



Music Prediction /Recommender System

15.11.2022

Team Members- Harsh Singh, Harsh Kumar,Harsh Raj,Harsh Kumar

Content

1. Overview
2. Goal
3. Specification
4. Benefits
5. Drawbacks

6. Algorithms used

7. Working of the Model

Overview

A recommender (or recommendation) system (or engine) is a filtering system which aim is to predict a rating or preference a user would give to an item, eg. a film, a product, a song, etc.

There are two main types of recommender systems:

- Content-based filters: [Medium post](#)
- Collaborative filters: [Medium post](#)

Goals

The aim of this project is to:

1. Generate a content-based music recommender system using a dataset of name, artist, and lyrics for 28500 songs in English-hindi(1950-2019) obtained from Kaggle. The data has been acquired from LyricsFreak through scraping by the author.

Specifications

Content-based filtering is one popular technique of recommendation or recommender systems. The content or attributes of the things you like are referred to as "content."

Content-based methods give recommendations based on the similarity of two song contents or attributes while *collaborative methods* make a prediction on possible preferences using a matrix with ratings on different songs.

Benefits Content-based Method

- The model doesn't need any data about other users, since the recommendations are specific to this user. This makes it easier to scale to a large number of users.
- The model can capture the specific interests of a user, and can recommend niche items that very few other users are interested in.
- Content-based methods are computationally fast and interpretable. Moreover, they can be efficiently adapted to new items or users.

Drawbacks of Content Based Method

- The model can only give suggestions based on the user's current interests. To put it another way, the model's potential to build on the users' existing interests is limited.
- Even though this could be helpful, the value of that recommendation is significantly less because it lacks the surprise component of discovering something completely new.
- One of the biggest limitations of content-based recommendation systems is that the model only learns to recommend items of the same type that the user is already using or, in our case, listening to.

Algorithms Used

1.Data-Preprocessing:-Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

Steps Involved in Data Processing:-

1. **Data Cleaning:-**The data can have many irrelevant and missing parts. To handle this part ,data cleaning is done. It involves handling of missing data,noisy data etc.
2. **Data Transformation:-** This step is taken in order to transform the data in appropriate forms suitable for the mining process. This involves following ways:-
 - i) Normalization
 - ii) Attribute Selection
 - iii) Discretization
 - iv) Concept Hierarchy Generation
3. **Data Reduction:** Since data mining is a technique that is used to handle huge amounts of data. While working with a huge volume of data,analysis became harder in such cases. In order to get rid of this, we use data reduction techniques. It aims to increase the storage efficiency and reduce data storage and analysis costs.

CountVectorizer

CountVectorizer is a great tool provided by the scikit-learn library in Python. It is used to transform a given text into a vector on the basis of the frequency(count) of each word that occurs in the entire text.

CountVectorizer creates a matrix in which each unique word is represented by a column of the matrix, and each text sample from the document is a row in the matrix. The value of each cell is nothing but the count of the word in that particular text sample.

Cosine Similarity

We will use the **Cosine Similarity** from Sklearn, as the metric to compute the similarity between two music.

Cosine similarity is a metric used to measure how similar two items are. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. The output value ranges from **0–1**.

0 means no similarity, whereas 1 means that both the items are 100% similar.

Working of The Model

To make the recommender system function the user would provide us the name of the song and the year it was released in now our function would input a list of dictionaries of the entered song and the dataset as parameters and would output a list of ten recommended songs by our model.

Thank you !