F6 report

Music Recommendation System

https://github.com/robinmurumets/Music-Recommendation-System

Background:

- The project focuses on creating a more personalized music recommendation system that suggests songs that the user would like, but also to help the user discover new music/artists. As well as furthering our own knowledge of the field and gaining new practical skills and understanding of the subject.

Business goals:

- Develop a recommendation system that predicts user song/artist preferences as accurately as possible
- Gain a better understanding of how the different recommendation systems work and how can they be combined to produce a more efficient product
- The recommendation system should recommend more niche artists and songs as well so the user can discover new genres, artists, songs

Inventory of Resources:

- It will be the two of us doing the project and for it we have the two laptops. We can also ask for help from our practice session teacher.
- Right now we have identified two datasets that we want to use for the project but for additional datasets that we might later find useful we still can use the million song dataset website that has links to additional data that might be useful for us. http://millionsongdataset.com/

Requirements, assumptions, and constraints

- One of the constraints is we have trouble estimating how long some tasks can take
- We need to state clear objectives for the project by 4th of December
- One of the constraints of this project is that we sometimes struggle with understanding what is the relevant data to use to achieve our objectives, but we do have access to a lot of useful data
- Another one of the constraints
- We will have applied relevant data mining methods on relevant data by 9th of December
- We would like to have our poster ready by 10th of December
- The deadline to present our project is 13th of December

Risk and contingencies

- One of the team members gets sick. We can delay some of our deadlines a little bit so the sick team member can catch up.
- Some of the tasks we have assigned to one of us will take longer than expected. We will reassess the divide of work and we will both work on the difficult tasks.
- Our laptops might not be able to compute the data. One of us has a more powerful computer at home that can work with bigger data.

Terminology

- Collaborative Filtering: Recommendation method based on user-user or item-item similarity.
- Content-Based Filtering: Recommendations based on item attributes such as song metadata.
- Triplets: a data representation where each record consists of three elements or attributes.

Costs and benefits do not need to be considered in the case of our prodject as we are doing it for educational purposes.

Outline data requirements:

The primary data requirement for this project involves the interaction between the users and songs. Specifically, the following are required:

- **User Data:** Information about which user listened to which song and the frequency of these interactions (listening count).
- **Song Data:** Metadata about songs, including:
    - Song title
    - Release year
    - Artist name
    - Song ID (to link with user data)

Possible Integration Opportunities: Combining collaborative filtering, content-based filtering (using song metadata such as genre and year), and popularity-based filtering (aggregate listening count).

Combining this data is crucial for a more robust hybrid recommendation system which can better analyze trends and user prefrences

Verify data availability

The data for this project is sourced from the publicly accessible OneMillionSongDataset, contributed by multiple researchers and organizations dedicated to advancing recommendation system research. This dataset includes:

- User-song interaction data: Contains user IDs, song IDs, and listening counts.
- Song metadata: Includes song IDs, titles, artists, release years, and additional descriptors.

Access to this dataset ensures that all required data is available for this project, although further exploration and potential cleaning might be necessary.

Define selection criteria

Currently, all features in both datasets appear relevant for the project:

- **User-song interaction data** provides the foundation for collaborative filtering.
- **Song metadata** is essential for content-based filtering and ensuring that the system can make diverse recommendations.
- **Listening counts** contribute to both collaborative and popularity-based filtering techniques.

For now, the full dataset will be used while in the exploratory and preparatory stages.

Describing data

- The data is contained in 2 csv files. One of the files contains the user_id,song_id and the listening count while the other contains the specific data of the song such as, song_id (to connect the 2 datasets), title, artist_name, year. This data seems suitable for the specific needs of this project, but if we find a more suitable dataset or something that would add to the project and enhance the accuracy of our trained model, then we would add it

Exploring Data

Initial inspection of the data reveals some issues that need solving:

- **Missing Data:** Missing values are present in certain fields. This could affect the quality and accuracy of content-based recommendations.
- **Duplicates:** Duplicate entries may exist, particularly in user interaction logs, and could distort metrics like popularity.
- **Value Errors:** Inconsistent or incorrect values (songs with wrong or not logical years or user interactions with impossibly high counts) need correction.

Detect outliers and anomalies in numerical data. Identify patterns and distributions of user interaction frequencies and song metadata.

Verifying data quality

While the dataset is generally robust, missing values need imputation or exclusion, depending on their criticality. It's important to keep the data consistent to avoid mismatches and problems in the suggestions.

Conclusion: The dataset is well-suited for the project's goals but requires some cleaning and integration to maximize its usability. As the project progresses, the dataset will be refined on an ongoing basis.

Project plan

- Make a simple recommendation system that uses the dataset that consists of triplets. 10 hours each
- Set clear objectives for how to improve that recommendation system 5 hours each
- Applying the second dataset to the project after we have stated clear objectives 10 hours each
- Making final improvements to the project and documenting our work in GitHub 5-10 hours each.
- Making a poster 1-2 hours each