Natural Language Processing
CSCE 5290-Section 003
Project Proposal
Team: Group-20

Project Title: Custom Song Classification and Analysis
(Unlock the Emotions and Tailor Your Music)

Team Members:

Name	Student ID	Mail ID
Naga Venkata Kanakalakshmi Murikipudi	11725119	nagavenkatakanakalmurikipudi@my.unt.edu
Naga Sai Sivani, Tutika	11703058	NagaSaiSivaniTutika@my.unt.edu

Motivation:

Nowadays, everyone is stressed out and searching for ways to unwind and get some peace of mind. One of the greatest things for many individuals to relax with is music. When we're feeling down, listening to the perfect music might help us feel better, forget our problems, and even improve our mood.

But with so many songs available, it's a challenging to choose the perfect song for our current state of mind. Scrolling through endless music can be annoying, regardless of our emotions—from happiness to sadness or just a need for some more energy.

So, Using Natural Language Processing we can built the system to help the people find music that matches to their mood, such as happy, energetic, sad, up-beat songs. Additionally, we are going to perform the sentiment analysis based on the song lyrics and make it easier for users to connect with music that speaks to them, provide caption songs so that everyone may more easily enjoy the music. Additionally, there are several languages listed in the big music list. hence, we will also categorize according to the language.

By doing this, we want to make it easy and quick for individuals to locate the perfect song, enabling them to use music to improve their everyday mood.

Significance:

The main significance of the project is providing an easy engage to the users to listen the music. As the music is a universal language which includes emotions can be expressed by lyrics - which are frequently ignored in the existing recommendation systems. We can improve the way music is categorized and suggested by using natural language processing (NLP) to analyze lyrical content. This allows us to go beyond basic acoustic parameters and concentrate on the true meaning of the song for better recommendations based on the emotions.

Objectives:

- Music Recommendation System: This dataset can be used to create a music recommendation system because it contains a variety of audio attributes like danceability, energy, and valence. Through the analysis of user preferences for specific genres and track attributes, the system can propose related tracks or even identify genres that users might find interesting.
- Song language classification: Provide a system that can identify and categorize songs automatically, giving users the option to filter music according to their preferred language.
- Automatic caption generation: Make music more accessible for those who require or prefer textual content while listening by adding captions automatically.
- Predict Emotion: Aim a functionality that predicts a song's emotional tone from its lyrics. This will provide listeners and music platforms with a greater understanding of the song's mood.
- Genre Evaluation: Analysts can learn more about how different music genres have changed over decades by examining trends in auditory characteristics like pace, danceability, and energy over time. This approach may provide light on the influences of various factors on the evolution of a genre and aid in understanding cultural transitions and changes within musical genres.
- Implement a Scalability to handle Corpus Data: Construct the system to manage massive datasets and provide scalability for millions of songs in many languages and genres.

Features:

1. Data Gathering: To achieve the objectives the required songs data need to gathered using sources like Kaggle, GitHub, Medium.
2. Data Preprocessing:
    - The Data gathered from all the sources need to clean and make sure no *dummy data, null values, and any unnecessary data*. Eliminate the punctuation from the songs because it can introduce *noise to textual data*.
    - Implementing the *Text Normalization* includes Lowercasing the data to have consistency during the analysis, *Handling the contractions* (e.g., "don't" → "do not") to perform better sentiment and emotion analysis.
    - In situations when language metadata is absent, use natural language processing (NLP) libraries, like *langdetect or polyglot*, to identify the language of the lyrics.
3. Exploratory Data Analysis:
    - The data need to be analyzed using various *data visualization techniques* to identify the patterns. Matplotlib and Seaborn are two tools we might use to visualize data. Numpy and Pandas will also be used for data analysis.
    - We are going to investigate word frequency within each mood group. To implement this using of *word clouds for each songs category* helps to show frequently using words in the songs.

4. Model Selection:
    - We may use various Machine Learning Algorithms: *SVM (Support Vector Machine), Logistic regression, Random Forest* to figure out the emotions of the song.
    - We may use keras layers, which is a widely used for implementing the *Deep Learning Models*, mostly used in the text classification.
5. Compute the performance based on the model: To ensure the model performance to have the accurate results, *Hyperparameter tuning and cross validation techniques* are used. *Performance metrics* like accuracy, precision, recall and f1score is computed.

Uniqueness:

In the Existing songs are group together based on features and using different clustering algorithms. As we want to enhance it in the better way taught to implement the above-mentioned features such as Provide recommendations to the users, Finding the emotions of the song using lyrics, Inject the Deep Learning techniques as per the requirement.

Milestones:

Milestone	Week	Task
1	Week -5	Data Gathering
2	Week -6,7	Data Preprocessing
3	Week -8	Exploratory Data Analysis
4	Week - 9	Language Recognization
5	Week -10	Analysis
6	Week -11	Model Selection and Training
7	Week -12	Evaluation
8	Week -12	Results and Visualizations
9	Week -12	Improvements If possible

Technical Features of the Project:

We construct the above-mentioned features and objectives we planned to use python and its libraries. Numpy, Pandas, matplotlib, seaborn, performance metrics, tokenization, stemming / lemmatization, exploratory data analysis, and data visualization techniques are some of the libraries we might employ for my project. We plan to use various machine learning, Deep Learning models.

About Dataset:

The dataset we have chosen is 960K Spotify Songs with Lyrics data. It's a consolidated dataset of 9 other popular Spotify songs dataset. The reason we have chosen this dataset is because it combines the attributes of songs received from official Spotify API and downloaded lyrics. The following are the links for our dataset with the source datasets.

Dataset link: https://www.kaggle.com/datasets/bwandowando/spotify-songs-with-attributes-and-lyrics

For this dataset we need to work with two files

- songs_with_attributes_and_lyrics.csv
- songs_with_lyrics_and_timestamps.csv.

The *songs_with_attributes_and_lyrics.csv* includes, attributes of songs generated from Spotify API and whole lyrics. There are 17 attributes in total. And has 955320 records ~ 960K.

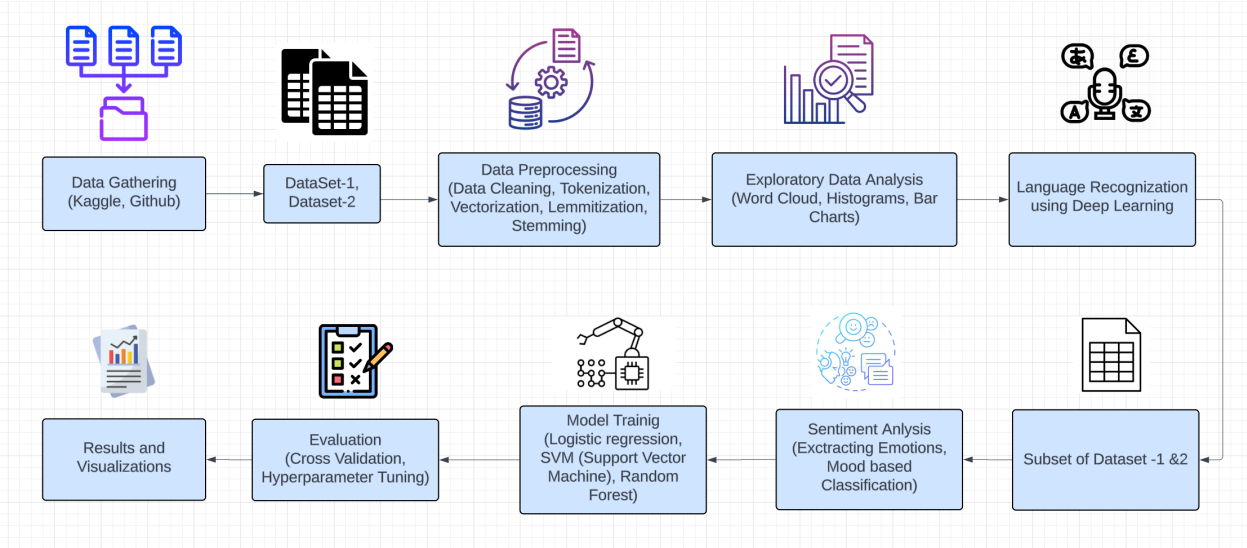
- Id – string to identify each song uniquely
- name – Name of the song
- album_name – Name of the album song belongs, blank if it doesn't belong to any
- artists – Name of the artist or set of artists.
- danceability – (0 or 1) perceptual measure if song is suitable for dancing
- energy – 0.0 to 1.0 perceptual measurement of activity or intensity
- key – ranges from (-1, -11), represents the central note or tonal center of music
- loudness – Measure of sound intensity, measured in decibels (dB)
- mode – if modality is Major (1) or Minor (0)
- speechiness –presence of spoken words in the track
- acousticness – presence of acoustic in track with confidence from 0 to 1
- instrumentalness – presence vocals (0 to 1)
- liveness – any presence of audience while recording (0 or 1)
- valence – positiveness in music (0.0 to 1.0)
- tempo – No.of beats per minute
- duration_ms – Duration of song in milliseconds(ms)
- lyrics – Lyrics of the song.

The *songs_with_lyrics_and_timestamps.csv* includes, 3 fields and 36519092 records. The 3 fields are id, startTimeMs and words.

Id – id of the song from lyrics data, not unique

- startTimeMs – start time in milliseconds of each verse given in words field
- words – verses of song

Visualization (Workflow):



References:

1. https://www.kaggle.com/datasets/bwandowando/spotify-songs-with-attributes-and-lyrics
2. https://github.com/gabminamedez/spotify-data/blob/master/data.csv
3. https://www.gigasheet.com/sample-data/spotify-dataset
4. https://huggingface.co/datasets/maharshipandya/spotify-tracks-dataset
5. https://mobidev.biz/blog/how-to-build-text-based-recommender-system-with-nlp
6. https://medium.com/@armandj.olivares/building-nlp-content-based-recommender-systems-b104a709c042
7. https://medium.com/@ghoshsou/mood-insights-from-lyrics-unveiling-the-emotional-landscape-of-music-715490b4a1a8