**Music Analysis**

**Team members:**

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | Pruthvi Gollahalli Niranjana | USC ID: 3154106640 | [gollahal@usc.edu](mailto:gollahal@usc.edu) |
| 2 | Swati Mutyala | USC ID: 4306529720 | [swatimut@usc.edu](mailto:swatimut@usc.edu) |
| 3 | Md Shadman Rafid | USC ID: 4106906001 | [rafid@usc.edu](mailto:rafid@usc.edu) |
| 4 | Aarju Goyal | USC ID: 9288448242 | [aarjugoy@usc.edu](mailto:aarjugoy@usc.edu) |

**Introduction:**Companies nowadays use music classification, either to recommend new music to their customers (Spotify, Soundcloud) or simply as a product (Shazam). Determining music genres is the first step in that direction. Machine Learning has been quite successful in extracting trends and patterns from the large pool of data. Audio is a very big part of NLP, and something that’s a growing industry. **The same principles are applied in this project.**

* ***Interesting / Challenging:***

There aren’t many recommendation systems for Non-English languages. This project is done in languages other than English. It would be challenging because there is a scarcity of datasets of Non-English songs based on lyrics, therefore we have to crawl and collect data.

* ***What existing work has been done:***

Music mood recognition and classification to build a genre classification system was done based on audio feature extraction (audio signals). Mood analysis and classification is used by many companies to improve their recommendation systems.

* ***Our Contributions:***

It is an extension of the previously mentioned works. We will be exploring the relationship between a song’s lyrics and its genre. We aim to build classifiers that would learn what lyrics are prevalent in given themes to predict a song’s genre solely through its lyrics. This could also be used to compare how well the moods can be analyzed based on audio signals (done in previous works/projects) and lyrics.

**Method:**

* ***Materials:***

There are very limited datasets for song lyrics in different languages other than English. We will be using a web crawler for the site **lyricswiki.com** to obtain the lyrics of the songs. We might use BeautifulSoup in Python for web scraping. For the basic model, we would start with a dataset of size ~1000 song lyrics. We have decided to use one of these four languages for our model - Spanish, French, Czech, Hindi.

* ***Procedure:***

1. Collecting and preprocessing the data. Preprocessing would include data cleaning, stemming, introducing stopwords, etc.
2. Using NLTK, BeautifulSoup, Doc2Vec, Feature extraction using tf-idf.
3. We would start with a basic Naive Bayer’s classifier that uses Bayes’ theorem to calculate probability of a lyric belonging to a genre and then maximize log-likelihood using Maximum A Posteriori (MAP) to improve accuracy.

* ***Evaluation:***

Manual checking is a tedious and time-consuming task requiring human resource. There is no gold standard dataset available for comparing performances of the developed systems. Thus, manual evaluation will be performed on the testing data and the results.

**References:** [(PDF) Retrieving Similar Lyrics for Music Recommendation System](https://www.researchgate.net/publication/321304341_Retrieving_Similar_Lyrics_for_Music_Recommendation_System)

<https://medium.com/better-programming/predicting-a-songs-genre-using-natural-language-processing-7b354ed5bd80>

**Division of Labor Between teammates:**

|  |  |
| --- | --- |
| Pruthvi Gollahalli Niranjana | Web crawling + Pre-processing data |
| Swati Mutyala | Web crawling + Naive Bayes classification |
| Md Shadman Rafid | Web crawling + Feature extraction |
| Aarju Goyal | Web crawling + Classification |

**Future Enhancements:**

1. Sentiment analysis on lyrics
2. Classify song lyrics based on other features like “**year of release”**
3. This can also be applied to speech recognition, beat recognition, emphasis, etc.