**Project Status Report**

**Team member: ezh825, dco668, mas608, szw910**

**(I) Project Objective**

Our task is to predict the winner and nominees of the GRAMMY award for Record of the Year 2017. The output of our project will be a rank of songs based on their probabilities to win the GRAMMY award. Taking 20 attributes that may exert any influence on the possibility of a song to win the GRAMMY into consideration, our project presents a detailed analysis of connections between the GRAMMY award and characteristics of songs. The result will not only help artists and producers understand how to work toward winning the award, but also yield insights to subtle differences between popular expectations and actual winners, especially in upset years.

**(II) Dataset for the Project**

We spent a lot of time writing scripts, using various APIs, and scouring the web to scrape data from various sites to obtain various attributes. The big data set we have compiled so far includes 5800 songs from 1958 to 2015. All of these songs were part of the Billboard Year-End Top 100 list. We can easily improve this data set by including songs not necessarily on that list in the near future. They also include all Record-Of-The-Year Grammy winners and nominees. Some example attributes out of the 20 we have collected include popularity, genre, word\_count, and even danceability.

The list of attributes is as following:

**4 attributes not included in the arff file:**

song\_title, spotify\_id, artist, year (from 1958 to 2015)

**15 numeric attributes:**

popularity (0-90), danceability (1-10), energy (0-10), loudness (-30 to -1),

speechiness (0-1), acousticness (0-1), instrumentalness (0-1), liveness (0-1),

valence (0-1), tempo (30-220), duration\_ms (90000-2000000),

word\_count (1-1400), reading\_ease (-2 to 150), polarity (-1 to 1), subjectivity (0 to 1)

**6 nominal attributes:**

genre (i.e. oldie, instrumental, rock, classic\_rock, country, jazz, pop, soul, funk, disco),

key (0, 1, … ,11), mode (0 for minor and 1 for major), time\_signature (1, 2, … , 5),

is\_winner (0 for No and 1 for Yes), is\_nominee (0 for No and 1 for Yes)

We plan to weigh data from recent years more when training our data by applying different weight in Logistic Regression or modify K-nearest neighbor later. At the initial period, we have decided not to consider the influence of years in order to check whether our model works. So we have scrambled the data to exclude the influence of time. When we use cross-validation, it will not take chunks of songs that are close together chronologically, which may

**(II) Models and Preliminary Results**