Project Proposal - Updated Explanation

Data Mining, Big Data, and Analytics - CMPS451

Spring 2024

Team 2
⟨Key, Value⟩Knights

Name	ID	Email
Ahmed Saad	1190184	ahmed.hussein00@eng-st.cu.edu.eg
Ali Hassan	2200011	alikhalaf1234.ah@gmail.com
Hazem Montasser	2200003	hazem.mohamed002@eng-st.cu.edu.eg

To begin with, we will start by listing the dataset features which we think are relevant in general to classify a song's genre.

Dataset Features

- **Danceability:**
  - Relevant for classifying music genres based on their suitability for dancing. Genres like disco or electronic dance music typically have higher danceability scores.
- **Energy:**
  - Helpful for distinguishing between high-energy genres (e.g., punk, metal) and low-energy ones (e.g., ambient, classical).
- **Key:**
  - Can aid in genre classification as different keys are often associated with different genres (e.g., rock songs often in major keys).
- **Loudness:**
  - Useful for identifying loud genres (e.g., rock, metal) versus softer ones (e.g., classical, ambient).
- **Speechiness:**
  - Important for distinguishing spoken word genres (e.g., audiobooks, podcasts) from purely musical ones.
- **Acousticness:**
  - Valuable for categorizing music into acoustic and non-acoustic genres, such as acoustic folk versus electronic dance music.
- **Instrumentalness:**
  - Helps differentiate instrumental genres (e.g., classical, jazz) from vocal-centric ones (e.g., pop, rap).
- **Liveness:**
  - Can assist in identifying live recordings versus studio recordings, which may be indicative of certain genres or performance styles.
- **Valence:**
  - Useful for classifying music based on emotional content, distinguishing between positive and negative affect in genres.
- **Tempo:**
  - Helpful for categorizing music into different tempo-based genres (e.g., slow ballads versus fast-paced dance music).
- **Language:**
  - Useful for categorizing music based on language-specific genres (e.g., identifying songs in Spanish for Latin genres).

Answers to Questions

1. The output will be the **genre** of the input song based on the audio features of the song. The audio features in the dataset were collected using the spotifyr package in R see `https://www.rdocumentation.org/packages/spotifyr/versions/2.2.4`

We won't require any signal processing since we consider it irrelevant to the project scope. The dataset is already prepared and can be utilized as is. Additionally, we can expand the dataset further if necessary using the available package.

2. Initially, we believed that relying solely on these features for song classification would be the most effective strategy. However, it might be advantageous to incorporate the song's lyrics into the model. Alternatively, we could develop a separate model capable of predicting the song's genre based solely on its lyrics. Here are a couple of NLP-based approaches:

- Approach 1: Combine the audio features and lyrics of the song in the model.
    1. Extract the audio features from the song.
    2. Use one or more of the following to extract extra features from the lyrics:
        - Word count
        - TF-IDF score for each word
        - Sentiment analysis (might prove difficult or unnecessary)
        - Pretrained word embeddings model (e.g.GloVe, see `https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1214/reports/final_reports/report003.pdf`)
        - Bag-of-Words model (based on previous experience, this model is computationally expensive to train as it requires a lot of memory, and we believe it will not yield the best accuracy)
    3. Combine the audio features and the extracted features in the model.
    4. Feed the concatenated input vector into a decision tree model or random forest model.
- Approach 2: Use clustering techniques to group the lyrics into different genres.
    1. Extract the audio features from the song.
    2. Use one or more of the following to extract extra features from the lyrics:
        - Word count
        - TF-IDF score for each word
        - Pretrained word embeddings model
    3. Concatenate the audio features and the extracted features in the model.
    4. Feed the concatenated input vector into a K-means clustering model. K will be the number of genres in the dataset.

This is a preliminary concept derived from brainstorming. While implementing the outlined steps may not always yield optimal results, we intend to proceed with a fallback approach initially. This approach will solely utilize audio features, disregarding lyrics altogether, and employ them in a decision tree or random forest model. If this approach fail to produce satisfactory outcomes, we will then explore leveraging lyrics to improve our results.