Project 01

Thu Thập Dữ Liệu Từ Web

1. Mô tả

Để thực hiện việc phân tích dữ liệu từ web, bước đầu tiên thường tập trung vào việc thu thập dữ liệu. Trong khuôn khổ Project này, chúng ta cùng thực hiện việc trích xuất dữ liệu từ web theo chủ đề.

2. Yêu cầu

Nhóm SV cần thực hiện các công việc sau:

- Xác định chủ đề cần lấy (không lấy hết dữ liệu trong 1 trang, có thể nhiều hơn 1 chủ đề)
- Thu thập dữ liệu theo đúng chủ đề từ ít nhất hai website trở lên
- Tạo cấu trúc dữ liệu phù hợp để lưu trữ
- Tự thiết kế crawler, không sử dụng các crawler sẵn có
- Tránh các vấn đề xảy ra đối với crawler đã được nêu trong lý thuyết
- Thể hiện sự phức tạp của crawler như lấy dữ liệu từ trang web động, lấy qua API, độc lập với cấu trúc của website, cho phép người dùng chọn chủ đề, …
- Đo tốc độ của crawler dựa trên thống kê số trang (webpage) xử lý trong đơn vị thời gian, kích thước dữ liệu lấy được theo chủ đề, thời gian lấy, …
- Đánh giá dữ liệu thu được như thể hiện các thống kê về dữ liệu lấy như kích thước, loại, đặc trưng, phân phối, tính đa dạng của dữ liệu…
- Tiền xử lý dữ liệu

3. Đánh giá

STT	Tiêu chí	Tỉ lệ
1	Trình bày chủ đề, lý do chọn chủ đề, trang web lấy hoặc trang web là hạt giống.	10%
2	Mô tả thuật toán, cấu trúc mã nguồn, các thành phần hệ thống	30%
3	Các vấn đề xảy ra đối với crawler và phương pháp xử lý	10%
4	Các tính năng phức tạp của crawler	10%
5	Đánh giá hiệu năng crawler	5%
6	Mô tả và đánh giá dữ liệu thu thập được.	5%
7	Tiền xử lý dữ liệu thu thập.	10%
8	Báo cáo rõ ràng các mục đã thực hiện, có thể hiện mức độ hoàn thiện của từng công việc	20%
Tổng		100%

4. Lưu ý

- Làm việc nhóm, mỗi nhóm tối đa 4 thành viên
- Thời gian: 3-4 tuần
- Nộp báo cáo trên Moodle
- Mã nguồn và dữ liệu thu thập được tải lên Google Drive/One Drive, mở quyền truy cập và đính kèm trong báo cáo.
- Ngôn ngữ tùy chọn nhưng khuyến khích Python. Một số thư viện có thể được sử dụng nhưng không liên quan trực tiếp đến crawler.
- Cấu trúc báo cáo cần bao gồm thông tin:
  - MSSV, họ tên từng thành viên
  - Phân công và kế hoạch thực hiện
  - Báo cáo theo các mục ở phần 3
  - Hướng dẫn cách thức biên dịch và chạy chương trình
  - Tài liệu tham khảo (nếu có)
- Đạo văn sẽ 0 điểm môn học