Report for Project 4

I use all features from segments_pitches and segments_timbre, I just use length of 500 for each line of my features because it is acceptable by R and it can help me save the time to run my code. Then I combine every row of my matrix to get a vector for each song, and the length of each vector is 12000. (For segments_pitches, the length is 6000 and for segments_timbre, it is 6000 as well. Then I combine them to get my final feature for each song.)

I used 2250 songs as training data and 100 songs as test data since our real test data will be 100 songs. Then I used LDA to find 10 topics and use ridge regression to fit the model. Training ridge model will cost about 15 minutes and when we use the predict model to predict the test data, the mean result is about 0.2068527.

If I use Lasso model to predict, the mean result is about 0.2041562 which is a little better than ridge.

But when I try to use 2000 songs as training data and 100 songs as test data, I find that the results become a little worse than before. They are 0.2452097 and 0.2441596 respectively for ridge regression model and lasso regression model.

I also try to use PCA to do feature extraction but it does not perform better than my original idea, and it will cost more time. So finally I decide not to use it.

When I use topic model to separate 2350 songs as 10 topics, it performs well. However, when the number of topic model becomes 20, the speed become slower and the accuracy rate does not improve. So finally I decide to use 10 topics.