Music Genre Analysis Report

1.Introduction

This report analyses music genres based on song keywords. The aim was to cluster similar songs together, extract meaningful insights, and create a model to predict genres of new songs given their keywords. The goal is to enhance music classification and improve recommendation systems.

2.Methodology

2.1 Data Vectorization

Technique Used: TF-IDF (Term Frequency-Inverse Document Frequency)

Justification: TF-IDF was chosen over Bag of Words (BoW) because it assigns higher weight to rarer words, which are more informative than common words. Additionally, the logarithmic transformation in TF-IDF reduces data scatter compared to BoW.

2.2 Dimensionality Reduction

Technique Used: Principal Component Analysis (PCA)

Implementation: Custom PCA class using NumPy

Dimensionality: Reduced to 2 dimensions for each keyword

PCA was used to transform high-dimensional TF-IDF vectors into a 2D space while preserving maximum variance. This facilitated visualization and clustering analysis, enhancing the efficiency of the overall algorithm.

2.3 Combining Embeddings

Method: Average of PCA-reduced vectors for all three keywords

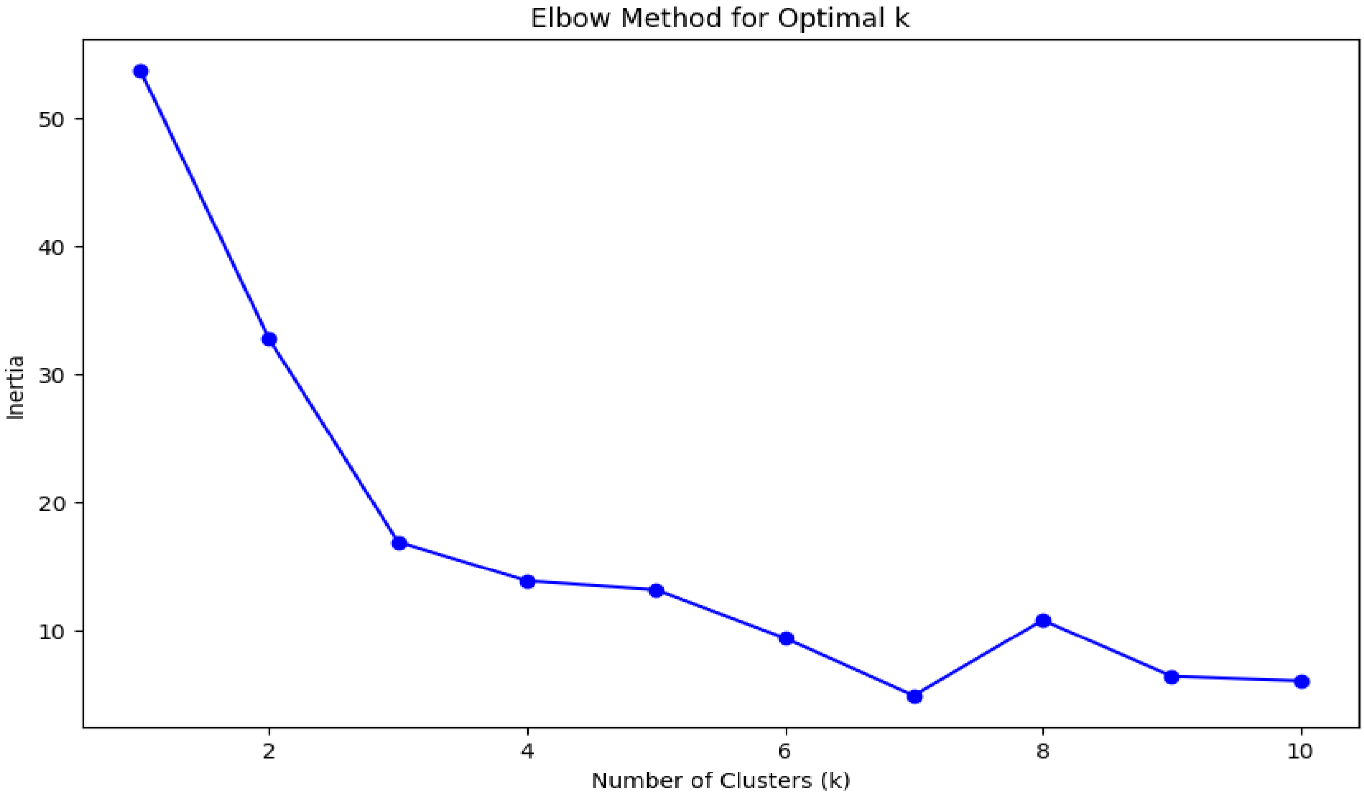
Justification: This simple approach gives equal weight (1/3) to each keyword, it gives a balanced representation of the song's characteristics. While more complex methods exist, this averaging technique offers a good trade-off between simplicity and effectiveness for our analysis

2.4 Clustering

Algorithm: K-means (custom implementation)

Optimal number of clusters (k) based on elbow method: 3

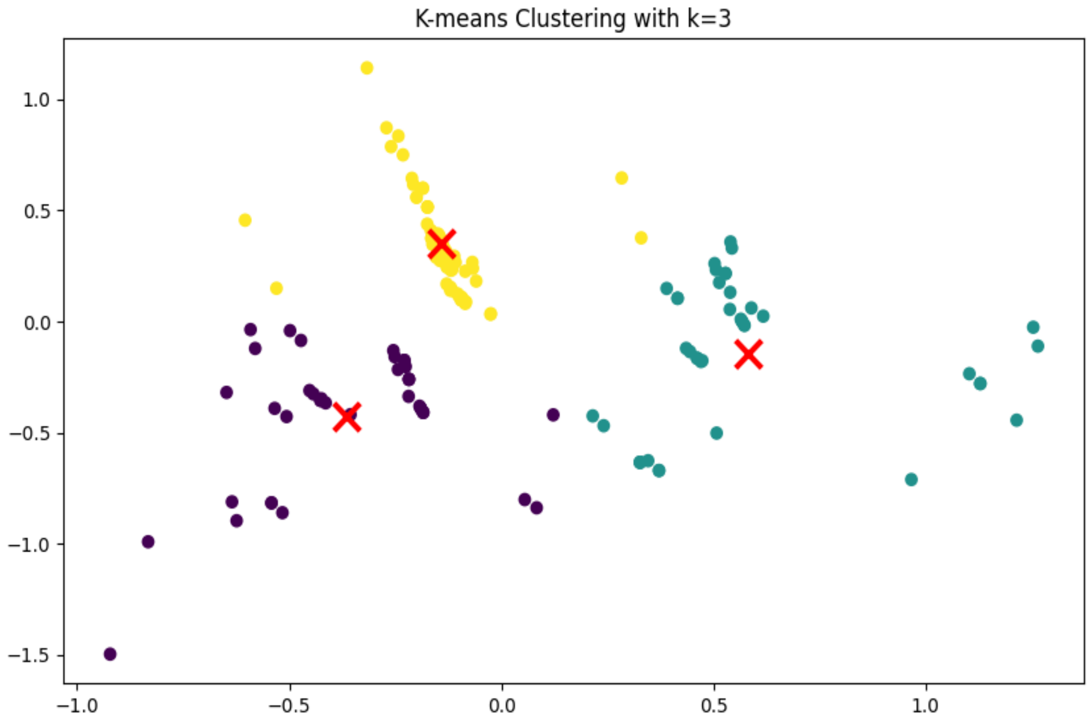
Justification for k: Based on the elbow method plot, which showed a significant decrease in inertia up to k=3, after which the decrease became more gradual although there is a decrease but there is not much steep. This suggests that three clusters provide a good balance between model complexity and explanatory power. I have tried changing the value of seed for most of seeds k=3 except for some values like 42 for which it is 4.



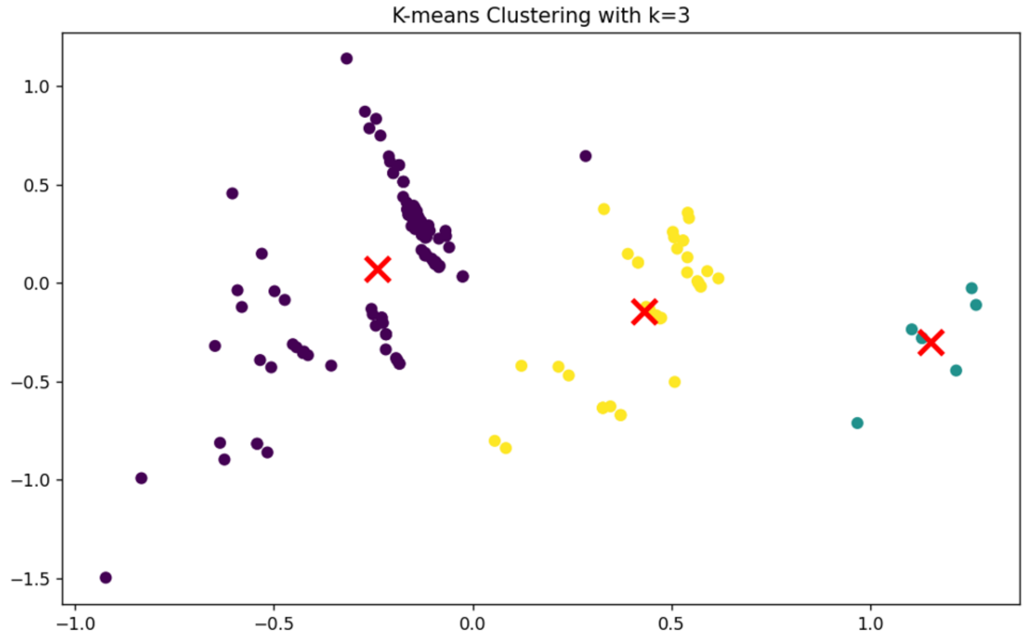
3.Results and Analysis

3.1 Cluster Distribution

The resultant cluster distribution is as follows which is similar for most of the seed values :



But, in the first run without any seed I got some very good clustering although I don't know which seed was chosen by computer in random, the first clustering I got is below:



3.2 Cluster Alignment with True Genres

As we can see in the bar graph which shows the distribution of genres across three clusters (Cluster 0, Cluster 1, and Cluster 2). The key observations are:

- Cluster 0: Dominated by the classical genre with a 30% other genres like country, hip hop are present but in smaller portions.
- Cluster 1: This cluster shows a balanced distribution of genres, with no single genre dominating the genres hip-hop and rock has highest percentage of 24.
- Cluster 2: Dominated by the pop genre, which has the a heavy count in this cluster but the rock genre also has non-negligible count.

3.3 Silhouette Score

- The average Silhouette Score is: 0.5406517063758718
- This suggests a reasonable clustering since a Silhouette score between 0.5 and 1 is suggests a reasonable clustering although not very good

3.4 Results of Genre Assignment for New Songs

- for the Keywords: ['piano', 'calm', 'slow'] the Predicted Genre is classical
- for the Keywords: ['guitar', 'emotional', 'distorted'] the Predicted Genre is rock
- for the Keywords: ['synth', 'mellow', 'distorted'] the Predicted Genre is hip-hop
- These are correct and reasonable prediction which suggests that this model is giving good and reasonable predictions

Conclusion

The analysis provides valuable insights into music genre classification, demonstrating the effectiveness of TF-IDF and PCA in clustering similar songs. It had performed pretty well given training data set is relatively a small data set. I have learnt many algorithms and ways of implementation of them in the process of completing this task.

By, E. Sai Vishesh Sharma

Roll No. : 23CS30019

THANKYOU