Content Based Music Recommender System
D. Krishna Teja
Advisor : Dr. Carol Romanowski
Rochester Institute of Technology

Abstract
With the digitization of music, it has paved path for many changes, some of the significant ones being no CD's, online streaming services, and too many options. With many options, the users are confused to select what they want to hear thus the need for a recommender system had arisen. Since manual annotation to a large set of music collections is arduous, Topic Modelling is preferred to automatically identify the underlying topics present in the text sources. Lyrics are sometimes way to confusing and poetic to find the latent topics. In this project, the interpretations of songs were used to get more accurate results.

Introduction
• One of the main interest of music listeners is to understand the subject of the song, but finding the subject of a song is a very hard task when compared to finding the metadata such as title, artist, genres and lyrics.
• Many researchers have attempted to extract subject information using several supervised and unsupervised techniques on the song lyrics. While they produced quite good results, but the lyrics contain many words which are ambiguous in nature and made it difficult for the machine to understand the context of it.
• The analyzing of the interpretations of songs is an active research topic. Hence decided to perform my algorithm on the interpretation of the songs.

Technologies Used

Data

Stop word removal : Added some additional words to the common English stop word list to better the accuracy of the system. Ex: Lyrics, song (These words do not add weight to the topics).
User Tag removal : Each comment has a user tag which is not required for the system hence wrote a regex to remove all such tags.

Fig 1: Methodology

Fig 2: Interpretation from Pink Floyd’s The Wall

Fig 3: Word Cloud of that album

Fig 4: Results of the topic modelling

Data Cleaning

Data collection and cleaning was the toughest part of the project. After collecting the data, we have performed word clouds on the data set to get a much better understanding. After looking at the word clouds decided to implement several data cleaning techniques such as Tokenization, Lower case conversion, Type checker, Stemming, Stop word removal, Punctuation removal, User Tag names removal, Digits removal.

Tokenization : Large corpus broken down into single words.
Type Checker : Used a dictionary to check if all the words are valid.
Stemming : Find the originate / base word ex: lovely -> love

Implementation

Topic Modelling: Devised a statistical model using LDA which allows the set of observations to be explained by unobserved groups that explain why there is semantic connection between the topics. LDA is an uni-gram method[2], so as to tweak the accuracy of the system n-grams technique was also implemented. There was a slight increase in accuracy but due to the tradeoff between space, time and accuracy, LDA was chosen as the primary algorithm.

Sentiment Analysis: The lyrics of the song is used to find the polarity of the song[1]. The Naive Bayes classifier was used to train on a list of positive and negative words list.

Conclusion & Future Work
• When tested with the tags from another site, got an accuracy of 89.4%.
• Some albums had a contrasting Sentiment and Tag.
• Implement an application which when integrated automatically goes through the user playlists and gets similar tracks
• Also implement a sarcastic factor to check how sarcastic the song are.
• Integrate with audio component of songs for more extensive and accurate classification and check for contrast with audio mood.

References

Contact : kxd8041@g.rit.edu

Contact : kxd8041@g.rit.edu

Abstract
With the digitization of music, it has paved path for many changes, some of the significant ones being no CD's, online streaming services, and too many options. With many options, the users are confused to select what they want to hear thus the need for a recommender system had arisen. Since manual annotation to a large set of music collections is arduous, Topic Modelling is preferred to automatically identify the underlying topics present in the text sources. Lyrics are sometimes way to confusing and poetic to find the latent topics. In this project, the interpretations of songs were used to get more accurate results.

Introduction
• One of the main interest of music listeners is to understand the subject of the song, but finding the subject of a song is a very hard task when compared to finding the metadata such as title, artist, genres and lyrics.
• Many researchers have attempted to extract subject information using several supervised and unsupervised techniques on the song lyrics. While they produced quite good results, but the lyrics contain many words which are ambiguous in nature and made it difficult for the machine to understand the context of it.
• The analyzing of the interpretations of songs is an active research topic. Hence decided to perform my algorithm on the interpretation of the songs.

Technologies Used

Data

Stop word removal : Added some additional words to the common English stop word list to better the accuracy of the system. Ex: Lyrics, song (These words do not add weight to the topics).
User Tag removal : Each comment has a user tag which is not required for the system hence wrote a regex to remove all such tags.

Fig 1: Methodology

Fig 2: Interpretation from Pink Floyd’s The Wall

Fig 3: Word Cloud of that album

Fig 4: Results of the topic modelling

Data Cleaning

Data collection and cleaning was the toughest part of the project. After collecting the data, we have performed word clouds on the data set to get a much better understanding. After looking at the word clouds decided to implement several data cleaning techniques such as Tokenization, Lower case conversion, Type checker, Stemming, Stop word removal, Punctuation removal, User Tag names removal, Digits removal.

Tokenization : Large corpus broken down into single words.
Type Checker : Used a dictionary to check if all the words are valid.
Stemming : Find the originate / base word ex: lovely -> love

Implementation

Topic Modelling: Devised a statistical model using LDA which allows the set of observations to be explained by unobserved groups that explain why there is semantic connection between the topics. LDA is an uni-gram method[2], so as to tweak the accuracy of the system n-grams technique was also implemented. There was a slight increase in accuracy but due to the tradeoff between space, time and accuracy, LDA was chosen as the primary algorithm.

Sentiment Analysis: The lyrics of the song is used to find the polarity of the song[1]. The Naive Bayes classifier was used to train on a list of positive and negative words list.

Conclusion & Future Work
• When tested with the tags from another site, got an accuracy of 89.4%.
• Some albums had a contrasting Sentiment and Tag.
• Implement an application which when integrated automatically goes through the user playlists and gets similar tracks
• Also implement a sarcastic factor to check how sarcastic the song are.
• Integrate with audio component of songs for more extensive and accurate classification and check for contrast with audio mood.

References

Contact : kxd8041@g.rit.edu

Contact : kxd8041@g.rit.edu