PROPOSAL NLP MINI-PROJECT: LYRIC-BASED GENRE CLASSIFICATION

DAVID VAN ERKELENS (10264019)
SHARON GIESKE (6167667)
ELISE KOSTER (5982448)

1. Introduction

This proposal describes the outline of a project researching the application of LDA and rhyme schemes to music (sub-) genre classification. Nowadays, many online music services exist, where users want music classified by genre. Manual genre classification is very time-consuming, especially with the large quantities of music available. However, if such a service misses this information, they could easily retrieve an estimated genre using lyrics-based classification.

Most earlier musical genre classification research is based on audio signals [3], which requires more storage and is more noise-sensitive. Earlier research classifying music using lyrics [1] looked at a multitude of features, but not topic models. LDA can be performed on song lyrics [2], however it has not yet been used for genre classification.
Also, using the LDA approach users may search for songs with similar lyrical themes, whose genres would not be very close otherwise. This will be an added service for users that feel strongly about certain topics.
Another interesting topic of research would be to model the correlation between different genres based on their lyrical content.

2. Objectives

The goal of this project is modeling the topic distributions for multiple music genres using lyrics.

Research questions:

- What are the most unique words and topics per musical genre?
- Are topic models a suitable feature for music genre classification using SVMs?
- Are topic models a suitable feature for subgenre classification using SVMs?
- Are rhyme schemes discriminative features of music genres?
- Which music genres correlate most and least?

3. Approach

To answer the research questions, the following steps will be used: first, a data set will be collected (possibly using a crawler, and online music websites) and processed (to clean it of noise and common words like 'the'). Then, LDA (topic model) will be implemented and used to create a distribution of topics per music genre. These distributions will then be used as a feature in training a multi-class SVM. If there is time after these results are obtained, a rhyme scheme library will be added and the rhyme schemes will be used as a feature in training another multi-class SVM. These results will be compared to the topic model classification and evaluated in a report. Also, a presentation will be prepared to concisely summarize the project and its results.

4. Deliverables

The components delivered upon completing this project will be a program that classifies a lyrics-document into a (sub-) genre as well as a report documenting the results of the project, the answers to the research questions and outlining possible future areas of research. Finally, a presentation reporting the work of the project, possibly featuring a live demonstration of the program, will be delivered.

5. Planning

Subject	Week
Gathering dataset and literature	week 1
Clean and visualize dataset	week 2
Implement LDA	week 3
Implement LDA with SVM	week 4
Add rhyme scheme feature	week 5
Start report and presentation	week 6
Finish report and presentation	week 7

References

[1] Michael Fell and Caroline Sporleder. Lyrics-based analysis and classification of music.

[2] Alen Lukic. A comparison of topic modeling approaches for a comprehensive corpus of song lyrics.

[3] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *Speech and Audio Processing, IEEE transactions on*, 10(5):293–302, 2002.