Cập nhật tiến độ Tuần 2

Triển khai các module theo kiến trúc microservices dùng giao tiếp thông qua Kafka, API

1. Frontier
    - Tham khảo luồng chạy lấy robots.txt của crawler4J. Customize lại để thu gọn các bước phù hợp với phần này. Ứng dụng lấy rules trong đó để cho phép đưa vào queue xử lý
    - Đã xử lý được thứ tự thực hiện url theo thiết kế
2. Fetcher
    - Xử lý nhận message từ Frontier gửi đến, thực hiện fetch lấy nội dung trang, kiểm tra content type, cập nhật domain metadata bằng api
    - Chuyển html qua kafka với thuật toán nén zstd
    - Bỏ lưu html ở S3, chuyển sang lưu luôn ở mongoDB, hiện tại chưa sử dụng, nhưng sau này có thể dùng để check dup (theo ngưỡng để xem có nên crawl lại không), hoặc dùng để check version của bài viết (hiện tại chưa tính đến)
3. Parser
    - Setup 2 bộ lọc Heuristics để kiểm tra trang hiện tại có cần parse nội dung không.
    - Triển khai 2 loại:
        - Content + Url Extracting (Chưa thực hiện với llm)
        - Sitemap Extracting: Dùng cho mục đích quét lấy category url, tránh phải tự điền seed url thủ công
    - Sử dụng 2 bộ lọc + Cache Aside cho việc kiểm tra xem có đưa url mới extract vào Frontier không.
4. Stroring

Mới set, nhưng chưa thực hiện lưu do chưa setup Kafka ở đây

5. Monitoring

(Grafana + Prometheus) Chưa thu thập đúng yêu cầu

6. Search API (Chưa làm)

Trong tuần tới:

- Triển khai thực hiện lấy config, lưu config cho một số domain (dự định dùng Gemini API)
- Stroring + Search API
- Expose metric thống kê (số url nạp vào frontier, số url là article / tổng số nhận vào, số url mới extract được, số url parse nội dung bị lỗi)