Study on lyrics-based music genre classification

Sai Shruthi Umakanth
Indiana University
919 E. 10thSt., IN 47403, USA
saumak@indiana.edu

Gayatree Tiwari
Indiana University
919 E. 10thSt., IN 47403, USA
tiwarig@indiana.edu

Abstract

Content based music genre classification is gaining significant momentum in the field of entertainment, be it marketing the product, or using it in recommendation engines, promotions on social media sites, etc. With the rise of digital music, it becomes simultaneously more accessible and ripe for analysis. In this paper, we are trying to identify topics among the host of songs we have. Mining the data for inherent topics runs parallel to the concept of genres, which have traditionally been the basis for classification. We'll be analyzing this solely based on the lyrics and later hope to extend this to include analysis over a given time period.

Author Keywords

Text classification, Data mining, LDA, Genre classification

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous; See http://acm.org/about/class/1998 for the full list of ACM classifiers. This section is required.

Introduction

As language evolves over time, so does its place in popular culture. One of the barometers for this is the music in any given time period. The cultural zeitgeist is captured not only in the music, but also the lyrics that make it up. Over time, different genres change styles, sometimes merging into one another, sometimes branching off into new sub-genres. The themes are part of what is defined by the lyrics, and they aren't restricted by genre. Certain words, idioms and other linguistic devices change meaning and are appropriated by different camps.

With the rise of digital music industry, 21st century has seen a tremendous increase in the growth of online music suggestion platform. The platform which currently has genre-based song base are specified by human experts as well as novice. Though there are more demand in identifying the genre of the music and in suggesting them to the users of these platform, there is a dearth in automatic classification of songs by genre. These difficulties are since songs, as a form of an art is ever evolving and changing and containing them into a predefined classification system seems to be difficult. Even though a song can be classified into one genre, same type of music in future cannot be done based on the same criteria, because of the gap in understanding the lyrics and music and in evolution of music over time.

In our project we wish to understand this phenomenon and see how lyrical influences on genre have changed over time. On a more mundane level, we also wish to understand the lyrical signatures of each genre.

Previous Research:

1. Mining Lyrics from Web

There are many websites such as lyrics.com, lyricfreak.com, which provides lyrics transcripts including details such as artists, language, year of release etc.. but had major issue in mining the data out of the site due to copyright issues. So, in order to avoid that, LyricsFly offers well documented XML file formatted content, for research purposes.

2. Semantics Analysis of Song Lyrics

This paper explores a technique which will automatically helps in understanding the artist similarity based on a lyric. For this, they have used a publicly available dataset and compared that with acoustic similarity technique. This paper heavily analyses on the connection between song lyrics and its impact in text analysis techniques.

3. Content based music-genre classification

This paper discusses a method of feature extraction in identifying the genre classification of music, through a unique methodology called DWCHs(Daubechies Wavelet Coefficient Histograms) where it uses feature extraction and multi-class classification as a base to stimulate automatic genre classification

Dataset:

After analyzing multiple dataset which will give us both the content in the form of lyrics and the year in which the song has been released, we decided to go with 380000+ from Metro Lyrics. It has:

- Personalized songs with thousands of tracks
- Collection of lyrics paired with artist and year in which the song has been released

- By size, it is one of the larger datasets which encompasses subsets of data

The content of the dataset are as follows:

- 99 MB of data
- 380000+ songs
- 18237 unique artists
- 1000+ unique terms

Approach

We got a data set consisting of the song name, lyrics, artist, year and genre. We had to do some data cleaning - removing observations that have lyrics that are too short or simply non-existent. Removing punctuations, numbers and other characters that don't belong came next. We then parsed the lyrics into tokens and stemmed them into the root words. This list of root words for each song was then converted to a dictionary containing frequencies of each individual words. The NLTK package's stemming was used, and it was found to have some weaknesses. Some words come out misspelt rather than stemmed. We compared with two other functions for stemming (the snowball stemmer & porter2stemmer) and also tried lemmatization before stemming. However, that did not improve the performance, so we're going to stick with the first stemmer we used.

The dictionary of word frequencies thus created forms the basis for our bag of words approach using LDA. However, not all the songs are in English, and we've hit a bit of a road-block getting rid of the non-English ones. Stemming and tokenization cannot be applied to those, because we're using English-based libraries. The detect() function of the langdetect package is being used to identify language. This function also has trouble parsing some of the text. Currently, we're trying to figure out why. Once we do, we will proceed with the LDA using the dictionary of stemmed word frequencies.

Methodologies Used:

- **Latent Dirichlet Allocation (LDA)**

Latent Dirichlet Allocation is a form of topic modelling technique, which classify text in a document to a specific topic. Topic modelling is a statistical modelling method for discovering the abstract topics in a collection of documents.

The main part of LDA is to split the our lyrics.csv file into topics. For that, we are using genism and NLTK libraries as a part of LDA methodologies.

- **Document feature extraction and classification**

Main part of feature extraction and classification involves TF-IDF, which is a abbreviation for Term frequency-Inverse document frequency. It uses all the

tokens in a dataset and locates the high TF tokens. I.e the most frequently occurring token in a document. Both these TF and IDF matrices for a particular document are multiplied and normalized to form TF-IDF of a document.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

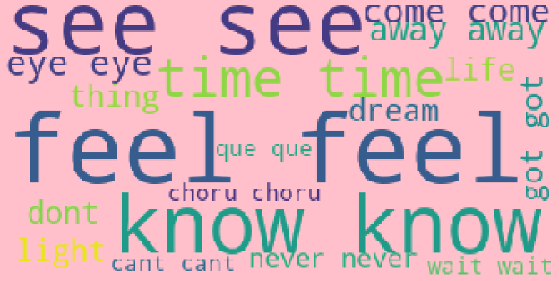
Apart from that, we are using Plotly, Lemmetizer, langdetect among others in analyzing and understanding the lyrics and in turn classifying them.

Statistical Techniques:

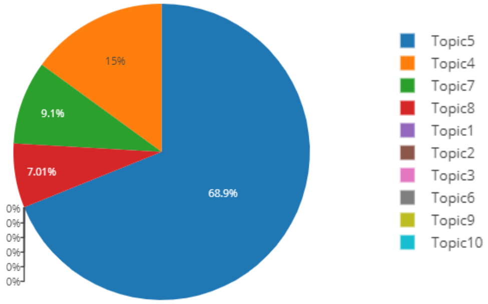
- As a first step in the process of classifying the text, we filtered out the non-english text from the lyrics part of the dataset. This is to ensure that the data is clean and to make sure that the classifier don’t label the wrong words to the actual genre classification.
- Next part, we filtered out the stopwords from the filesystem to remove the unnecessary words out of the file, so that, while classifying, only the main words can be targeted.
- Once data cleaning are done, we moved onto the text modelling. In our case an LDA model with two topics was developed. After computing the topic probabilities for all songs, we can see if this unsupervised learning, distinguish or reveal associations between music genres.
- By these step by step method, we are trying to bridge the gap between the proposal and the final result by identifying the popularity of genre and the unique one among all.
- So, in answering to the question, by finding out the most popular genre for the current data, we are setting up an environment, where we can apply large dataset with year and geographical field, to find out the demographics of the genre across globe and by that, we can apply the same model to different languages once, instead of just analyzing english lyrics.

Discussion and Result:

Based upon the above-mentioned methodologies, we were able to mine topics from the dataset targeting the lyrics column. From there, we extracted overlapping word from each topic and using WordCloud, we were able to identify the most frequent words in each of the topic.



This word cloud shows a glimpse of the top 10 words for topic #4. This can be used to compare the word frequencies of different topics at a glance. The words have been sized in proportion to how often they occur in the dataset for this particular topic.



Conclusion

In conclusion, we have found that implementing Latent Dirichlet Allocation using a bag of words approach reveals underlying topic clusters in the data that wouldn't be obvious from genre alone. This ties back to the discussion about how genres morph over time, and the results of such studies could be used to mine the shifting trends in lyrics that bring certain genres closer together and drive others further apart.

In the future, we might expand the study to include the time dimension. This will change it from a static view of current topics to a dynamic analysis of topics, which could be used to classify future lyrics into new buckets. This might be useful for related market research in the music industry. Having the finger on the pulse of trends is not just for retail, it can also be valuable information to study the change in linguistic preferences

Reference

Mayer, Rudolf, et al. "Combination of Audio and Lyrics Features for Genre Classification in Digital Audio Collections." *Proceeding of the 16th ACM International Conference on Multimedia - MM 08*, 2008, doi:10.1145/1459359.1459382.

Li, Tao, et al. "A Comparative Study on Content-Based Music Genre Classification." *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval - SIGIR 03*, 2003, doi:10.1145/860435.860487.

Logan, B., et al. "Semantic Analysis of Song Lyrics." *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*, doi:10.1109/icme.2004.1394328.

Hall, David, et al. "Studying the History of Ideas Using Topic Models." *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP 08*, 2008, doi:10.3115/1613715.1613763.

Li, Susan. "Topic Modeling and Latent Dirichlet Allocation (LDA) in Python." *Towards Data Science*, Towards Data Science, 31 May 2018, towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24.

"Multicore LDA in Python: from over-Night to over-Lunch." *Pragmatic Machine Learning*, rare-technologies.com/multicore-lda-in-python-from-over-night-to-over-lunch/.