CSE392 Course Project Proposal
Team Name: New Lions Prophet
Members: Isaac Duarte, Yinuo Zhang, Anthony Girard

**Introduction**

The genre of Hip-Hop music has grown to be more popular than ever. As a type of music that mainly express through lyrics, it has influenced generations of people, including us. Thus we will make a model that is going to predict which year/time period a given Hip-Hop song belongs to, and also generate word clouds of the lyrics for each of the decade since Hip-Hop originated. We believe this analysis will help people better understand how Hip-Hop evolved over the past few decades. It is a reflection of how the society has changed, as Hip-Hop lyrics largely depict the problems in the society and how they influenced people. From NLP perspective, Hip-Hop lyric is a corpus that is drastically different from other language corpuses we have seen so far. We would like to take the challenge and find out if NLP methodology works as well here.

**Methods**

The topics that we plan to integrate into our projects are the following: POS tagging, Logistic Regression, TensorFlow and RNN. The workflow is shown below:
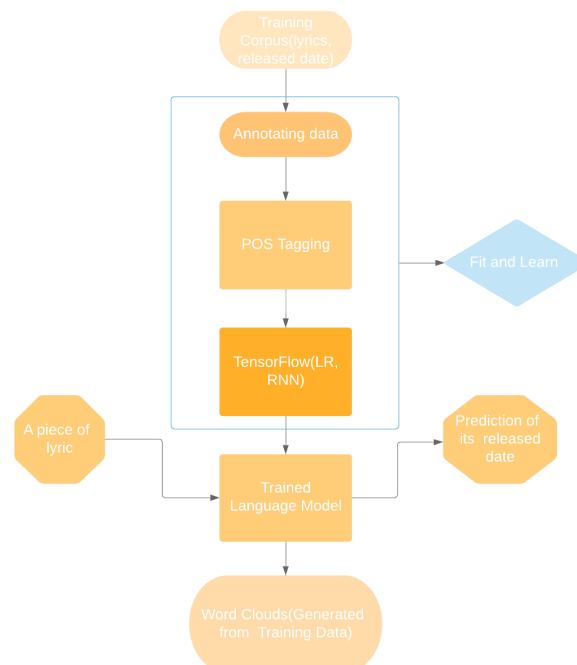


Figure 1. The Workflow of Our Project

The core of the training process is using TensorFlow to train our model. We plan on training our model in the following fashion:

The processed lyrics will first be transformed into neural word embeddings using methods such as Word2Vec. Then these vectors will be passed as input to a bidirectional Long Short-Term Memory neural network (RNN using LR). This neural network is expected to use categorical cross entropy as the loss function, the 'adam' optimizer, and accuracy as its metric. Additionally dropout regularization will be implemented to prevent overfitting.

The targeted output will be a classification of the song — specifically, the predicted year of the songs release. This will be done through a softmax output layer. We will also try to predict a range of years that the song is released, in order to make better comparison in accuracy. The work flow is described below:

- Lyrics embedded using word2vec
- Specify expected # of layers, activation functions, cost function, batch size, epochs, etc.
- Bidirectional LSTM
- Softmax output layer
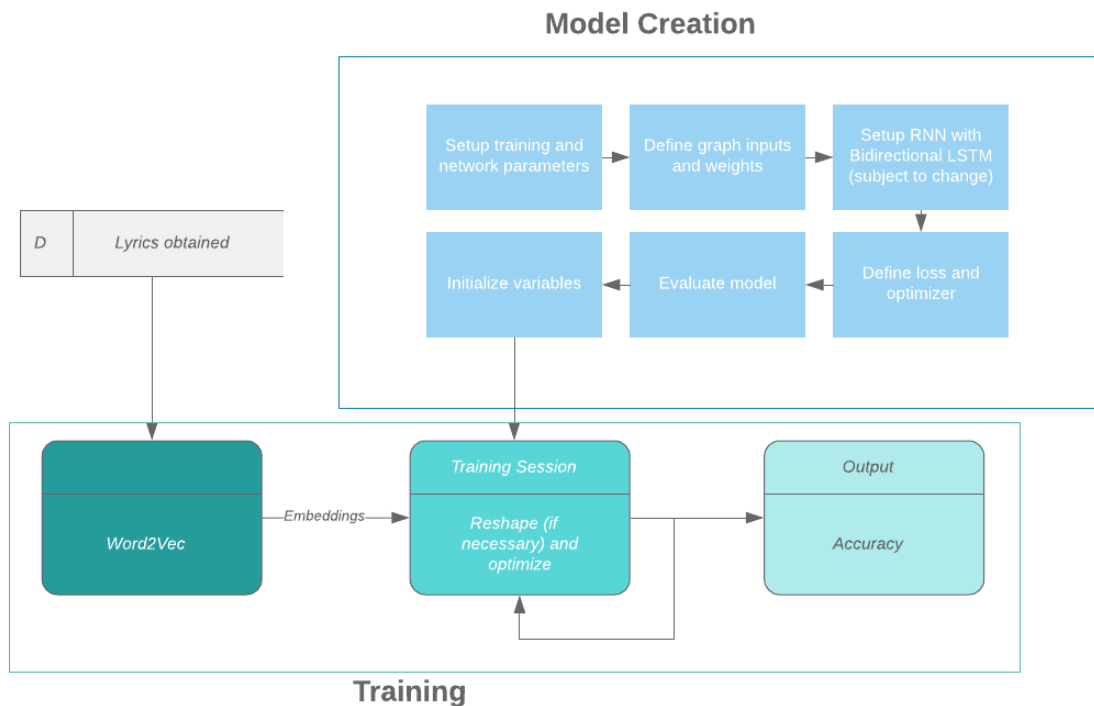- Predicts year of release



Figure 2: Model Training Workflow

**Data**

The data we are going to use is from Genius, MusixMatch and AZLyrics. The former will supply us with released date of a song, author and all the information except the actual lyrics, and the latter two will be used to obtain the actual lyrics of songs. There are existing APIs for the

first two sources, however we need to find out how to annotate the lyrics with released date ourselves. This is subject to change during production.

**Evaluation**

Results will be evaluated using a standard F1, precision, and recall score. The model will be tested on different time periods and ranges, and potentially other genres (although the primary focus is rap/hip-hop music). The model will predict a specific year which will be placed into a corresponding range, within the range its accuracy will be measured (is this song actually from this year?), but on a more general scale it will also be evaluated on how accurate it places a song into the correct range (is this song from the year 2000-2005?). Measuring it this way will account for trends, where a certain time period could have many songs with similar content.

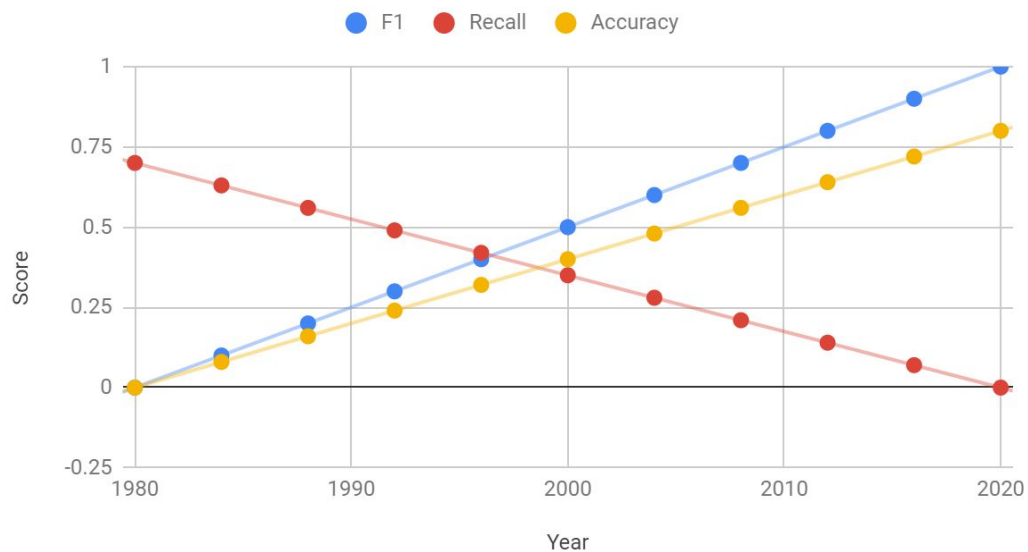|  | Precision | Recall | F1 |
|---|---|---|---|
| 2010-Present |  |  |  |
| 2000-2009 |  |  |  |
| 1990-1999 |  |  |  |
| 1980-1989 |  |  |  |
| General |  |  |  |

Table 1: Mock Evaluation Table



Figure 3: Visualization of results

**Team Plan**

This is a draft plan of what we will do individually, it may change during the implementation stage. Detailed collaboration information will be displayed [here](#).

- Isaac: Model training and code development
- Yinuo: Model testing and analyzing
- Anthony: Data processing and collection

*format*