READ ME

Purpose

Sentiment analysis is a known task in Data Science. There have been multiple projects done to determine the polarity of sentiments on reviews (ie Hotel Reviews, Imdb Reviews). However, mostly academic papers present the case of analyzing lyrics and music. We will try to understand the difference of lyrical dataset on typical review datasets. The purpose of this project is to create models that can distinguish the emotion create by a song, using its lyrics or musical data. We train different Machine Learning Algorithms with different data, comparing them, and declaring the best.

Data

We obtain two Datasets for text data, and one dataset for musical Data.

For the textual data, the first dataset is obtained by the famous ISEAR movie reviews Dataset which contains 50.000 reviews on films, with their sentiment class. The second is a 57000+ lyrics dataset which doesn't contain sentiment classes.

Links found here:

http://ai.stanford.edu/~amaas/data/sentiment/ [1]

https://www.kaggle.com/mousehead/songlyrics

Methodology

The music dataset contains musical features and is obtained by Harvard dataverse [2].

We perform sentiment analysis on the lyric dataset, with VADER and SentiWordNet to obtain sentiment classes. We perform typical NLP tasks such as tokenizing and lemmatization, to create vectors with our textual data. Then, we try different ML algorithms to choose for our model.

We perform the same procedure on the IMDB dataset, which doesn't require sentiment analysis.

After carefully examining our parameters for vectors and ML Algorithms, we create 3 different models for textual data. Every model will be tested on every dataset.

To create the music model, we standardize our data. Since our dataset has 4 sentiment classes, we create 3 different subsets with 2 classes. Then, we create ML model for every subset, after deciding between different ML algorithms.

Results

The results showed that the classifier emerged by the SentiWordNet Analysis is the best. This fits our original assumption that even though reviews and lyrics express sentiment, they do it in a different fashion. Our lyrical classifiers performed sub-optimally on the IMDB reviews dataset, and vica-versa. We will continue to research to find the best lyrical and musical features to create ML models.

Files

The file **ImportsDefinitions.py** contains the methods and libraries used for the project. It is loaded on top of each module.

In **Vader Sentiment Analysis** and **SentiWordNet Sentiment Analysis**, we obtain the data, we perform the sentiment analysis described on the title. Then we create vectors using different metrics and finally, we create ML models. The same process is performed in **Imdb-Movie-Reviews-Sentiment-Analysis.**

These 3 files, create lists and vectors used by the **Text Data Clf Comparison Final** to determine the best model.

In **Music-Sentiment-Analysis** we obtain the dataset, create the 3 ML models and compare them on the 3 different subsets, determining the best.

References

[1]: Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).

[2]: Eerola, Tuomas, 2016, "Music and emotion dataset (Primary Musical Cues)", https://doi.org/10.7910/DVN/IFOBRN, Harvard Dataverse, V1,