TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHÊ THÔNG TIN

BÁO CÁO

ĐỒ ÁN 1 – THU THẬP DỮ LIỆU



Giáo viên hướng dẫn: Lê Ngọc Thành

Sinh viên thực hiện

Lê Tấn Đạt - 19127353

Lê Hoàng Thịnh Phước – 19127518

Lê Tiến Trí - 19127593

*Ngày 12, tháng 11 năm 2021*

1. Phân công công việc

|  |  |  |  |
| --- | --- | --- | --- |
| Tên Thành viên | Công việc | Hoàn thành | Đánh giá(%) |
| Lê Tấn Đạt |  |  |  |
| Lê Tiến Trí |  |  |  |
| Lê Hoàng Thịnh Phước |  |  |  |

1. Mô tả về dữ liệu và lấy dữ liệu trên SoundClound
2. API\_Data:
3. *Quy trình lấy dữ liệu*

* Ta khai thác dữ liệu dựa vào API do SoundClound cung cấp , ở đây nhóm sử dụng 3 API để có thể thu thập dữ liệu của *users, tracks, playlists*
* Users: <https://api-v2.soundcloud.com/users/><entity\_id>?client\_id=<client\_id>
* Tracks: <https://api-v2.soundcloud.com/users/><entity\_id>/tracks?client\_id=< client\_id>
* Playlists: <https://api-v2.soundcloud.com/users/><entity\_id>/tracks?client\_id=< client\_id>
* Vấn đề: Trang SoundClound không còn cho đăng kí ClientID, ta chỉ có thể lấy ClientID bằng việc inspect trình duyệt SoundClound và tìm kiếm ClientID ở trong đó

+ Sử dụng selenium để mở google chrome và mở SoundClound để có thể inspect lấy html về. Ta thấy được ClientID được lưu trữ ở thẻ div với id = “g\_id\_intermediate\_iframe” để lấy dãy kí tự string là ClientID.

Graphical user interface, text, application, email

Description automatically generated

* Tiếp theo ta sẽ sử dụng requests các API để có thể lấy dữ liệu về và đưa vào 3 file user.csv, track.csv và playlist.csv

+ Ta sẽ tạo 1 header kèm với requests để có thể tránh lỗi 403 do requests quá nhiều lần. Đồng thời cũng thiết lập thời gian sleep time để chương trình sleep

+ Khi ta lấy được dữ liệu về ta tiến hành xóa 1 số thuộc tính không cần thiết như:

* user trong track và playlist thay vào đó ta chỉ cần lưu userID
* Xử lí thuộc tính tracks ta sẽ thay thế bằng các ID của cách track đó

1. *Mô tả dữ liệu lấy được*

* Mô tả dữ liệu: Nằm trong file jupyter notebook ( phân tích dữ liệu lấy được )
* Mối quan hệ của dữ liệu:

Diagram

Description automatically generated

1. Crawl\_Data:
2. Kết quả lấy được
3. API\_Data:
4. *Số lượng records trong csv*

|  |  |  |
| --- | --- | --- |
| Tên file | Số lương records | Số lượng thuộc tính |
| User.csv | 1500 | 38 |
| Track.csv | 4535 | 63 |
| Playlist.csv | 2093 | 32 |

1. *Đánh giá dữ liệu*

* Vấn đề bias:
* Dựa vào việc ta lấy dữ liệu của 1500 user đầu tiên đồng thời track và playlist của họ, ta thấy được việc lấy sample như vậy là không bias vì cách lấy dữ liệu đó thuộc dạng Systematic Random và đảm bảo cách user trong đó có độ ưu tiên là như nhau
* Mẫu có những thông tin về thống kê:

Các trường dữ liệu về playback\_count, like\_count,repost\_count,comment\_count

1. Crawl Data:
2. Tài liệu tham khảo

* <https://developers.soundcloud.com/docs/api/explorer/open-api>
* <https://stackoverflow.com/questions/54753238/soundcloud-application-registration-form-is-closed>