Predicting A Song’s Genre Using NLP

*Abstract*—This paper attempts to explore the relationship between the lyrics in a song and its genre through NLP and NLTK. Utilizing a very large dataset of song lyrics we apply logistic regression to learn which words play an importance in different genres

Keywords—Genre Prediction, Logistic Regression, NLP, NLTK

# Introduction

Grouping music is basic for online music web-based features like Spotify, YouTube Music and iTunes, as it permits them to give a rich experience. Having the option to consequently order music into various genres is valuable to these web-based services as it diminishes the time and exertion expected by record labels or streaming services to categorize music accurately. In this paper, we will aim to explore the relationship between a song’s genre and its lyrics. Genre classification by lyrics solely presents itself as a natural language processing (NLP) issue. A huge part of NLP is to assign meaning and labels to text and this is identical to genre classification of the lyrics. The goal is to build a classifier that can learn what lyrics are common in certain themes to predict a song’s genre solely through its lyrics.

In order to do this process, we used a dataset from Kaggle [1] which contained about 380,000+ lyrics from MetroLyrics. The dataset contains 362,000 songs with six columns for each song: lyrics, artist, genre, title, year released, and song index. Every song inserted into the database were released between the years 1970 and 2017.

# Issues With Dataset

## Lyrics pulled from MetroLyrics

First, Mishra straightforwardly pulled the verses from MetroLyrics. This is an issue in light of the fact that MetroLyrics assembles its information base through client input. While they guarantee to have a strong screening process utilizing interior commentators, this was a sufficient worry for us to specify. Client input permits human blunders, like spelling words in an unexpected way, particularly everyday language. A few words are likewise written in their slang spellings rather than their word reference spellings, which could prompt miscalculations.

## Songs are not in English

We discovered a few tunes in various different dialects, and tragically, the data frame doesn't give a section demonstrating the language in the verses. There are a couple of issues with this:

Words in different dialects might be spelled indistinguishably from specific English words and afterward would be sorted as a similar word, which might delude the analysis. A word with similar significance however in two unique dialects, could be ordered as two distinct words. This is deceiving as they ought to be viewed as something similar. We just eliminated English prevent words from the verses; we couldn't eliminate stop words from various dialects. However, all the non-English songs are invaluable for interpreting genres based on lyrics because they cannot be processed accurately.

## Skewed Dataset

The dataset was vigorously slanted towards Rock (131,377 tunes; 36.27%). Different kinds, like R&B and Indie, drifted around 1-3% of the example with near 6,000 melodies altogether. This distinction in number inclinations our model to foresee Rock more precisely than different kinds.

We attempted to make a decent dataset so we could prepare our model with similar number of tunes for every class; in any case, the class Indie just comprises of ~2000 melodies in the cleaned dataset, implying that a reasonable dataset would just comprise of under 20,000 tunes. Ruling against disposing of a huge number of qualified data of interest, we chose to keep our dataset with no guarantees and produce results regardless of this impediment. Chart, bar chart

Description automatically generated

1. Genre Distribution

# Cleaning Data

Because of the constraints of our dataset, we needed to do some messing around to set up the data so it can be analyzed. We needed to see any formatting prior or examples in the text and investigate inconsistencies. We needed to consider cautiously about how much and what sort of data to exclude with each cleaning step.

## Exclude unnecessary information

Initially our dataset had the accompanying data for every tune: Song Index, Title, Year Released, Artist, Genre, Lyrics. To group melodies by verses, we need to zero in on Genre and Lyrics. We thought the title and year delivered shouldn't demonstrate class, accepting creation of every kind stays steady over the long run. A musician could likewise compose melodies of different sorts, so we overlooked the artist features for this analysis.

A few tunes were missing genres, so we barred those from the dataset. We eliminated tunes with the class characterizations "N/A" (29,800 melodies) and "Other" (23,700 melodies) since we only needed information that had the both lyrics and a genre. Furthermore, a few genres were either not clear cut or not very verse weighty so we avoided these tunes from our dataset. The avoided kinds were Other, Electronic, and Jazz.

## Convert Genre Categorization and Tokenize Lyrics

* “Folk”, “Bluegrass”, “Country” 🡪 “Country”
* “Rock”, “Metal”, “Indie”, “Punk” 🡪 “Rock”

A lot of the genres were very identical to one another. So in hope of making the analysis more efficient, we jointed all of the “sub genres”.

We then tokenized each word in the verses to set up our dataset for impending data cleaning steps, for example, eliminating stop words and changing verse arranging. Utilizing the NLTK bundle, slashing up the heft of lyrics into individual words in a rundown was clear.

## Normalize Formatting

Not all verses were organized to something similar; there was a blend of capitalized and lowercase words, and a few included extraordinary characters. We changed all verses to lowercase and eliminated any unique characters. The verses contained stop words that we didn't think would be valuable to incorporate. Stop words were not significant for preparing our model as they are very normal and don't give important marks of significance or feeling. We eliminated each of the 179 prevent words from the lyrics.

A portion of the words in the verses were various types of a similar base word. For instance, strolling, strolled, and strolls are types of “stroll”. We changed all varieties of these words to a similar base word (for this situation, stroll) to more readily adjust our information with the end goal that various tenses or plural structures are deciphered as a similar word. This depends on the understanding that words in these structures convey similar importance with regards to the context of the lyrics.

# Results

## Logistic Regression

Logistic regression was used as the model with test size of 30% and training size of 70% of the sample. The model’s accuracy was done by using accuracy\_score from sklearn. This function computed the subset accuracy, and the predicted labels are identical to the labels that were given. Accuracy is defined as (True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives) [2]. On top of that, a classification report was generated to measure the precision. Logistic Regression speculates the possibility of the lyrics in the song assigned to every genre rather than greedily choosing a single genre. They are all allowed to be interdependent to one another. We were able to generate an accuracy score of 40-50%.

Table

Description automatically generated with medium confidence

1. Accuracy score

## Analyzing results

The verses of a tune and its type are not profoundly connected. This is presumably in light of the fact that a considerable lot of the most well-known words in every genre are additionally the most widely recognized words in other genres. For instance, “make” shows up as one of the most widely recognized words in each genre utilized by the program. That's what this intends, however certain tunes are named various sorts, they should sing about exact same things. This would additionally imply that the genre of a tune depends all the more intensely on its hear-able parts like instruments and melodic estimations (ie. pitch, musicality, and so on) instead of on its artistic ones like verses.

The genres in the training dataset are not so unmistakably characterized as at first suspected. As the genres were foreordained in the dataset by MetroLyrics, it is conceivable that clients transposed genres mistakenly. Regardless, a precision scope of 40-50% proposes that while verses may not without hesitation decide the genre of a tune, there is some separation in melody verses between genres. This implies almost certainly melodies in a specific genres are likelier to be about a particular theme than tunes having a place with an alternate genre.

# Future Works

There are numerous open doors for additional research and upgrades in this task. It is putting it mildly to say we are amateurs in NLP — there was a lot of we still can't seem to learn in a quickly developing field. At the hour of composing, we wrote down certain things we were keen on attempting:

Attempt bi-grams, n-grams: Instead of considering single words, we could count groupings of words; for example think about phrases notwithstanding individual words as single features. Refine the pre-handling step: Investigate ways of recognizing and avoid exceptions, to identify and discard non-English verses as well as 'good for nothing' lyrics like oh, na and so on. Look into more models that could create higher preciseness: Word2Vec Feature Extraction, Neural Networks, Decision Trees… We presently can't seem to attempt the more uncommon strategies in our undertaking, which could deliver models with better precision. Investigate the collaboration between year released, song title, artists, and genre. By making more data representations and running our ongoing procedure through additional highlights, we hope to recount a firm and adroit story.

Using sound file elements on top of lyrics to group songs to their individual genres would also improve efficiency. Ongoing research totally disregards the hear-able elements of a tune, which would be important to arrange genres. We need to attempt to characterize song lyrics in view of year of release. A few genres may be more well known in certain years than others, so adding year delivered as an element may be enlightening.

A Sentiment analysis on verses; grouping verses as negative, positive or impartial may be intriguing to explore the normal feelings of a genre.

##### References

1. 380,000+ lyrics from MetroLyrics dataset. Retreived from <https://www.kaggle.com/gyani95/380000-lyrics-from-metrolyrics>
2. Scikit-learn: Machine Learning in Python, Pedrosa *et al*., JMLR 12, 2011