GBA Crawler

1. Project Overview

This project implements a Python-based news crawler system aimed at automating the collection and analysis of articles from major news sources in the Guangdong-Hong Kong-Macao Greater Bay Area. The system identifies important news related to social issues, economics, and livelihoods using keywords and event triggers, and generates daily reports.

2. Key Features

Multi-source Crawling: Supports information gathering from news websites in Guangzhou, Shenzhen, Zhongshan, Jiangmen, and more.

Content Analysis: Identifies significant news events through keyword matching and date detection.

Deduplication Mechanism: Uses hash values to check for content duplication in news articles.

Scheduled Operation: Automates execution to ensure regular updates of news data.

Report Generation: Produces HTML formatted reports for easy reference and analysis.

3. Technologies Used

Web Requests: Uses the requests library for webpage fetching.

HTML Parsing: Utilizes BeautifulSoup for content extraction.

Chinese Language Processing: Employs jieba for word segmentation and part-of-speech tagging.

Task Scheduling: Implements APScheduler for managing scheduled tasks.

Date Parsing: Uses the dateutil library to handle date formats.

4. Future Outlook

Future enhancements could include:

Expanding to more news sources to cover additional regions.

Integrating machine learning models for sentiment analysis of articles.

Providing an API interface to allow other systems to access crawled news data.