CS432/532: Final Project Report

Project Title: Correlating Sentiment and Popularity in Music Data

Team Member(s): Serdar, Emre

I. Problem

In the evolving landscape of music streaming and social media, the popularity of a song is influenced by various factors that go beyond its musicality. Traditional measures such as downloads and sales have been supplemented by online engagement metrics like YouTube views and likes. Furthermore, the emotional content of a song's description or lyrics, quantifiable through sentiment analysis, may also play a role in its reception. This project aims to dissect the complex interplay between a song's perceived sentiment and its popularity metrics to determine any existing correlations. By analyzing the sentiment scores derived from song descriptions against YouTube views and likes, I aim to uncover patterns that could provide deeper insights into the factors contributing to a song's success or failure. Additionally, the project seeks to explore whether a song's inherent characteristics, such as its danceability, are reliable indicators of its popularity, thereby providing a more nuanced understanding of the digital music landscape.

II. Software Design and Implementation

A. Software Design and NoSQL-Database and Tools Used

The project commenced with the acquisition of raw data pertaining to various songs, encompassing both metadata and engagement metrics. To ensure data integrity and relevance, a Python script was developed and utilized to clean and preprocess the dataset. The script focused on removing duplicates, filling in missing values where appropriate, and normalizing data formats for consistency.

Once cleaned, the data was imported into MongoDB, a NoSQL database, chosen for its flexibility in handling large volumes of unstructured data and its robust querying capabilities. The schema-less nature of MongoDB allowed for the storage of documents in a JSON-like format, making it an ideal choice for dealing with the heterogeneous nature of the data collected.

B. Parts that I have implemented

Data Cleaning with Python: The initial dataset was processed using Python, leveraging libraries such as Pandas for data manipulation and NumPy for numerical computations. The script not only cleansed the data but also enriched it by calculating sentiment scores using the Sentiment npm package, which provided a quantifiable measure of the emotional content of the song descriptions.

Data Storage in MongoDB: The cleaned and enhanced data was then stored in MongoDB. The database's non-relational structure was particularly well-suited to accommodate the varying attributes associated with each song, including sentiment scores, views, likes, and danceability.

Data Analysis and Visualization: The data analysis was executed using JavaScript and Node.js, with express routes fetching data from the MongoDB collection. The retrieved data was then used to generate visualizations:
A bubble chart for sentiment analysis, plotting sentiment scores against YouTube views, with bubble size representing likes. This chart was rendered using Chart.js, a flexible JavaScript charting library. Two scatter plots for popularity analysis, mapping danceability against YouTube views and likes, respectively, to identify any correlation between a song's danceability and its popularity.

API Endpoint Creation: Express.js routes were created to serve as API endpoints that provided the necessary data for the front-end visualizations. These endpoints allowed for real-time data retrieval and rendering of the charts upon page load.

Front-end Development: The visualization interface was built using HTML, CSS, and JavaScript. The front-end design aimed at providing a user-friendly experience where the data could be interactively explored. The main webpage served as a dashboard, linking to individual pages for each analysis type – sentiment analysis and popularity analysis.

Integration with Backend: AJAX calls were set up to fetch the data asynchronously from the backend endpoints, allowing the web pages to load quickly and the data to be updated dynamically without the need for page refreshes.

III. Project Outcome

A. Sentiment Analysis with BubbleChart Visualization

The first analysis focused on understanding the sentiment of song lyrics in correlation with YouTube views and likes. For this, we utilized a bubble chart visualization, where the x-axis represented the sentiment score, the y-axis represented YouTube views, and the bubble size denoted the number of likes. This multi-dimensional approach allowed us to observe not just the positive or negative nature of the

sentiments but also how these sentiments potentially impacted viewer engagement. Larger bubbles indicated songs with more likes, providing immediate visual feedback on their popularity. The logarithmic scale on the y-axis was applied to manage the wide range of view counts, ensuring smaller values remained visible and comparative.

The outcome of this analysis revealed that songs with higher sentiment scores did not necessarily correlate to a higher number of views or likes, suggesting that the sentiment of the lyrics alone isn't a strong predictor of a song's popularity. This is a pivotal insight for music producers and artists, indicating that while sentiment may play a role in content engagement, it is not the sole factor driving viewer numbers or likes. The bubble graph is attached below.

Figure 1: Sentiment

B. Danceability and Popularity Correlation

The second analysis comprised two parts, both examining the relationship between a song's danceability and its popularity, but with popularity measured in two ways: views and likes. Two separate scatter plots were created to depict these relationships.

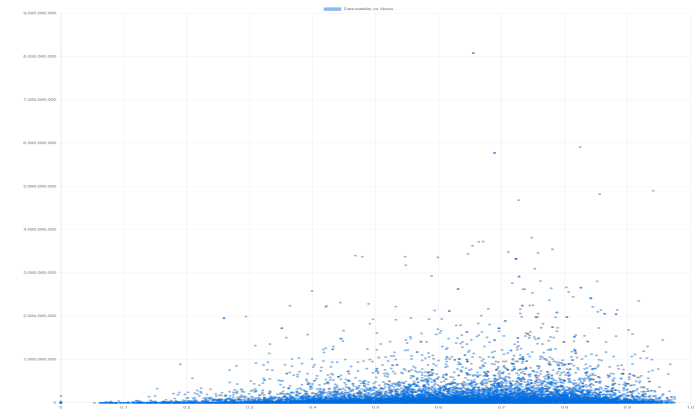

Figure 2: Correlation Dance-Views

The first scatter plot charted danceability against YouTube views. Our findings indicated a weak correlation between danceability and views, as evidenced by a correlation coefficient of 0.09. This suggests that while a song's danceability may be a contributing factor to its views, it is not a definitive indicator of its popularity.

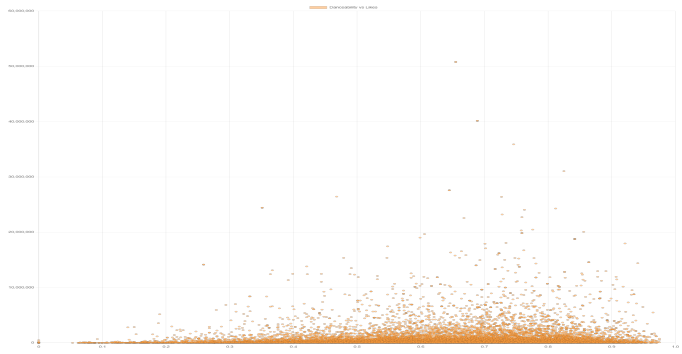

Figure 3: Correlation Dance-Likes

The second scatter plot compared danceability against YouTube likes. Here, the visualization was separated to more clearly articulate the data points. While the correlation was slightly stronger than the first case, it still remained in the lower quadrant, suggesting only a moderate relationship between a song's danceability and the number of likes it receives.

Overall, the analyses underscore the multifaceted nature of song popularity. They point to the fact that while attributes like danceability and sentiment scores do influence viewer interaction, they are part of a broader array of factors that drive the popularity of a song on platforms such as YouTube.

References

[1] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135. doi:10.1561/1500000011

[2] T. Li and M. Ogihara, "Music Data Mining: An Introduction," in Music Data Mining, CRC Press, 2011, pp. 1-10.

[3] Hu, X., & Downie, J. S. (2007). Exploring mood metadata: Relationships with genre, artist and usage metadata. Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007).

[4] J. H. Lee et al., "How Similar is Too Similar? Exploring Users' Perceptions of Similarity in Playlist Evaluation," in Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014), 2014.