Audio and Music Processing Lab

Report - Collaborative Filtering with ListenBrainz

UPF Barcelona, SMC 2024-2025

Madhav Jaideep

github repo: https://github.com/MadhavJ06/amplab-collaborative-filtering.git

Data pre-processing

- Processed data from ListenBrainz, 12 files for each month in 2024. Extracted user_id and recording_msid (MessyBrainz ID) from the files and stored them in user-msid.csv file. The next step was to process the mapping file and for this I used only the "exact_match" and "high_quality" matches in order to have a more efficient, smaller and also memory-efficient mapping. This smaller mapping covers around 55% of the total data given, and is good enough for our task as it's more reliable data for later tasks. Processed canonical redirects, where the original MBIDs were maintained in cases where no redirect existed.
- In order to process the dataset with the artist data, when a recording has multiple artists, I chose to take only the first artist in the list as it is most possibly the original artist of the recording. From this we create a mapping with the artist IDs. Similarly the artist IDs were mapped to their respective names, this mapping consists of the recording, artist MBIDs and artist name with over 27 million processed recordings.
- Final step of the data pre-processing was to create the final data file that contained the user ID, artist ID, artist names and play counts. Around 7.9 million unique user-artist pairs were present in this data. Additionally in order to have the data structure prepared for the ListenBrainz model, the data structure was modified to have proper headings and it's content.
- In order to understand and make sure the code is accurate, I kept verifying the data structure and content after each preprocessing steps to make sure we have the data computed correctly. Checking items present was also helpful in order to understand the same.

Collaborative Filtering

- Following the implicit tutorial, a collaborative filtering model is built using our extracted data. The artist MBIDs are mapped to their respective artist names.
- **Artist similarity (Recommending similar artsits):**
    - Using the model we give a query artist MBID to find similar artists to that. The model seems to perform very well for this task. A couple of artists that I like were given as query and it returned pretty accurate results, at least

according to my personal preferences. Notably, I gave the ID of an artist "Jeremy Zucker" who makes music in the type of indie electro-acoustic pop and the model predicted other artists who very closely could be associated with this type of music and genre, including some other artists that I listen to as well. The model mostly gives accurate results and one way we can maybe see it struggle is if you have a very subjective view of a particular artist and expect a specific type of music, like for example some people might categorize Pink Floyd as being more psychedelic-space-rock and not the generic alt rock genre, which is what the model predicted for Pink Floyd, which although is understandable.

    - When compared with other platforms like spotify or Last.fm, the results were very similar to those and predicted around 90% of the same artists as results.

- When investigating the data model, from analysis we see that from the final extracted we have around 7.9 million entries, with over 15k unique users. Alternatively, when looking at the documentation for the Lastfm 360k dataset, we see that there are around 360k unique users, true to its name. From this, we can assume that the Lastfm dataset, offers a broad and diverse representation of user listening behaviours, making it better suited for large-scale modelling and generalization across different profiles. Although our extracted data from ListenBrainz has only 15k users, it contains 7.9 million entries, indicating dense interaction data, hence providing to the good performance of the model.
- Finally, when analysing the most common artists in the dataset, it is noted that all the top or most common artists listened to are concentrated towards more contemporary western pop and hip-hop genres, with artists predominantly from the United States and UK. This could suggest that the system may exhibit a bias towards promoting more mainstream western music, overlooking a more diverse range of artists, possibly even artists that could be more to the user's liking from other regions of the world. One way to mitigate this would be to have a dataset with a wide variety of genres and musical styles from all around the world with a more balanced representation.

Use of LLM/AI tools

- I mostly used ChatGPT in order to understand how to perform certain tasks required for the data preprocessing steps to get the right structure and format of data. In certain places used it to help identify and fix errors, at times the errors were identified incorrectly and did not give me the appropriate fix required, so had to spend time going through the code and giving a more detailed and informed prompt to understand the issue.