# Proposal for Music Predictive Analysis Project based on Lyrics

LEONARD MWANGI[1,*]

[1] School of Informatics and Computing, Bloomington, IN 47408, U.S.A.
[*] Corresponding authors: lmwangi@iu.com

project-01, March 27, 2017

**Being certain that lyrics of your song will lead to the next greatest hit would boost confidence to a lot of amateur artists who are faced with fears of never making it thus never attempting to make good their creativity. With Machine Learning (ML) this can be a thing of the past, these artists would have the ability to let ML models determine the viability of their lyrics becoming the next hit based on history of other songs that have made it to top. Through training, the model can certainly determine the outcome of different songs which will be depicted in this project.**



**Keywords:** Cloud, I524

https://github.com/lmundia/sp17-i524/tree/master/project/S17-IO-3013/report/report.pdf

## 1. INTRODUCTION

When faced with the decision to forward their song to a recording company, amateur artists find it daunting due to uncertainty of whether their song would be recorded and if it is if it will make them wealthy. Having ability to run the lyrics through a predictive analysis process that would determine the viability of the song making it would be a huge win and confidence booster to many artists. That prediction is achievable by use of machine learning and creating a model that takes already greatest his and trains it to determine what makes the song successful. This would be done by analyzing the lyrics, the locality and time of release.

In this project, we will utilize machine learning to help determine the viability of a song becoming the next greatest hit based on the lyrics, time of production, locality and the artist. The project will utilize the greatest hits of all time [1] to train a model which will then be used to analyze larger dataset of random songs [2] and provide an in-depth analysis of the next possible hit. The project will utilize a Hadoop cluster deployed on Chameleon Cloud using CloudMesh to accomplish this analysis.

The following components will be utilized to accomplish the project:

- Ansible
- Apache Mesos
- Apache Spark
- MongoDB
- Million Song Dataset
- Billboard charts
- Python Scripts

## 2. COMPONENTS ROLES

### ANSIBLE

Will be used to install software packages and define roles to different nodes in the cluster.

### APACHE MESOS

Will act as the scheduler for the environment.

### APACHE SPARK

Due to Sparks ability to parallel process, we'll utilize it to process the dataset to achieve the required performance while providing in-depth analysis.

### MONGODB

MongoDB will be used as the repository for the dataset.

## 3. MILLION SONGS DATASET

This is a freely-available community maintained dataset [**?** ]. The dataset will be used by ML as the source of random songs that will be analyzed for results. This project will utilize a subset of the dataset due to time and size of our development environment.

## BILLBOARD CHARTS

In conjunction with Million Songs Dataset, Billboard charts [**?** ] will be used to determine the greatest hits of all time, which will be used to train the model on how to determine a great hit.

## PYTHON SCRIPTS

Scripts will be used to train the model and determine the next greatest hit.

## 4. CONCLUSION

Ability for amateur artists, artists and record labels to quick determine the viability of a hit is paramount to their success and missed chances due to inexperience, fear of unknows, bad song or acting when time is not ripe can be costly. Machine learning has the ability to change these outcomes, a well-trained model can help determine with high accuracy where the song will end up.

## REFERENCES