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 services offer a wide range of music, that users may want to sort by genre. However, genre information is not always available, and manual genre classification is very time-consuming, especially with the large quantities of music available. Thus, a tool that can accurately classify music into a genre would be an economically efficient way to enhance the user experience.

Most earlier musical genre classification research is based on audio signals [**?**], which requires a lot of storage compared to text documents and is more noise-sensitive. Earlier research classifying music using lyrics [**?**] looked at a multitude of features, and performed quite well, but did not use topic models nor classify (sub-) genres. LDA has been performed on song lyrics [**?**], but not in the context of genre classification.
An interesting application of the LDA approach is finding 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 modeling the correlation between different genres based on their lyrical content (topic distribution).

2. Objectives

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

Research questions:

- What are 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?
- Can songs be correlated with regard to their lyrical similarity using topic models?

3. Approach

The project will progress as follows: first, a data set will be collected (possibly using a crawler, and online music websites) and processed (to remove 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 left 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, evaluated in a report and presented.

4. Deliverables

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

5. Planning

The project takes places over the course of 7 weeks, scheduled according to the table below.

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