KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
227 Nguyễn Văn Cừ, Phường 4, Quận 5, TP.HCM
Điện thoại: (08) 38.354.266 – Fax: (08) 38.350.096

Khoa Học Web

PROJECT 2

MỐI QUAN HỆ CỦA DỮ LIỆU

Biên soạn:
Lê Ngọc Thành

1. Nội dung

Tìm hiểu mối quan hệ giữa các trường dữ liệu từ dữ liệu thu thập từ website thực tế.

2. Yêu cầu

Project được thực hiện theo nhóm. Thời gian và cách thức nộp, xem trên Moodle.
Nội dung cần nộp:

- Báo cáo trình bày trong file .doc/.docx/pdf chứa:
  - 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.
- Souce code kèm hướng dẫn chạy nếu thực hiện trong môi trường khác Jupyter Notebook hoặc python gốc.
- Dataset nếu có điều chỉnh so với Project 1 thì cần mô tả thêm.
- Ngôn ngữ lập trình bắt buộc: Python
  - Cho phép sử dụng các thư viện đã được giới thiệu trong lý thuyết.

3. Yêu cầu chi tiết

Nhóm thực hiện tiền xử lý dữ liệu trước khi chuyển sang pha tiếp theo. Dữ liệu gốc và dữ liệu đã điều chỉnh cần lưu lại và nộp kèm trong bài nộp.

Cụ thể trong project này, nhóm được yêu cầu thực hiện các nhiệm vụ sau:

- Thực hiện các thống kê dữ liệu cơ bản như trung bình, độ lệch chuẩn, kỳ vọng, phương sai và một số các công thức thuộc về thống kê dữ liệu.
- Sử dụng nhận xét, code/thuật toán để thể hiện trực quan các mối quan hệ giữa các trường dữ liệu
  - Nhóm thảo luận và chọn ra các trường 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 xem xét liệu trong dữ liệu có quan hệ nhân quả không (cause-effect). Cần chứng minh thông qua các phép trực quan dữ liệu.
  - Nhóm triển khai nhiều mối quan hệ nhất có thể và phủ được nhiều loại biểu đồ đã học. Tối thiểu 10 quan hệ với 6 dạng biểu đồ khác nhau.

Nhóm giữ lại các dữ liệu để có thể thực hiện tiếp cho các bài sau.

4. Những giới hạn

- Bài project này được giới hạn trong môi trường lập trình. Nhóm **không** sử dụng phần mềm như Tableau để minh họa hoặc đưa phần đó vào phần bổ sung cộng điểm.

- Một số thư viện như numpy, pandas, seaborn, matplotlib nên sử dụng.

5. Đánh giá

- Các tiêu chí đánh giá:
    1. Tiền xử lý dữ liệu (5%)
    2. Thống kê dữ liệu (10%)
    3. Chọn lựa, giải thích, trực quan các trường và các mối quan hệ giữa chúng (35%)
    4. Rút ra ý nghĩa hợp lý sau mỗi dữ liệu được trực quan (20%)
    5. Xem xét trên nhiều quan hệ, nhiều góc nhìn khác nhau (10%)
    6. Báo cáo trình bày bố cục và định dạng hợp lý, rõ ràng (20%)

Lưu ý: nếu số quan hệ quá ít thì sẽ xem xét giảm tỉ lệ ở mức 3 và 4.

6. Qui định

- Bài không có báo cáo và code 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 thư mục bài làm là MSSV1_MSSV2_MSSV03_..., 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 1 năm.
- Bài giống nhau sẽ 0 điểm môn học.

7. Liên hệ

Mọi thắc mắc trong quá trình thực hiện vui lòng gửi mail về lnthanh@fit.hcmus.edu.vn