Audio & Music Processing Lab - 2023

Collaborative Filtering - Thomas LE ROUX

1 - Generate the csv

Our goal is to generate a csv file containing the information formatted like this :

```
user_id,artists_mbid,artist_name,count
17240,0004d643-96e3-486a-bafe-eb792871ce9f,Siddhartha,2
17240,0008af7d-2aa1-4b4d-80af-b3b64ee3cac6,Miklós Rózsa,81
17240,000fc734-b7e1-4a01-92d1-f544261b43f5,Cocteau Twins,3
```

To do this, we divided the work of processing the data in a lot of small steps :

1. csv file linking the user_id with the recording_msid of every song he listened in the year : dict_user-msid.csv - 1.8Go
2. csv file linking the recording_mbid with the artists_mbid : recording_mbid2artists_mbid.csv - 1.73Go
3. csv file linking the every recording_msid with the artists_mbid that made the song : recording_msid2artists_mbid.csv - 3.34Go
4. csv file linking the user to every artist he listened to : user_id2artists_mbid.csv - 1.26G
5. csv file linking the user to every artist he listened to - counting for every artist : user_id2artists_count.csv - 144Mo
6. Final csv file with the right template : user_id2artists_count_better_formatted.csv - 194Mo

Results :
At the end, we obtain a nearly 200Mo file with the right template.

Some details :

- For the long and massive data processing, there is some printing being done that gives the user some information about where the process is.
- For certain links, there is an error rate detection made with try/except. When this system was needed, we had a 3% error rate.
- We chose to link the user with every artists of every track he listened
- The choice has been made to only keep the exact and high quality matches

2 - Collaborative Filtering using Implicit

For this task, we use the Implicit library.
The choice has been made to not use the hdf5 format as it's not needed when the format required has been understood.
To be able to have the artist name and to be able to query an artist name or an MBID for the 'related artists' part, we need to build dictionaries linking with our values.

After several try and iterations, we obtain the following results :

For this user : 6278, the collaborative filtering recommends :

	artist	score	already_liked
0	Alias Conrad Coldwood	0.528886	True
1	Pentagon	0.388292	True
2	永田権太	0.281805	True
3	Toby Fox	0.275514	False
4	Richard Jacques	0.274436	False
5	祖堅正慶	0.273329	False
6	Jackal Queenston	0.262407	False
7	steventhedreamer	0.261463	False
8	Darius	0.260600	False
9	RushJet1	0.258647	False

We can't conclude a lot of information on this, however there are some things to comment about the 'similar artists recommandation'.

We work for example with Mall Grab. He is an Australian DJ, producer and also label owner.

He was first known by being one of the leading figures of the lo-fi house, a genre that emerged in 2016. This was one of his first songs :
https://www.youtube.com/watch?v=wyEEvnZw3Z8&ab_channel=OOUKFunkyOO

He really touched several genres of electronic music, going from more soulful joyful disco-house (https://www.youtube.com/watch?v=GHmHY2Wuh_Q&ab_channel=MallGrab) to more grime/drum and bass (https://www.youtube.com/watch?v=9YST-05c03s&ab_channel=MallGrab) to more techno, flirting with rave and hardcore vibes (https://www.youtube.com/watch?v=3_al-rdXySQ&ab_channel=MallGrab, https://www.youtube.com/watch?v=GVm7YHfnZi8&ab_channel=MallGrab).

We compare our results with 3 lists of 'similar artists' (based on Spotify, LastFM and Deezer).

We can see that, even if there are differences, our system seems to be pretty accurate. It also represents the various sides of Mall Grab :

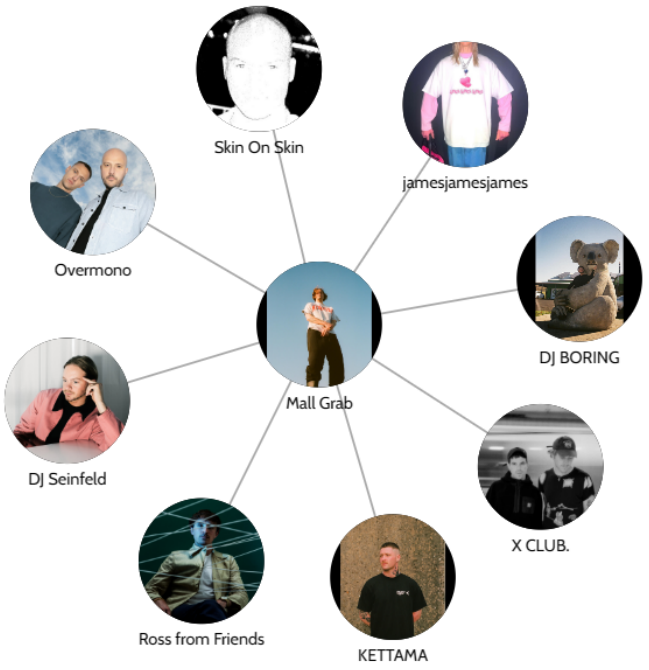
- Lofi House :
  - DJ Seinfeld
  - Baltra
  - DJ Boring (somewhat lo-fi house thanks to his groundbreaking Winona https://www.youtube.com/watch?v=ar9qA3WJInk&ab_channel=Slav)
  - DJ Swagger
- House (going lo-fi but also more hard):
  - Harrison BDP
  - Riovh
  - Overmono
- 90's hard rave / transe :
  - Dax J
  - Asquith

I would say that every ‘similar artist’ is somewhat coherent to what I would think. Some of them differ, but we find a lot of similarities.

Our results

	artist	score
0	Mall Grab	1.000000
1	Overmono	0.830906
2	DJ Seinfeld	0.821291
3	DJ Boring	0.814272
4	Harrison BDP	0.809491
5	Asquith	0.805530
6	Dax J	0.802230
7	Riohv	0.798247
8	Baltra	0.795541
9	DJ Swagger	0.793326

Musicroamer



Last FM :

- DJ Boring
- Skin On Skin
- Kettama
- Fjaak
- COMPUTER DATA
- Subjoi
- Demuja
- Denis Sulta
- Ross From Friends
- Baltra

Deezer :