

Lab 01: Thu nhập dữ liệu và trực quan hóa dữ liệu

Võ Nhật Tân

1 Quy định chung

- Bài làm được thực hiện theo nhóm, tối thiểu 2 người và tối đa 3 người.
- Thành viên không tham gia sẽ không có điểm bài tập này.
- Bài làm phải tuân thủ theo yêu cầu đề án.
- Các nguồn tài liệu tham khảo (nếu có) cần ghi đầy đủ trong báo cáo ở mục Tài liệu tham khảo.
- Các công cụ hỗ trợ trong việc code, viết báo cáo như ChatGPT, Github Copilot, ... chỉ dùng như một công cụ tham khảo; nội dung phải được kiểm tra và chỉnh sửa phù hợp với bài toán đặt ra. Nếu phát hiện sử dụng công cụ AI để sinh nội dung quá nhiều, hoặc nội dung không phù hợp/sai lệch thì sẽ trừ tối đa 50% số điểm (tùy mức độ).
- Bài giống nhau sẽ **0 điểm môn học**;
- Báo cáo sao chép nội dung từ các nguồn khác không trích dẫn sẽ **0 điểm phần thực hành**;

2 Giới thiệu đề án

2.1 Giới thiệu đề án

Hiện nay, các nền tảng nghe nhạc trực tuyến như Spotify, Zing MP3, Nhaccuatui, SoundCloud, ... là một trong những kênh để nghe nhạc và khám phá các xu hướng âm nhạc mới. Trên các nền tảng này chứa dữ liệu khổng lồ về các thông tin bài hát, nghệ sĩ, thể loại, Tuy nhiên, dữ liệu này thường phân tán và chưa được khai thác hiệu quả.

Đề án của chúng ta sẽ tập trung vào việc thu thập, làm sạch và phân tích dữ liệu từ các nền tảng này. Chúng ta sẽ khám phá mối quan hệ giữa các yếu tố âm nhạc, xác định xu hướng thị trường và phát hiện những đặc điểm nổi bật trong hành vi nghe nhạc của người dùng bằng cách áp dụng các kỹ thuật xử lý dữ liệu, phân tích thống kê và trực quan hóa. Nhờ đó, xây dựng một bức tranh về thị trường âm nhạc trực tuyến tại Việt Nam.

2.2 Nhiệm vụ đề án

Trong đề án này, các bạn sẽ thực hiện các nhiệm vụ sau:

2.2.1 Thu nhập và tiền xử lý dữ liệu

- Nhóm có thể chọn một hoặc nhiều nền tảng âm nhạc trực tuyến Zing MP3, Nhaccuatui, SoundCloud, để thu nhập dữ liệu (có thể sử dụng API, web crawling, để thu nhập dữ liệu)
- Dữ liệu sau khi thu nhập sẽ được qua các bước tiền xử lý (nếu cần) như làm sạch dữ liệu, chuẩn hóa, và được lưu dưới dạng **.csv** (có thể lưu thành nhiều file nếu cần)
- File dữ liệu sạch cần chứa ≈ 3000 records có ý nghĩa phân tích.

2.2.2 Trực quan hóa dữ liệu

1. Phân tích cơ bản về dữ liệu: giới thiệu dữ liệu, cỡ mẫu, cấu trúc, phân tích thống kê, xử lý dữ liệu,
2. Xác định mục tiêu phân tích và lựa chọn các trường dữ liệu
 - Xác định rõ mục tiêu phân tích và lựa chọn các trường dữ liệu phù hợp để trực quan hóa
 - Cần xem xét các mối quan hệ giữa các trường để đưa ra lựa chọn phù hợp nhất.
 - Mỗi nhóm cần xác định ít nhất [**Số lượng thành viên**] * **2** mục tiêu phân tích và trực quan.
3. Lựa chọn biểu đồ thích hợp và giải thích lý do
 - Biểu đồ cần phải phù hợp với tính chất của trường dữ liệu.
 - Nhóm sử dụng đa dạng các biểu đồ để làm rõ mục tiêu đã đề ra.
 - Toàn bài phân tích cần bao phủ hết các biểu đồ đã được học.
4. Nhận xét và đưa ra kết luận dựa trên các biểu đồ

3 Yêu cầu đồ án

- Nhóm thực hiện các nhiệm vụ trên trong môi trường lập trình Python, không sử dụng các phần mềm khác như Tableau, PowerBI, ... để minh họa; Các thư viện được sử dụng như *numpy*, *pandas*, *seaborn*, *matplotlib* được phép sử dụng; các thư viện ngoài thì cần ghi rõ trong báo cáo lý do sử dụng.
- Nhóm tổ chức tất cả các code Python trên một file Jupyter Notebook (**.ipynb**) và sử dụng các cell Markdown để mô tả rõ quá trình thực hiện.
- Nhóm không được sử dụng dữ liệu có sẵn trên Kaggle, mà phải tự thu nhập bằng các cách như web crawling, gọi api, ... Code để thu nhập cần có trong file Jupyter Notebook.
- Có thể sử dụng các thuật toán máy học trong việc phân tích và thể hiện trên biểu đồ.
- Trong quá trình phân tích, có thể dùng **dữ liệu của nhóm khác** nhưng cần ghi rõ lý do trong cả báo cáo và phần code (nếu mục đích phù hợp sẽ được cộng điểm, ngược lại sẽ trừ điểm tùy mức độ).
- Viết báo cáo ngắn gọn và đầy đủ toàn bộ quá trình thực hiện; kết quả và nhận xét. Báo cáo tối đa là 24 trang (không bao gồm trang bìa, mục lục và tài liệu tham khảo). Nội dung file code và báo cáo phải khớp với nhau. Nội dung báo cáo bao gồm:
 - 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 các bước thực hiện (kèm hình ảnh), thuật toán, chạy ví dụ, nhận xét. Trình bày đơn giản, có hình minh họa.

4 Quy định nộp bài

- Nhóm sẽ cử đại diện 1 người nộp bài.
- Khi nộp bài, cần chạy lại các cell (không cần chạy lại code thu nhập dữ liệu) và không xóa output của các cell (đảm bảo biểu đồ đúng như báo cáo).
- Bài nộp là một file nén đặt tên là [**MSSV_1, MSSV_2, MSSV_3**].zip sẽ gồm:

1. Thư mục chứa dữ liệu/Link tới dữ liệu
 2. File code: `[MSSV_1, MSSV_2, MSSV_3].ipynb`
 3. File báo cáo: `[MSSV_1, MSSV_2, MSSV_3].pdf`
- Trong trường hợp dữ liệu quá nặng, các bạn chỉ upload dữ liệu lên server ngoài như Google Drive, ..., nộp link và giữ link public ít nhất trong 2 năm; còn file code và báo cáo vẫn nộp trên Moodle.

5 Liên hệ

Mọi thắc mắc trong quá trình thực hiện vui lòng gửi mail về vntan.work@gmail.com