes principales:

- - Un modelo de datos (Article) que representa cada noticia.
- - Un gestor de base de datos (DatabaseManager) que controla el almacenamiento en SQLite.
- - Un manejador de archivos (FileManager) que exporta los datos a CSV.

Los datos son recolectados mediante solicitudes HTTP y procesados con BeautifulSoup.

Automatización con Docker y GitHub Actions

- Dockerización del proyecto: Se creó un Dockerfile que contiene el entorno necesario para ejecutar el script principal (main.py).
- Subida a Docker Hub: Se configuró el repositorio en GitHub para construir y publicar automáticamente la imagen en Docker Hub (estebanadso/bbc-scraper).
- Configuración del workflow en GitHub Actions:
- - Activación del workflow mediante eventos push en la rama main.
- - Autenticación en Docker Hub usando secretos del repositorio.
- - Construcción de la imagen Docker y su ejecución dentro del workflow.
- - Montaje de un volumen para almacenar los archivos .csv y .db.
- - Uso de git-auto-commit-action para subir los archivos generados al repositorio.

Resultados y conclusiones

Gracias al uso de contenedores y la automatización con GitHub Actions, se logró implementar un sistema que extrae artículos automáticamente, los almacena y los

publica en cada commit al repositorio. El proceso se ejecuta en un entorno Docker controlado, lo que asegura portabilidad y consistencia.

Este enfoque representa una mejora sustancial frente a la ejecución manual o mediante ambientes virtuales, ya que permite una ejecución más robusta y sin dependencias externas. También se aplicaron buenas prácticas DevOps, como CI/CD, uso de secretos y control de versiones.

Este proyecto ha permitido reforzar competencias clave como el uso de contenedores, automatización de procesos, manejo de flujos de trabajo con GitHub Actions, y extracción de datos con Python. La solución es escalable, reutilizable y aplicable a otros contextos de análisis de datos.

Referencias (formato APA 7ª edición)

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