Đồ án cuối kỳ môn Nhập môn khoa học dữ liệu

THÔNG TIN NHÓM

STT: 12

THÀNH VIÊN:

- Vũ Đăng Hoàng Long (18120203)
- Nguyễn Huỳnh Đại Lợi (18120198)

PHÂN CÔNG CHÍNH

Long

- Quản lý các báo cáo.
- Tiếp nhận các kết quả truyền thông từ nhóm, sau đó tìm hiểu các hướng để giải quyết vấn đề và gợi ý thảo luận trong nhóm.
- Thực hiện phân công.

Lợi

- Tìm hiểu các vấn đề liên quan đến kỹ thuật (code).
- Tiếp nhận phân công và truyền thông lại kết quả thu được.

QUÁ TRÌNH LÀM VIỆC

15/01/2021

Hoàn thành.

13/01/2021

- Thầy yêu cầu nộp đồ án vào ngày 14/01/2021.
- Thử nghiệm xong mô hình.

Phân công

Tất cả cùng hoàn thành báo cáo và slide để nộp.

12/01/2021

- Hoàn thành chuẩn hóa toàn bộ dữ liệu.
- Thử nghiệm tiền xử lý thành công.

Phân công

- Lợi: xây dựng mô hình học hôm sau.
- Long: tiến hành báo cáo phần khám phá và tiền xử lý.

10/01/2021

- Lợi sửa được laptop.
- Task tiền xử lý khá chậm chạp.

Phân công

- Lợi: tiếp tục tìm hiểu xây dựng mô hình học.
- Long: tối ưu khâu tiền xử lý để có thể xử lý nhanh chóng toàn bộ dữ liệu.

08/01/2021

- Lợi bị hư laptop.
- Hoàn tất thu thập báo (hơn 800.000 bài).

Phân công

Tạm thời Long sẽ tiếp quản toàn bộ công việc của Lợi.

SÁNG 07/01/2021

- Hoàn tất khắc phục xong các vấn đề cản trở khai thác dữ liệu.
- Hoàn thành notebook phần khai thác dữ liệu.

CHIỀU 06/01/2021

Tạm thời hoàn tất khai phá.

Phân công

- Long: tiếp quản quá trình khai phá và tiền xử lý.

- Lợi: chuyển sang thử nghiệm xây dựng model.

CHIỀU 05/01/2021

- Hoàn thành khai thác được khoảng 47.000 bài báo.
- Câu hỏi trước đó thiếu tính ứng dụng do hầu hết người viết báo ra đều thực hiện gắn nhãn cho bài báo rồi mới đăng tải.

Quyết định:

Thay đổi câu hỏi thành: Với một đoạn hội thoại bất kỳ, làm sao để biết đoạn hội thoại này nói về chủ đề gì?

Phân công:

- Long: tiếp tục thực hiện khai thác hàng loạt. Tiến hành viết báo cáo (hơi muộn vì trước đó phải mất thời gian khám phá).
- Lợi: tiến hành khai phá dữ liệu và truyền thông lại kết quả khai phá cho Long (quy trình khai thác) để khắc phục các vấn đề trong dữ liệu khai thác được.
- Sau khi khắc phục xong quá trình khai thác dữ liệu, cả nhóm sẽ chuyển qua tiền xử lý và tìm hiểu các kiến thức cần thiết để có thể phân tích dữ liệu này.

SÁNG 05/01/2021

Hoàn thành xây dựng code để khai thác dữ liệu từ nguồn báo Tuổi Trẻ Online (https://tuoitre.vn/). Tuy nhiên cách khai thác cũ sử dụng Selenium để khai thác thiếu hiệu quả và tiêu hao quá nhiều tài nguyên máy nên đã phải tìm hiểu lại cách mới.

Phân công:

- Long: thực hiện thay đổi phương thức khai thác để hiệu quả hơn và treo máy để khai thác dữ liệu qua đêm.
- Lợi: tạm thời nghỉ ngơi, chờ nhận dữ liệu để chuyển qua bước khai phá dữ liệu và tiền xử lý.

04/01/2021

Nhận xét quá trình khám phá:

- Thay đổi góc nhìn để tìm câu hỏi: từ góc nhìn "Mình có thể giải quyết vấn đề gì?" sang góc nhìn "Là một người dùng web phổ thông, mình có vấn đề gì có thể giải quyết?". --> Với góc nhìn này, các câu hỏi đặt ra sẽ phần lớn phục vụ cho nhu cầu của người dùng hơn là nhu cầu của doanh nghiệp hay nhu cầu vĩ mô của toàn xã hội.

- Dữ liệu dạng chữ rất dồi dào (báo mạng, mạng xã hội, blog,...).

Quyết định:

Khai thác dữ liệu báo mạng vì đây là nguồn dữ liệu có chuyên môn và sử dụng ngôn ngữ chuyên nghiệp, bên cạnh đó các bài báo đều có chủ đề để có thể phục vụ cho tác vụ phân lớp.

Câu hỏi:

Làm sao để biết một bài báo bất kỳ thuộc chủ đề gì?

Phân công:

- Lợi: tìm hiểu các vấn đề thiên về kỹ thuật (code). Nhiệm vụ trước mắt là tìm hiểu và khai thác dữ liệu báo mạng.
- Long: quản lý các báo cáo, tiếp nhận các kết quả truyền thông từ Lợi, sau đó tìm hiểu các hướng để giải quyết vấn đề và thực hiện phân công. Nhiệm vụ trước mắt là hỗ trợ Lợi trong khai thác dữ liệu báo mạng.

01/01/2021

Nhận xét quá trình khám phá trước đó:

- Rất khó để có thể tìm được data giải quyết các vấn đề mà nhóm nghĩ ra (môi trường, y tế, giao thông, thời tiết,...).
- Các data dạng bảng tính thường chỉ tìm được tại các cổng lưu trữ dữ liệu. Tuy nhiên loại này thì phải tải về chứ không phải sử dụng api hoặc parse html.

Phân công:

Tiếp tục tìm hiểu, mail hỏi ý kiến giáo viên xem sử dụng loại dữ liệu này được không.

26/12/2020

Mới nhận thông tin đồ án từ giáo viên. Phân công tất cả thành viên khám phá những thông tin sau:

- Dữ liệu có thể thu thập được.
- Các câu hỏi có khả năng trả lời được.