Dylan Mohsen

DS 210

Final Project Proposal

1.

The dataset I chose is the Top Spotify Songs of 2023 from Kaggle. This dataset is interesting because it includes various features for each song such as artist, genre, streams, and more. These features can be utilized to construct a graph where songs are vertices connected by edges based on shared characteristics or direct collaborations between artists.

2. Problem Statement and Insights

The primary question this project aims to answer is: "How are top songs of 2023 interconnected based on their musical features and artist collaborations?" Insights can come from identifying clusters of similar songs, exploring the influence of artists within the network, and understanding genre diversity among the top tracks of the year.

3. Project Steps and Milestones

Data Preparation and Graph Construction (1-2 weeks)

- Objective: Prepare the dataset for analysis and construct a graph.

- Parse the dataset to identify relevant features.
- Define criteria for connecting songs (e.g., shared artist, similar genre, or feature similarity).
- Build the graph using Rust with nodes as songs and edges as shared attributes.

Implementation of Graph Analysis Algorithms (2-3 weeks)

- Objective: Implement and apply graph analysis algorithms.

- Implement algorithms like Breadth-First Search and Dijkstra’s for shortest paths to explore distances.
- Calculate centrality measures to identify influential songs and artists.
- Apply a clustering algorithm to group similar songs.

Analyze results

Testing and Validation

Objective: Test and validate the results obtained from the graph analysis.

- Cross-validate findings with external music industry reports.
- Use subsets of data (e.g., specific genres or months) to test the robustness of findings.

Each component of this project (data preparation, algorithm implementation) will be independently tested using unit tests in Rust to ensure accuracy and reliability of the functions.

I know this project idea might be a little difficult and high achieving to do with a dataset like the one I chose, so I would appreciate feedback on if this idea is feasible and if I can continue or not. Thank you!