CSC10108 – TRỰC QUAN HÓA DỮ LIỆU

LAB-01: TRỰC QUAN HÓA DỮ LIỆU VỚI PYTHON

(Cập nhật lần cuối: 07/03/2023)

1 THÔNG TIN CHUNG

Loại bài tập	Bài tập thực hành
Thời gian thực hiện	
Hạn nộp bài	
Hình thức thực hiện	Làm việc nhóm
Hình thức nộp bài	Nộp thông qua Moodle FIT
Biên soạn	Nguyễn Thị Thu Hằng (NTT Hằng) Nguyễn Bảo Long (NB Long) Lê Nhựt Nam (LN Nam)
Giáo viên hướng dẫn thực hành	Nguyễn Thị Thu Hằng (NTT Hằng) Nguyễn Bảo Long (NB Long) Lê Nhựt Nam (LN Nam)
Thông tin liên lạc	nam.lnhut@gmail.com (LN Nam) baolongnguyen.mac@gmail.com (NB Long) ntthuhang0131@gmail.com (NTT Hằng)

2 NỘI DUNG

Trong bài tập thực hành này, sinh viên tìm hiểu mối quan hệ giữa các trường dữ liệu thực tế thông qua việc trực quan hóa dữ liệu. Sinh viên thực hiện tìm kiếm dữ liệu đã được công khai trên trang Kaggle datasets: https://www.kaggle.com/datasets về một chủ đề mà nhóm sinh viên quan tâm, hứng thú. Dữ liệu phải được cấu trúc hóa thành một bảng gồm ít nhất 5 trường dữ liệu và 1000 dòng.

Lưu ý: **Sinh viên cần lựa chọn các tập dữ liệu mới gần đây.**

A. Thu thập dữ liệu

Nhóm sinh viên cần trình bày ngữ cảnh và động cơ lựa chọn thực hiện trên tập dữ liệu đã chọn.

- Ngữ cảnh, câu chuyện gì khiến nhóm sinh viên thực hiện việc tìm kiếm 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? Phương pháp thực hiện là gì?

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

Nhóm sinh viên thực hiện tiền xử lý và khám phá tập 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 (dạng số, dạng phân loại) được phân bố như thế nào?
- Có cần phải tiền xử lý dữ liệu hay không và nếu có thì nhóm sinh viên cần phải xử lý như thế nào?

C. Khám phá mối quan hệ trong dữ liệu

- Nhóm sinh viên thảo luận và chọn ra các trường dữ liệu dựa trên các phân tích khám phá dữ liệu để thể hiện trực quan bằng các loại biểu đồ đã học.
- Việc chọn biểu đồ cần giải thích tính phù hợp với tính chất trường dữ liệu. Có thể sử dụng nhiều hơn 1 loại biểu đồ cho trường dữ liệu nhưng cần giải thích lí do.

- Việc thể hiện quan hệ phải tích hợp dần dần nghĩa là từ đơn giản đến phức tạp, từ một trường đơn đến quan hệ giữa nhiều trường, ...
- Ngoài quan hệ độc lập, nhóm sinh viên xem xét liệu trong dữ liệu có quan hệ nhân quả không (cause-effect). Ví dụ: Khi xem xét dữ liệu COVID-19, thì liệu có thể có mối quan hệ giữa tỉ lệ ca nhiễm tăng với số ca chết không, ... Cần chứng minh thông qua các phép trực quan dữ liệu.
- Nhóm sinh viên không cần phải làm hết tất cả các quan hệ nhưng nhiều nhất có thể và phủ được nhiều loại biểu đồ đã học.
- **Khuyến kích sinh viên sử dụng những biểu đồ thú vị cung cấp thông tin hữu ích từ dữ liệu.**

3 NHỮNG GIỚI HẠN

- Bài tập thực hành này được giới hạn trong môi trường lập trình Python đơn giản. **Nhóm không sử dụng phần mềm như Tableau để minh họa.**
- Một số thư viện như NumPy, Pandas, Seaborn, Matplotlib có thể được sử dụng. Các thư viện khác muốn sử dụng cần phải hỏi ý kiến của giáo viên thực hành.
- Dữ liệu không tô màu để hiểu ở lab này.
- **Có thể chạy một số thuật toán học máy đơn giản để hiểu thêm về dữ liệu nhưng không bắt buộc.**

4 HÌNH THỨC NỘP BÀI

Bài tập thực hành được thực hiện theo nhóm. Thời gian và cách thức nộp, xem trên Moodle. Sinh viên tổ chức thư mục nộp bài như sau:

- **Thư mục docs**: chứa báo cáo trình bày trong file .doc/.docx/pdf (Khuyến khích sử dụng định dạng .pdf). Sinh viên trình bày cáo đảm bảo các nội dung như: Thông tin nhóm: tên nhóm, mssv...; Mức độ hoàn thành tổng thể của mỗi yêu cầu; Mức độ hoàn thành của từng thành viên; Chi tiết thuật toán, chạy ví dụ, nhận xét; Khuyến khích trình bày đơn giản, có hình minh họa.
- **Thư mục source_codes**: chứa mã nguồn bao gồm Jupyter notebooks, Python script của nhóm, có kèm hướng dẫn sử dụng (đối với mã nguồn khác Python).

- **Thư mục datasets**: chứa đường dẫn đến tập dữ liệu (nếu quá nặng), đường dẫn đến tập dataset đã tiền xử lý. Khuyến kích sử dụng các kho lưu trữ như OneDrive, hay Google Drive.

5 HÌNH THỨC ĐÁNH GIÁ

Tiêu chí	**Tỷ lệ**
Thu thập và tiền xử lý dữ liệu.	5%
Chọn lựa, giải thích, trực quan các trường và các mối quan hệ giữa chúng.	50%
Rút ra ý nghĩa hợp lý sau mỗi dữ liệu được trực quan.	20%
Xem xét trên nhiều quan hệ, nhiều góc nhìn khác nhau.	10%
Báo cáo trình bày bố cục và định dạng hợp lý, rõ ràng.	15%
Có những phân tích, trực quan hóa bằng những biểu đồ mới lạ và rút ra những thông tin hữu ích. Sử dụng mô hình học máy cơ bản.	5%
Tổng	**105%**
Lưu ý: nếu số quan hệ quá ít thì sẽ xem xét giảm tỉ lệ ở mức 2 và 3.	

6 CÁC QUY ĐỊNH

- Bài không có báo cáo sẽ không chấm.
- Thành viên không tham gia sẽ không có điểm.
- Các nguồn tài liệu tham khảo (nếu có) cần ghi đầy đủ trong báo cáo ở mục Tài liệu tham khảo. **Lưu ý cần phân biệt giữa tham khảo và đạo văn**.
- Đặt tên thứ mục bài làm là MSSV1_MSSV2_MSSV03_MSSV04_ MSSV05_Lab01, với MSSV là mã số sinh viên, nén toàn bộ bài nộp thành 1 tập tin trước khi nộp. Nếu kích thước >20MB thì upload lên server ngoài như Google Drive, ..., nộp link và giữ link public ít nhất trong 2 năm.
- **Bài giống nhau sẽ 0 điểm môn học.**