Trường Đại học Khoa học Tự nhiên

227 Nguyễn Văn Cừ, P4, Q5, TP HCM

ĐA1:Thu nhập dữ liệu & EDA

24/10/2010

Tổng quan

Trang **Soundcloud** là một trang website cho phép người dùng upload và chia sẻ các bài hát. Chúng ta sẽ khám phá cộng đồng gồm các nghệ sĩ, ban nhạc, podcast và người sáng tác âm nhạc trên trang Soundcloud thông qua đồ án này.

Nhiệm vụ: Thu nhập các thông tin từ trang **https://soundcloud.com/** và đưa ra các thông tin/đánh giá từ tập dữ liệu đã thu nhập được.

Mục tiêu

1. **Làm quen và biết cách thu nhập dữ liệu (50%):** sử dụng cả hai cách để thu nhập dữ liệu: *sử dụng api* và *parse HTML*
2. **Phân tích dữ liệu (EDA) (50%):** Thực hành cơ bản các bước phân tích dữ liệu từ bộ dữ liệu đã thu nhập được.

Chi tiết

Thu nhập dữ liệu

Các dữ liệu cần phải được thu nhập bao gồm:

- Track
- Playlist
- User

Bạn sẽ sử dụng hai cách API và parse HTML trực tiếp từ trang SoundCloud.

Phân tích dữ liệu

Từ dữ liệu đến thông tin chi tiết

Trước khi bạn khám phá dữ liệu, hãy viết ra danh sách ngắn về những gì bạn mong đợi thấy trong dữ liệu: sự phân bố của các biến chính, mối quan hệ/quan trọng giữa các cặp biến, v.v. Danh sách như vậy về cơ bản là một dự đoán dựa trên hiểu biết hiện tại của bạn về dữ liệu

Bây giờ hãy phân tích dữ liệu. Lập bảng, tóm tắt, bất cứ điều gì cần thiết để xem nó có phù hợp với mong đợi của bạn không.

- Danh sách kiểm tra phân tích dữ liệu: Danh sách kiểm tra này có thể được sử dụng như một hướng dẫn trong quá trình phân tích dữ liệu hoặc như một cách để đánh giá chất lượng của một phân tích dữ liệu được báo cáo.
- Trả lời những câu hỏi về bộ dữ liệu:
    1. Bạn đã xác định số liệu trước khi bắt đầu?
    2. Bạn đã hiểu ngữ cảnh cho câu hỏi và ứng dụng?
    3. Bạn đã xem xét liệu câu hỏi có thể được trả lời với dữ liệu có sẵn không?
- Xóa dữ liệu
    1. Bạn có xác định được dữ liệu bị thiếu không?
    2. Các kiểu dữ liệu khác nhau có xuất hiện trong mỗi bảng không?

3. Kiểm tra các ngoại lệ?

- Phân tích khám phá
    1. Bạn có thực hiện các trực quan của đơn biến (histogram, distplot, boxplot) không?
    2. Bạn đã xem xét mối tương quan giữa các biến (scatterplot, jointplot, kde plot, correlation matrix)?
- Trình bày:
    1. Bạn đã dẫn dắt một cách ngắn gọn, dễ hiểu cho mọi người về vấn đề của bạn?
    2. Bạn đã giải thích dữ liệu, mô tả câu hỏi cần quan tâm?

Yêu cầu

1. Code

   Làm trực tiếp trên các file notebook .ipynb. Các bạn có thể sử dụng jupyter lab trong quá trình thực hiện nếu thấy thuận tiện.

2. Dữ liệu thu nhập

   Để tại hai thư mục:

   - Api_data
   - Crawl_data: sử dụng parse HTML

   Mỗi thư mục cần phải có các file track.csv, playlist.csv, user.csv

Với việc sử dụng parse HTML, bạn cần phải thu nhập được trường (cột) dữ liệu **nhiều nhất** có thể.

Playlist có thể có nhiều track, chỉ cần để một cột dữ liệu: trackIds là 1 string danh sách các id. Ví dụ: playlists[1]["tracks"] = "345,376,389". Khi sử dụng chỉ cần tách các số trong string ra là có một danh sách các track của 1 playlist.

Chú ý: Các file của thư mục API thường sẽ nhiều cột dữ liệu hơn so với việc parse HTML.

3. Phân tích dữ liệu

Chú ý: Chỉ làm trên bộ dữ liệu được thu nhập bằng cách sử dụng parse HTML.

Cần phải đưa ra các câu hỏi từ bộ dữ liệu. Phải hiểu bộ dữ liệu để đặt câu hỏi cho chính xác và có khả năng trả lời được câu hỏi. (Ví dụ: loại nhạc nào được yêu thích, user nào hot nhất,...). Câu hỏi đặt càng hay sẽ được đánh giá cao + điểm cộng

Xử lý dữ liệu: Clean data, Exploratory analysis (phân tích, khám phá)

Mỗi một một thể hiện (khám phá) cần phải có biểu đồ (line, circle, bar, plot,...).

4. Chú ý

Bài làm giống nhau 0 điểm cả môn.

Ghi rõ nguồn tham khảo đầy đủ.

Không cần viết báo cáo. Chỉ cần thể hiện trong **các file notebook**.

5. Nộp bài

Nén thành một file MSSV.zip nộp qua moodle.

Thông tin liên hệ

TA:Hoàng Xuân Trường

Nếu có thắc mắc gì liên hệ qua:hxtruong6@gmail.com