

# Capstone Project - Album Recommendation System: Final Report

## 1. Introduction

- **Problem statement:** Develop a recommendation model that suggests similar albums based on user input of a favorite album.
- **Approach:** Utilize data wrangling, exploratory data analysis (EDA), and unsupervised learning techniques to build and evaluate three recommendation models.
- **Findings:** Assess the performance of each model based on consistency of genres and artist recommendations.

## 2. Data Wrangling

- Identified inaccuracies in numerical features (e.g., danceability, instrumentalness) and concluded they are likely aggregated data of album songs.
- Average Rating feature found to be ineffective for distinguishing albums due to the dataset being a compilation of critically acclaimed albums.
- Number of ratings and reviews, although useful for distinguishing popular albums, not desirable as it biases against recommending obscure albums.

## 3. Exploratory Data Analysis

- Identified 'genre' and 'description' columns as the most promising features, containing detailed textual information about album genres, subgenres, moods, and themes.
- Utilized NLTK for vectorizing words in both columns to analyze their frequency distribution and explore cluster formation.

## 4. Model Development

- Three models were constructed:
  - **Model 1:** Cosine similarity with a linear kernel and TfidfVectorizer
  - **Model 2:** Non-negative Matrix Factorization (NMF) with TfidfVectorizer
  - **Model 3:** Cosine similarity with Word2Vec

## 5. Model Performance Evaluation

- Evaluation based on a list of albums from various genres:
  - Bob Dylan - Blonde on Blonde (Folk)
  - The Beatles - Abbey Road (Pop Rock)
  - Metallica - Master of Puppets (Metal)
  - Jay-Z - The Blueprint (Hip Hop)
  - Ramones - Ramones (Punk)
  - Bob Marley and The Wailers - Catch a Fire (Reggae)
- **Model 1** demonstrated the best results with consistent genre recommendations and repeated suggestions of albums by the same artist.
- **Model 2** offered more genre variety while still predicting albums by the same artists.

- **Model 3** performed poorly, yielding inconsistent genres and sometimes failing to predict albums by the same artist.

## 6. Conclusion and Recommendations

- Overall, the TfidfVectorizer model with Cosine Similarity (Model 1) proved to be the most reliable and accurate for the album recommendation system.
- Recommendations for further research:
  - Incorporate user feedback and ratings into the recommendation system to personalize suggestions based on individual preferences.
  - Explore the use of collaborative filtering techniques to enhance the accuracy and relevance of album recommendations.
- Recommendations for the client:
  - Implement the TfidfVectorizer model with Cosine Similarity (Model 1) as the primary recommendation algorithm for the system.
  - Consider integrating user feedback and ratings to continuously improve and personalize the recommendations.