Đồ án 1 - Thu thập dữ liệu

Tổng quan:

Trang Soundcloud là một trang web cho phép người dùng upload và chia sẻ các bài hát. Chúng ta sẽ thu thập thông tin về 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/ bằng parse HTML và API

Mục tiêu:

Làm quen và biết cách thu thập dữ liệu của một trang web bằng cả parse HTML và API.
Lưu ý trong đồ án này các bạn cần tự thiết kế / tổ chức các file notebook chứa mã nguồn, sao cho gọn gàng và dễ theo dõi nhất.

Chi tiết

Các dữ liệu cần phải được thu thập bao gồm:

- User
- Playlist
- Track

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

- Các file user, tracklist,.. nên có mối liên hệ với nhau
- Số lượng mỗi file cần có ít nhất tầm > 1000 records
- Các bạn có thể dùng bất kỳ thư viện nào của python

Gợi ý 1: API có dạng:
https://api-v2.soundcloud.com/<số nhiều của entity>/<entity_id>?client_id=<client_id>

trong đó entity = {user, track, playlist}
Cụ thể client_id là gì các bạn có thể tham khảo document của Soundcloud
Gợi ý 2: client_id có thể tìm bằng chức năng Inspect của trình duyệt

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 thập

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.

Sau khi thu thập dữ liệu, các bạn mô tả sơ lược trong notebook về thông tin dữ liệu thu thập được (số mẫu thu thập, ý nghĩa các đặc trưng,...)

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

3. Bảng phân công công việc

Đầu notebook của mỗi nhóm cần có bảng danh sách tên và phân công công việc của các thành viên.

Các bạn lưu ý chia việc hợp lý, mỗi thành viên cần đảm nhận lượng công việc tương đương nhau.

4. Lưu 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, đặt tên theo cú pháp sau và nộp qua moodle:

<MSSV1>_<MSSV2>_<MSSV3>_<MSSV4>_<MSSV5>.zip

Thông tin liên hệ

TA: Lê Minh Nhật
Nếu có thắc mắc các bạn vui lòng liên hệ qua: minhnhatvt2@gmail.com