Weather Scaper Document

1. Giới thiệu:

* **weather\_scraper** là một tool để thu thập dữ liệu thời tiết từ nguồn <https://www.accuweather.com/vi/vn/ho-chi-minh-city/353981/current-weather/353981?unit=c>, sử dụng các công cụ hỗ trợ như GitHub Actions Workflow, thư viện jsoup.
* **Mục đích:**
* Lấy dữ liệu thời tiết theo ngày, tại 1 khung giờ nhất định
* Xuất dữ liệu ra file weather\_log để lưu trữ
* Dùng dữ liệu để sử dụng cho dự án học tập
* **Ngôn ngữ:** Java
* **GitHub của project:** <https://github.com/BaoTam368/weather_scraper>

1. Các công cụ hỗ trợ

* **Jsoup**: là một thư viện Java dùng để làm việc với HTML thực tế. Nó cung cấp một API rất tiện lợi để truy xuất URL, trích xuất và thao tác dữ liệu,
* **Github action**: là công cụ tích hợp sẵn trong github hỗ trợ tự động hóa chạy script định kỳ(như cron job)

1. Cấu trúc thư mục và các thành phần

weather\_scraper/

├── .github/

│ └── workflows

│ └── weather.yml ← file yml tập hợp các bước mà │ GitHub sẽ chạy tự động khi có sự kiện │ xảy ra

├── .settings/

├── src/

│ └── main/

│ └── java/

│ └── weather\_scraper ← source code chính

├── target/

├── pom.xml

├── README.md

└── weather\_log.txt ← log để lưu trữ

* **Giải thích :**
* ***src/main/java/weather\_scraper:*** nơi chứa class chính của project
* ***pom.xml***: file cấu hình Maven, liệt kê dependencies, plugins, cấu hình build
* ***weather\_log.txt:*** file log dữ liệu output
* ***target/:*** thư mục output sau khi build
* ***.github/:*** tự động hoá quá trình chạy
* .***settings/:*** cấu hình project

1. Chức năng chính:

* Lấy dữ liệu thời tiết từ trang web (sử dụng Jsoup).
* Trích xuất các thông tin quan trọng (ví dụ: nhiệt độ, thời gian, tình trạng thời tiết).
* Ghi dữ liệu vào file log.

1. Quá trình chạy

* Chương có URL mặc định là: <https://www.accuweather.com/vi/vn/ho-chi-minh-city/353981/current-weather/353981?unit=c>
* File yml sẽ chạy theo schedule đã cài đặt sẵn là mỗi ngày là 8H30 sáng (UTC – 7)
* Quá trình chạy của File yml:

1. Lấy code từ repo
2. Cài đặt JDK
3. Build project bằng Maven
4. Chạy class Java chính (weather\_scraper.java)
5. Commit và push file weather\_log.txt trở lại repo

* Quá trình chạy class Java chính:

1. Gửi request đến URL.
2. Lấy nội dung HTML.
3. Trích xuất dữ liệu thời tiết.
4. Ghi vào weather\_log.txt với định dạng:

Thời gian| Nhiệt độ | UV | Gió | Độ ẩm| Điểm sương | Khí áp | Mây| Tầm nhìn | Trần mây

1. Những hạn chế hiện tại

* Phụ thuộc vào cấu trúc HTML của trang (nếu thay đổi sẽ lỗi).
* Mới chỉ hỗ trợ 1 địa chỉ cụ thể.