ISMT S-117 Assignment 4: Project Outline

Version: 0.1

Date: July 23, 2020

Projectgroup: Erik Buunk

Working title: Predicting the genre of a song

Research Question

Can we use Natural Language Processing for determining the genre of a song, just by looking at the lyrics? We all have our ideas about genres and what the topics are they are singing about. Pop songs are about love, Hip Hop about gold, money and women, punk about youth and angry words against the current status quo and hard rock is all about fantasy stories. Let's see if we can use computer analytics on the lyrics of the song to determine this true or even possible.

The main Research Question will be:

> *Can we predict the genre of a song by analyzing the song lyrics, and what is the quality of that prediction?*

This will not be an easy task. We need the computer to interpret the lyrics, find patterns and make a prediction. The genres will not always be clearly defined and neither are the topics. Pop songs can be about politics. And there are multiple hard rock songs about love. The question is, can a computer make that distinction?

Determining the genre may not even be the most important goal of the research. It would be interesting to see if we can determine the artist by the lyrics of the song. Or if we bring the idea further: is this song an original or is there some copyright infringement. These more elaborate research questions are very interesting, but the datasets for this are not readily available. Or the datasets are too small. Determining the artist of a song is hard problem:

- One artist does have relatively a few songs
- There a many artist
- Songs are not always originals, but may be a cover.

The setup for this project can be used to explore the initial prediction of a limited set. If the method is predicting well, the knowledge or software can be transferred to more complex data predictions: more genres or one of the other ideas mentioned above. Fail Fast. Fail Often. We need to know if the method is feasible before we start with more complex options.

For an indication of the current baseline performance see next paragraph.

Dataset

The following primary dataset will be used:

- https://www.kaggle.com/neisse/scrapped-lyrics-from-6-genres

This dataset contains:

- Artist information. Information about almost 3000 artists, with labeled genres, number of songs, Link (to directory in the website). This Link will be used to link to the lyrics CSV
- Dataset with Lyrics with around 128000 lyrics. Fields: Link, songname, lyrics, language.

That the artists are pre labeled with genres and the lyrics are already scraped from the web, so this saves a lot of time.

Not the whole dataset will be used:

- There are Portugese and English songs (55%) in there. I will focus on the English songs.
- The genres are not equally distributed over the dataset. The majority of the artists are: Pop (25%), Rock (25%) Hip Hop (16%) . Of these artists the songs are 50% Rock, 34% Pop and 17% Hip Hop. The other genres (such as Samba) are generally the songs in Portugese.

This means with no additional information the current baseline is the random choice of genre, so we have around 33% chance that we have a correct answer. Or about 50% if we alway choose Rock.

If additional lyrics are needed there is this dataset available: https://www.kaggle.com/albertsuarez/azlyrics, but the artists need to be matched to get the genres.

Exploratory analyses

As exploratory analysis the following will be included:

- Check distribution of the data. (genres, artists, lyrics, languages)
- Data analysis: is the data usable, what needs to be cleaned. Visual inspection.
- Create a base pipeline, word counts/TFIDF
- Top words, word counts leading to a basic tokenizer.
- Cosine similarity between document genres.

Methodology

Even though the data is close to the purpose of this research, there is still a lot of noise in the data. For example: the lyrics include words such as “CHORUS”, or Guitar TABs. These will either have to be cleaned, or the tokenizers have to be tuned to limit the input of this noise.

The topic models (LDA and NMF) are interesting for this task. One would expect that topics of the lyrics to be indicative of the genres. These models will be tuned (determine optimal size of topics, iteration on word selection, choosing TF-IDF/word count). The topics will be evaluated by eye and by pipelining this into a classifier (SVM) and determining the accuracy.

The second part will focus on using document level embedding of the BERT model. The pretrained BERT embeddings can have very good performance. Since these model take more of the context into account than GloVe embeddings or just the word counts.

The document level BERT embedding will need to be generated and after. These embeddings will be used in the same classifier as. If time allows, other embedding may be used (for example Elmo or DistilBERT).

Deployment strategy

The deployment strategy will be described, since the available time doesn’t allow a complete implementation. The idea is a backend system where the model can be trained. On the other side we have a consumer website where users can select lyrics or enter (copy) their own. This input is sent back to the backend. The backend makes a prediction and stores the lyrics. The prediction is shown to the user and can click if the prediction was correct or not. If the prediction was wrong, the user can select the correct genre. This new information will be stored for future training.

Training of the model could be automated as long as the performance does not decrease.

Questions

My main question is: Is the scope of the project what is expected and if not, what needs to be added? I have tried to keep the scope small and closely aligned with the course. I have 2 weekends to finish the project, so this is my expectation of what will be feasible for me to finish, within the time constraints.