PROPOSAL NLP MINI-PROJECT

DAVID VAN ERKELENS, SHARON GIESKE, ELISE KOSTER

1. Introduction

This proposal describes the outline of a project researching the application of LDA and rhyme schemes to music (sub-) genre classification. Musical genre classification based on lyrics is a useful tool for the many digital music services available nowadays. If they have no genre information they can estimate it using this tool.
Most musical genre classification research was based on audio signals, which is more computationally expensive, requires more storage and is less noise-sensitive.

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

A program that classifies a lyrics-document into a (sub-) genre, a report documenting the results of the project and outlining possible future areas of research and a presentation full of funny jokes and cool pictures (there are no cats on the pictures, only word clouds, filled with high-level concepts and dreams).

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