Bộ quét Đọc Báo Crawler

**TÀI LIỆU CHO DEVELOPER**

**A.Giới thiệu chung**

Bộ quét Đọc Báo Crawler là bộ quét hiện được sử dụng cho toàn bộ các sản phẩm của Bảng Tin. Bộ quét này được thiết kế với một số nguyên tắc:

* Toàn bộ cấu hình quét được lưu trữ dưới dạng plain text (YAML) để dễ dàng lưu trữ, chia sẻ.
* Quá trình quét diễn ra theo chu kỳ (vd: 10p, 15p…) được thiết lập thông qua crontab
* Quét đa luồng (multiprocess) để tăng tốc độ quét và tối ưu phần cứng server.
* Thuật toán quét tổng quát được cấu hình thông qua file config. Chỉ cần dùng file config là có thể quét được báo chí, website, Facebook.
* Sử dụng Selenium+Firefox để quét các trang cần Login (vd Facebook) hoặc các trang dùng AJAX để nạp dữ liệu
* Tích hợp sẵn một số cơ chế để tránh bị block: vd giãn cách thời gian giữa các request, dùng browser quét để qua mặt các trang load AJAX
* Tích hợp sẵn bộ lọc bài trùng (dựa trên title, href, content)
* Tích hợp sẵn API server để lấy dữ liệu qua API
* Tích hợp sẵn các thuật toán phân tích, thống kê keyword để xác định trend
* Tích hợp output dữ liệu ra một số kênh phổ biến: file json, redis, rabbitmq, wordpress, elasticsearch
* Tích hợp cùng frontend để hiển thị dữ liệu (tham khảo: <https://theodoibaochi.com>), cùng tool CLI để thao tác dữ liệu qua terminal (https://github.com/hailoc12/docbao-research-cli)

Ưu điểm:

* Kiến trúc monolith tương đối đơn giản, triển khai dễ dàng lên một server
* Quét được chi tiết nội dung bài viết
* Chỉ cần cấu hình là có thể quét được 80-90% các website/blog tin tức trên Internet
* Tích hợp nhiều know-how liên quan tới qúa trình quét. VD: tự động parse chính xác định dạng date time trên nhiều website, chống block
* Test bug, chạy ổn định trong thời gian dài

Nhược điểm:

* Khó scale ra nhiều server khác nhau
* Quét theo chu kỳ cố định, do đó chưa tận dụng tối đa thời gian sử dụng server. Nếu để chu kỳ quá thấp, có thể xảy ra tình trạng quét lặp, quét trùng.
* Chưa thể quét realtime
* Mỗi bài báo chỉ được quét một lần, do đó chưa thể cập nhật dữ liệu dạng timeseries.

Một số tài liệu bổ sung nên đọc:

[Cơ chế quét + hướng dẫn config bằng tay](http://danghailoc.com/post/2019-02-11_huong_dan_config_docbao/)

[Hướng dẫn tạo config sử dụng tool](https://drive.google.com/file/d/1sgTb5Fv8hgrAfWGU3PYrMWQNfOIUzovv/view?usp=sharing)

**B. Cấu trúc file và các đối tượng quan trọng**

- frontend: folder chứa toàn bộ file cần thiết để tạo dashboard hiển thị dữ liệu từ backend (giống trang https://theodoibaochi.com)

- backend: folder chứa toàn bộ dữ liệu liên quan tới backend

+ crawl.py: file khởi động quá trình quét. Có thể setting các kênh đẩy dữ liệu qua file này

+ input/config.yaml: file chứa toàn bộ cấu hình phục vụ cho quá trình quét. Có thể edit file này qua tool config\_manager (giới thiệu bên dưới)

+ data/: folder này chứa toàn bộ các file database. Bản chất là pickle các đối tượng Data Manager, Keyword Manager

+ lib/: folder chứa toàn bộ các thư viện của dự án

* + docbao\_lib.py: chứa đối tượng Docbao\_Crawler, là đối tượng thể hiện logic quét chính của toàn bộ ứng dụng. Đọc file này là hiểu được tổng quan qúa trình chạy của app
  + data.py: chứa mô tả class Article, là đối tượng chứa nội dung bài quét. Ngoài ra còn có class ArticleManager, là class đóng vai trò như database và cung cấp các utility liên quan tới Article
  + keyword.py: chứa mô tả class Keyword, là đối tượng thể hiện từ khoá. Noga2i ra còn chứa class KeywordManager, là class đóng vai trò như database về keyword và cung cấp các utility liên quan tới Keyword
  + config.py: chứa mô tả class WebConfig, là đối tượng đại diện cho config để quét một trang xác định, ngoài ra còn có ConfigManager, là đối tượng chứa config chung của toàn bộ ứng dụng
  + utils.py: chứa các hàm utility sử dụng trong các thư viện khác.
  + browser\_cralwer: wrap chức năng của selenium+firefox, để hỗ trợ quét sử dụng trình duyệt.
  + cdn.py: chứa các hàm dùng để convert ảnh & thay thế link ảnh gốc bằng link ảnh đã upload lên CDN của Bảng Tin
  + redis\_database.py, wordpress.py, elasticsearch\_client.py, rabbitmq\_client.py: các file hỗ trợ đẩy dữ liệu từ docbao crawler lên các database tương ứng.