Analysis of song lyrics to match genre

Assignment 1: Basic text processing

Jernej Janež (63130077), Rok Marinšek (63130146), Luka Podgoršek (63130189)

November 10, 2018

1 NLP task

For our assignment we decided to analyze song lyrics, extract keywords that correspond to specific genres and try to classify song by its lyrics to corresponding genre. First we found a dataset that cointained song lyrics. We preprocessed data, trained and tested a model and presented results with graphs. Some similar solutions already exist but perform similar task with neural networks or some other more complex methods. With this assignment we wanted to find out if our aproach can provide satisfactory results by using simple natural language processing techniques.

2 Data

We searched the internet for appropriate dataset. We found many different but in the end decided to use *380,000+ lyrics from MetroLyrics dataset* found on kaggle portal. This dataset had the attributes we needed to solve our task.

Dataset contained following attributes:

- song title,
- year,
- artist
- genre,
- lyrics.

2.1 Data preparation

Data we found was stored in *.csv* file. Because it contained more than *380 000* entries we decided to analyze songs that we're released in 2016 (latest songs in dataset). Afterwards we filtered songs to match predefind genres. We selected *Hip-Hop, Pop and metal* and ended up with *6845* different songs. Then we removed lyrics that we're shorter than 100 words and longer than 1000 words. This way we removed outliers in data.

When we finished data preparation and selection we focused on the text preparation. First we removed special characters from text with regular expressions, converted words to lowercase and removed punctuations. Finally we removed non-english songs. This way we ended up with **5613** different songs.

#	Genre	Number of different songs
0	Hip-Hop	2180
1	Metal	814
2	Pop	2619

Table 1: Number of songs per genre

In the end we saved filtered data into *.csv* file. In our model class we used this file as input to train our model. You can also use this file to replicate our results.

3 Model

To train our model we used preprocessed file. Model is build with logistic regression.

3.1 Train, test data and metrics

To train our model we used 80% of data and 20% to test our model. To measure score and performance of our model we used following metrics:

- accuracy,
- precision,
- recall,
- and f1 score.

In development phase we also played with regularization factor. We used above mentioned metrics to determine best regularization factor. In the end we set it to 1 (TODO UPDATE TO PROPER VALUE).

3.2 Resources, tools and corpora

We used several different python libraries. Pandas was used for data structures and data purging. Nltk corpus was used to determine stopwords and for lematization. Langdetect library was used to remove non-english lyrics. To build our model we used sklearn and preseted results with matplotlib.

4 Algorithm-Model description

5 Results

6 Github repository

Github repository: https://github.com/marok39/onj-02

Appendix

A Tmp

```
# comment
```