Classification of Rap Artists by Lyrics

INM430 Coursework

Thomas Martin 16th December 2018

Total 500 words

I. Introduction

Domain

This project aims to create a model to accurately classify rap artists by their lyrics (or features generated from their lyrics) alone. Analysis of song lyrics is a popular research topic as they form an interesting area of natural language processing (NLP) research, as they have characteristics different to prose such as a greater emphasis on rhyme, structure, and repetition.

Research tends to focus on one of two tasks,

- Supervised: can songs be accurately categorised by a given piece of metadata e.g. artist
- Unsupervised: can the relationship between songs be characterised by a given piece of metadata e.g. genre

This paper is falls into the former category.

This

- Why lyrics?
- Why rap artists?
- Related literature

Related Work

Analysis of lyrics has proved to be a powerful analytical tool in the classification of songs at the level of artist, genre, and release date.

Lyrics only

Classifiers typically used include logistic

However, previous studies do not generally limit themselves by both music genre and release date

In general, previous work has used lyrics for the purpose of

Frequently follows an unsupervised task resulting in classification of artists

Michael Fell and Caroline Sporleder. 2014. "Lyrics-based Analysis and Classification of Music." Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp.620–631

Questions

- List of analytical questions to tackle

List of objectives

- Accurately classify song by artist using lyrics
- Effectiveness of word embeddings vs bag-of-words
- Find set of (derived?) features for an effective classifier
- Get comparable performance to benchmark

Other data exploration points * Differences by year? * Differences by location? * Differences by sub-genre?

II. Approach

Data Collection

Due to legal issues surrounding the use and distribution of lyrics ...

The data for this project was taken from a couple websites. To find the set of relevant artists, the script `get_charting_albums.py` finds all albums to have featured on the Billboard Rap Albums chart between January 2000 and November 2018. From this, I wanted to find a set of the top 10 most prolific artists in this time period (why???). To do this I found the full track list for each album with `get_tracklist_for_albums.py`, and grouped the total number of released tracks by artist. For these ten artists, I found the matching lyrics for each song from Genius.com using `get_song_lyrics.py`. This set of lyrics was further refined as not all lyrics were successfully returned from the website following this procedure due to parsing errors. In total, the lyrics 679 tracks for 10 artists were found.

The number of tracks for artist ranged from 39 for E-40, to 101 for Eminem

All scripts for the data collection process are in the `scripts` folder.

Analysis Strategy

With a data collected, I planned to do the following,

- Pre-process the lyrics following typical NLP techniques. Using Sklearn

Plan

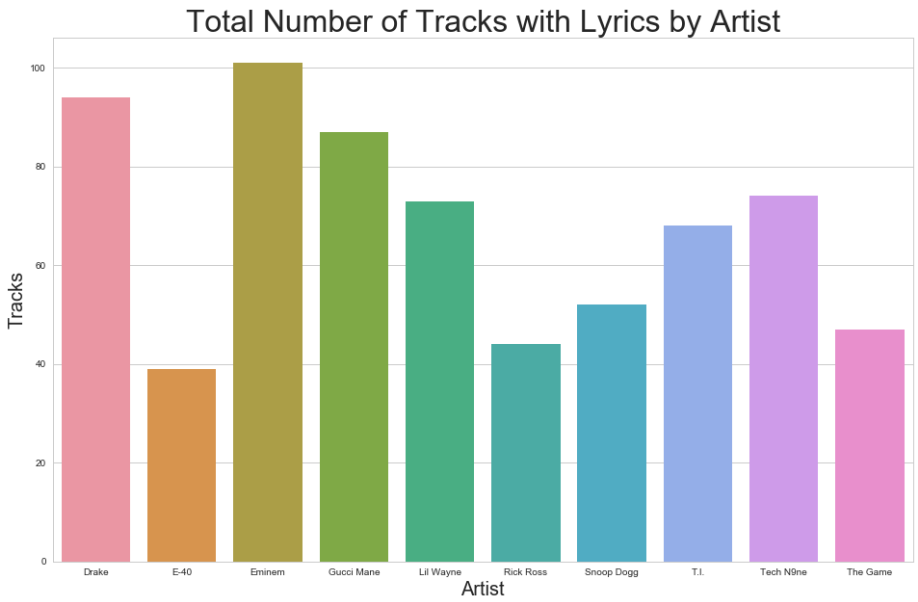

Figure 1: Total number of tracks with lyrics by artist

Data Preprocessing

Normalisation includes * Standardise casing * Removal of whitespace * Removal of punctuation * Removal of stopwords * (Did not perform lemmatisation - why?)

Vectorisation * TF-IDF * Doc2Vec

Other feature engineering

Why choice of models

Total 1000 words ## III. Analysis of Results

Analysis

Link to html computation notebook . . .

Evaluation metrics * Accuracy * Precision * Recall * F Measure

For best performing classifier produce confusion matrix for test results

Conclusions

Reflections

- Dataset contained unbalanced classes - is this evident in wrongly classified artists?
- Overall size of dataset may be an issue?
- Problems with lyrics collected?
  - User entered text - not necessarily reliable
  - Inconsistencies in text format
  - Song lyrics may also include other
- Not directly comparable to benchmarks due to scope of study: just one genre, only songs from 2000
- What features were not considered? Is this important? Why/why not?
- Reproducibility of study due to availability of data set

Relate to original objectives and motivations

References