Understanding the World of Music through Big Data Analysis

Varun Nagpal
3rd Year BTech, DSAI
IIT Guwahati

Yash Sharma
4th Year BTech, EP
IIT Guwahati

***Abstract*—The "Big Data Analysis of Music" project is driven by the growing intersection of music and data in the digital age. In this project, we aim to explore and analyze various music datasets to gain valuable insights into musical trends and patterns.**

Our focus is on leveraging data to understand how music genres evolve, how user preferences shape the industry, and how musicians and platforms can benefit from data-driven strategies. We believe that music, beyond its artistic value, holds substantial potential for data-driven exploration.

In the following sections, we will outline our objectives, datasets, deliverables, and expected impacts, as we delve into the world where music and data converge.

I. Motivation

Music is an integral part of human culture, and the advent of digital technology has led to an explosion in the amount of music data available. This explosion in data encompasses various aspects of music, including the vast libraries of songs, albums, and artists available online, as well as data related to music consumption, user preferences, and trends.

This wealth of data presents a unique opportunity to gain valuable insights into the world of music through the application of big data analysis techniques. By analyzing vast datasets of musical compositions, user listening habits, and music industry trends, we aim to achieve the following **objectives**:

- Understand the evolution of music genres and music industry as a whole over time.
- Identify patterns in music consumption and preferences across different demographics.
- Discover factors that contribute to the success of music tracks and albums.
- Analyze the commonalities and distinctions in music genres across diverse geographical regions.
- Develop recommendations for musicians, music platforms, and industry professionals based on data-driven insights.
- Enhance our understanding of the cultural and emotional impact of music.

To summarize, we want to analyze music from different aspects, using different combination of the multiple datasets we have found.

Additionally, we would also like to try the The Spotify Million Playlist Dataset Challenge (Continuation of the RecSys 2018 challenge). The evaluation task is automatic playlist continuation: given a seed playlist title and/or initial set of tracks in a playlist, to predict the subsequent tracks in that playlist.

II. Datasets

A. *The Million Songs Dataset*

A freely available collection of audio features and metadata for a million contemporary popular music tracks.

B. *Spotify's The Million Playlist Dataset*

A dataset and open-ended challenge for music recommendation research

C. *The Yahoo! Music dataset*

It comprises various components, encompassing Yahoo! Music User Ratings for Musical Artists, Yahoo! Music User Ratings for Songs with accompanying Artist, Album, and Genre details, Yahoo! Music ratings for songs selected by users and those selected randomly, as well as Yahoo! Movies User Ratings and Descriptive Content Information, Yahoo! Delicious Popular URLs and associated Tags, among others.

D. *MusicBrainz*

MusicBrainz is an open music encyclopedia that collects music metadata and makes it available to the public.

E. *Additional Datasets*

And more for cross dataset analysis (for eg The NRC Emotion Lexicon for emotion analysis)

III. Deliverables

Our project aims to deliver the following outcomes:

- Comprehensive data analysis reports that explore trends, correlations, and insights from the music datasets.
- Interactive visualizations to present key findings in an easily understandable manner.
- Predictive models for music popularity and trends based on historical data.
- Recommendations for musicians, record labels, and streaming platforms to optimize their strategies.

Some examples of the types of analysis we plan to add in our dashboard are:

- **General trends of the music industry**: In this section, our objective is to uncover the connections between song popularity, artists, genres, and time. By delving into the

evolution of the music industry, the impact of genres on song and artist popularity, and the temporal trends in artist and genre searches, we can identify valuable genres, track album sales trends, and analyze artist search patterns, among other insights.

- **Relation between Music, genre and emotion**: In this section of our dashboard, we explore how music genres evoke distinct emotions. We analyze lyrics, sentiment, and user interactions to uncover emotional patterns, helping users understand the emotional impact of different music genres
- **Relation between Geography and Music**: We believe that music is, to some extent, linked to its geographical origins. Consequently, we plan to examine the correlation between music and geography. Our analysis may encompass geographical references in lyrics and reviews, as well as the geographical distribution of various artists, among other factors.
- Additionally, we anticipate uncovering various other captivating insights from our data analysis.

A. *Predicted Tech Stack/ Tools*

- **Backend Data Processing**:
  - Hadoop HDFS: For storing and managing large music datasets.
  - Apache Pig: For data preprocessing and ETL tasks.
  - Apache Spark: For advanced data processing and analytics.
- **Data Visualization and Exploration**:
  - Apache Zeppelin: An interactive notebook for data exploration, visualization, and reporting.
  - D3.js: A JavaScript library for creating custom and interactive data visualizations.
  - Chart.js: A library for creating charts and graphs.
  - Other visualization tools like Tableau and Power BI.
- **Frontend Visualization**:
  - React, Angular, or Vue.js: A frontend framework for building the user interface of your music data dashboard. A library for creating charts and graphs.
- **Languages** involved include HTML, CSS, JS, Python, Scala, SQL, Pig Latin, Bash, Git etc.

IV. IMPACT

- **Informed Decision-Making:** Data analysis reports empower stakeholders to make informed decisions in areas like genre selection, artist signings, and marketing strategies.
- **User Engagement and Retention:** Interactive visualizations enhance user engagement and retention on music platforms by offering personalized insights.
- **Efficient Music Licensing and Royalties:** The music industry can benefit from more efficient licensing and royalty distribution processes by identifying which songs are most popular and where they are being streamed. This can lead to fairer compensation for artists and content creators.
- **Recommendation Systems:** Recommendations based on data analysis benefit musicians and streaming platforms, improving content suggestions and user experiences.
- **Optimized Strategies:** Recommendations help stakeholders optimize creative and marketing efforts for better outcomes.
- **Competitive Advantage:** Data-driven strategies provide a competitive edge in a dynamic music industry.
- **Educational Resources:** Project outcomes can serve as educational resources, fostering learning and innovation in music and data analysis fields.
- **Social and Cultural Impact:** Music is a significant cultural and social force. This project can impact how people discover, enjoy, and engage with music, influencing their cultural experiences and connections with others.