ĐỒ ÁN THỰC HÀNH
CSC17104 – LẬP TRÌNH CHO KHOA HỌC DỮ LIỆU

1 TỔNG QUAN

Trong môn học này với các bài labs, sinh viên đã được cung cấp những kiến thức và kỹ năng về làm thế nào để sử dụng các công cụ khoa học dữ liệu như: các lệnh Linux, cách sử dụng Git và Github, Conda, Jupyter Notebook, Markdown, ngôn ngữ lập trình Python cùng với một số thư viện hỗ trợ như Matplotlib, Numpy, Pandas. Thông qua các bài tập thực hành, sinh viên đã học cách sử dụng các công cụ này ở một cấp độ đủ sâu, và đã áp dụng chúng trong một quy trình khoa học dữ liệu từ đưa ra câu hỏi có ý nghĩa cần trả lời, thu thập dữ liệu (trong phạm vi của môn học này: dùng dữ liệu có sẵn, hoặc có thể thu thập một cách tương đối là dễ dàng), khám phá dữ liệu, tiền xử lý dữ liệu, phân tích dữ liệu bằng cách phân tích đơn giản với các tính toán và trực quan hóa để trả lời cho các câu hỏi về thông tin *nằm trong* dữ liệu, và truyền thông các kết quả thu/ ra quyết định.

Để thực hiện tốt một quy trình này, một nhà khoa học dữ liệu cần phải:

- Có kiến thức sâu sắc, sử dụng thành thạo các công cụ hỗ trợ trên máy tính.
- *Luôn luôn bình tĩnh, khách quan, và thành thật.*

Như là một phần không thể thiếu của một môn học, đó chính là **đồ án thực hành**. Đồ án thực hành sẽ được thực hiện trong 4-5 tuần (tùy tình hình), và tiến hành theo hình thức làm việc nhóm. Các phần tiếp theo sẽ trình bày chi tiết về nội dung, cách thực hiện, yêu cầu, và hình thức bảo vệ đồ án.

2 NỘI DUNG ĐỒ ÁN THỰC HÀNH

Sinh viên thực hiện tìm kiếm dữ liệu đã được công khai, ví dụ như trên **Kaggle** về một chủ đề mà nhóm sinh viên quan tâm, hứng thú. Sinh viên thực hiện một quy trình khoa học dữ liệu đối với dữ liệu được chọn: khám phá dữ liệu, xác định các câu hỏi có thể trả lời được bằng dữ liệu, tiền xử lý dữ liệu và phân tích để trả lời cho mỗi câu hỏi. Dữ liệu phải được cấu trúc hóa thành một bảng gồm ít nhất 5 thuộc tính (trường dữ liệu) và 1000 dòng (records).

A. Thu thập dữ liệu

- Dữ liệu mà nhóm sinh viên là về chủ đề gì và được lấy từ nguồn nào?
- Người ta có cho phép sử dụng dữ liệu như thế này hay không? Ví dụ: cần kiểm tra thử License của dữ liệu là gì?
- Người ta đã thu thập dữ liệu này như thế nào?

Lưu ý: Không được phép sử dụng dữ liệu đã thu thập được ở môn học **Nhập môn Khoa học Dữ liệu**. Nếu phát hiện, đồ án thực hành sẽ được **điểm 0**.

B. Khám phá dữ liệu (thường đan xen với pha tiền xử lý dữ liệu)

- Mỗi dòng có ý nghĩa gì? Có vấn đề các dòng có ý nghĩa khác nhau không?
- Mỗi cột có ý nghĩa gì?
- Mỗi cột hiện đang có kiểu dữ liệu gì? Có cột nào có kiểu dữ liệu chưa phù hợp để có thể xử lý tiếp hay không?
- Với mỗi cột, các giá trị được phân bố như thế nào?

C. Đưa ra các câu hỏi có ý nghĩa cần trả lời

Nhóm sinh viên cần phải đưa ra ít nhất từ 4-5 câu hỏi mà có thể trả lời bằng dữ liệu. Tất cả các câu hỏi phải có ý nghĩa (Lợi ích của việc tìm ra câu trả lời là gì?). Và bạn cần phải chú ý đến độ khó của câu hỏi, không nên quá dễ.

Trong tập tin notebook, với mỗi câu hỏi, sinh viên cần trình bày:

- Câu hỏi là gì?
- Nếu trả lời được câu hỏi thì sẽ có lợi ích gì?

D. Tiền xử lý và phân tích dữ liệu để trả lời cho từng câu hỏi

Với mỗi câu hỏi:

- Có cần phải tiền xử lý dữ liệu hay không và nếu có thì nhóm sinh viên cần phải xử lý như thế nào?
  - Text: vạch ra các bước thực hiện một cách rõ ràng và dễ hiểu sao cho nếu người đọc không đọc code thì vẫn có thể hiểu được cách nhóm sinh viên tiền xử lý.
  - Code: cài đặt các bước đã vạch ra ở trên. Nhóm sinh viên cũng cố gắng viết code cho rõ ràng và dễ đọc.

- Nhóm sinh viên phân tích dữ liệu như thế nào để ra đượccâu trả lời cho câu hỏi?
  - Text: tương tự như trên.
  - Code: tương tự như trên.

Lưu ý: Ở đây, chúng ta sẽ quan tâm nhiều đến kỹ thuật xử lý (tính toán, trực quan hóa, …) để giải quyết câu hỏi đề ra.

E. Tổng hợp lại quá trình thực hiện đồ án

Sau khi hoàn thành đồ án, mỗi nhóm viết một báo cáo để đánh giá lại công việc như sau:

- Từng thành viên: Bạn đã gặp những khó khăn gì?
- Từng thành viên: Bạn đã học được gì?
- Nhóm của bạn: Bạn sẽ làm gì nếu có nhiều thời gian hơn?

LÀM VIỆC NHÓM

Mỗi nhóm phải sử dụng Git và GitHub để kiểm soát phiên bản và tương tác với các thành viên khác một cách hiệu quả. Mỗi giai đoạn hoặc tác vụ phải có nhánh riêng của nó thay vì tập trung mọi thứ cho nhánh chính (main/master branch). Nhóm sinh viên cần đảm bảo các yêu cầu sau:

- Một kế hoạch cho từng nhiệm vụ được lập cẩn thận (Ai sẽ thực hiện nhiệm vụ? Mất bao lâu để giải quyết nó?)
- Khối lượng công việc được cân bằng giữa các thành viên (Lịch sử cam kết trong Github sẽ cho thấy điều đó)
- Mỗi thành viên phải hiểu tường tận công việc của các thành viên khác trong nhóm.

Kế hoạch và lịch trình phải được theo dõi bằng các công cụ như Notion và Trello. Mỗi nhóm cần thể hiện chiến lược tổng thể và công việc của từng thành viên trong slide báo cáo cuối cùng.

3 CÁC YÊU CẦU THỰC HIỆN ĐỒ ÁN

Sinh viên chú ý các yêu cầu sau:

a) Tổ chức thư mục cho đồ án: Các file notebooks phải được tách biệt rõ ràng cho từng giai đoạn, từ thu thập dữ liệu, tiền xử lý dữ liệu, phân tích đến xây dựng mô hình, đánh giá và phân tích kết quả.
b) Việc trả lời câu hỏi cần được thể hiện thông qua các hình vẽ biểu đồ trực quan và giải thích có tính hợp lý và thuyết phục của sinh viên.
c) Phải có giải thích rõ ràng cho mọi cell code trong file jupyter notebook. Tức là, mỗi cell code nên có một cell markdown kèm theo để giải thích.

4 HÌNH THỨC NỘP BÀI VÀ VẤN ĐÁP ĐỒ ÁN

Mỗi nhóm sẽ thiết lập một GitHub repository trong một cài đặt riêng tư (chế độ private). Thùng chứa sẽ được công khai một ngày trước hội thảo để người hướng dẫn và các cá nhân được chọn có thể xem lại tất cả các công việc. Điều quan trọng cần lưu ý là điểm số sẽ bị ảnh hưởng nếu nhóm bỏ ra ít nỗ lực (ví dụ: ít hơn 10 commit trên GitHub repository) hoặc sử dụng các thủ thuật vào những ngày cuối cùng (ví dụ: dùng một commit duy nhất để hoàn thành đồ án, số lượng commit quá nhiều vào những ngày cuối). Các file cần nộp cho final version trước khi seminar:

- Tất cả file jupyter notebook, các mã nguồn Python nếu có
- Slide trình bày báo cáo (dạng .pdf)
- File .pdf phân công công việc của nhóm
- Dữ liệu đã thu thập (có thể sử dụng Google Drive, One Drive, … và để link trong file .txt)

Vào ngày vấn đáp, mỗi nhóm sẽ có ít hơn/ khoảng trong **15 phút** để trình bày (Trợ giảng sẽ quyết định thứ tự của người trình bày) và **10 phút** cho phần Hỏi & Đáp. Hơn nữa, bài thuyết trình nên tập trung vào công việc một cách rõ ràng, những kỹ thuật và phương pháp thực hiện, thay vì chỉ tập trung vào mã nguồn. **Khi phát hiện hành vi không trung thực, toàn bộ đồ án sẽ bị 0 điểm**. Mọi tài liệu trực tuyến có thể được sử dụng làm tài liệu tham khảo cho các chủ đề của bạn, nhưng cần phải trích dẫn đầy đủ. Tất cả sinh viên được tự do thảo luận về chủ đề của mình với bất kỳ nhóm/ bạn học nào trong lớp, nhưng công việc của nhóm phải được triển khai và diễn giải theo cách hiểu của riêng nhóm sinh viên.

5 TIÊU CHÍ CHẤM ĐIỂM

Tiêu chí	**Tỷ lệ**
Quy trình làm việc với dữ liệu (thu thập, tiền xử lý, phân tích, mô hình hóa dữ liệu) phù hợp.	40%
Xây dựng các câu hỏi và cung cấp được các insight quan trọng từ dữ liệu.	20%
Cài đặt và giải thích quá trình cài đặt một cách tường minh cho các câu hỏi đã đề ra.	20%
Trình bày báo cáo (báo cáo viết + slide + phần thuyết trình)	10%
Vấn đáp (trả lời các câu hỏi của giáo viên thực hành + của các thành viên trong lớp)	10%
Điểm cộng (Chủ đề thú vị, giải pháp hay, những case study thú vị, …)	10%
Tổng cộng	**110%**

6 LIÊN HỆ

Nếu có bất kỳ thắc mắc, sinh viên liên hệ với nhóm giáo viên thực hành

- Trần Đại Chí: ctran743@gmail.com
- Lê Nhựt Nam: lenam.fithcmus@gmail.com