Group Proposal

Submitted by Team 5:
JiWoo Suh
Sanjana Godolkar
Upmanyu Singh

What problem did you select and why did you select it?

We have chosen the task of predicting song popularity. Analyzing various features, with a primary focus on lyrics, we aim to develop a model that accurately predicts the popularity of songs. The goal is to understand how natural language processing (NLP) techniques can enhance the prediction based on textual data.

What database/dataset will you use?

Our dataset consists of song-related attributes, including track details, artist information, lyrics, popularity, album data, playlist details, and various musical features such as danceability, energy, key, loudness, and more. The dataset covers a wide range of genres and subgenres, providing a diverse and comprehensive source for our analysis.

What NLP methods will you pick from the concept list? Will it be a classical model or will you have to customize it?

The primary NLP methods include text classification, sentiment analysis, and other text-based analyses applied to the 'lyrics' column. Initially, we plan to employ classical NLP models, exploring their effectiveness in predicting song popularity. As the project progresses, we remain open to customizing our approach based on insights gained during the analysis.

What packages are you planning to use? Why?

Our toolkit will include NLTK, Scikit-Learn, Pandas, Numpy, and potentially TensorFlow. NLTK and Scikit-Learn will support natural language processing tasks, while Pandas and

Numpy will aid in data manipulation and visualization. TensorFlow may be employed for custom model development if required.

What NLP tasks will you work on?

The primary NLP tasks involve analyzing song lyrics for sentiment, theme, and other relevant information. Text classification will categorize songs based on their popularity levels. We will explore how these NLP tasks contribute to the overall prediction accuracy when combined with other musical features.

How will you judge the performance of the model? What metrics will you use?

Model performance will be evaluated using multiple metrics, including accuracy, precision, recall, and F1-score. These metrics will provide a comprehensive understanding of the model's performance in predicting song popularity across different categories. For the lyrics generation model, human judgment would be used with the cosine similarity with the real lyrics to evaluate the performance.

Provide a rough schedule for completing the project.

The project is expected to take approximately 3-4 weeks to complete, including data preprocessing, model training and evaluation, and presentation of results.

The schedule for completing the project is as follows:

Week 1: Data pre-processing and exploration

Week 2-3: Training and evaluation of Model

Week 4: Presentation of results and conclusion.