**Báo cáo đồ án 2**

**Bảng phân công công việc:**

|  |  |  |  |
| --- | --- | --- | --- |
| **MSSV** | **Họ và tên** | **Công việc** | **Mức độ tham gia** |
| 19127330 | Lê Tâm Anh | * Đặt câu hỏi về mối tương quan giữa các trường dữ liệu của Track và User. * Vẽ biểu đồ trực quan cho Track và User. | 33% |
| 19127326 | Vương Thành An | * Đặt câu hỏi về mối tương quan giữa các trường dữ liệu của Track và Playlist. * Vẽ biểu đồ trực quan cho Track và Playlist. | 34% |
| 19127281 | Trần Minh Thiện | * Đặt câu hỏi về mối tương quan giữa các trường dữ liệu Playlist và User. * Vẽ biểu đồ trực quan cho Playlist và User. | 33% |

***Trả lời những câu hỏi về bộ dữ liệu:***

1. Bạn đã xác định số liệu trước khi bắt đầu?

* Trong quá trình thu thập dữ liệu thì các thuộc tính mình đã có thể quan sát trước khi quyết định chọn để lấy về và lấy tất cả những gì mình có thể lấy được
* Dữ liệu api mỗi file khoảng 2000 mẫu
* Dữ liệu html mỗi file khoảng 1000 mãu

1. Bạn đã hiểu ngữ cảnh cho câu hỏi và ứng dụng :

* Nhìn vào bộ dữ liệu em đã thấy được mình cần tìm hiểu về các mối quan hệ tăng giảm, tỉ lệ thuận của các trường dữ liệu và từ đó rút ra được ý nghĩa của các trường dữ liệu và áp dụng vào thực tế

3.Bạn đã xem xét liệu câu hỏi có thể được trả lời với bộ dữ liệu có sẵn hay không ?

Sau khi quan sát dữ liệu được lấy về em tự tin có thể giải quyết câu hỏi dựa vào dữ liệu đang có

***Xóa dữ liệu:***

1/ Dữ liệu có bị thiếu không?

- Dữ liệu khá đầy đủ

- Các thuộc tính của mỗi loại(user, playlist, track) rất đa dạng đặc biệt là dữ liệu được lấy về kiểu API

- Dữ liệu lấy về kiểu HTML đảm bảo số mẫu đủ nhiều khoảng 1000 nhưng có số cột dữ liệu ít hơn nhiều so với dữ liệu API tuy nhiên đó đều là các cột dữ liệu quan trọng gần như có thể lấy để phân tích và trực quan.

- Còn dữ liệu lấy về kiểu API tuy nhiều trường dữ liệu nhưng phải chọn lọc tính toán thì mới đưa ra được các trường dữ liệu hợp lí để phân tích

2/ Mỗi bảng có các kiểu dữ liệu khác nhau? Có kiểu dữ liệu nào chưa phù hợp?

- Giữa 2 bảng dữ liệu lấy về theo 2 kiểu HTML và API thì đa số các cột dữ liệu quan trọng thì sẽ giống nhau nhưng bảng dữ liệu api thì có nhiều cột dữ liệu hơn dữ liệu HTML tuy nhiên chỉ có thể tận dụng thêm vài cột dư liệu để phân tích còn lại thì đa phần có nhiều cột dữ liệu dư thừa không cần thiết

- Số mẫu dữ liệu api cũng nhiều hơn dữ liệu HTML do HTML tốn nhiều thời gian để lấy hơn

- Cột duration của playlist và track có kiểu dữ liệu lấy về là string khi lấy về phải chuyển về dạng timedelta

- Cột date\_release của playlist và track track có kiểu dữ liệu lấy về là string khi lấy về phải chuyển về dạng datetime

- Cột created\_date của user có kiểu dữ liệu là string phải chuyển về dạng datetime

3/ Kiểm tra các ngoại lệ :

* + Kiểm tra các giá trị None nên xem xét để xóa đi hoặc sửa lại .
  + Sau đó đối với những ngoại lê khác khi trực quan hóa thì sẽ quyết định xóa để có thể trực quan hóa tốt hơn và thấy được các sự tương quan giữa các mối quan hệ dữ liệu rõ hơn