Big Data Final Project Proposal

1. Project title
   1. Song Preference Prediction
2. Data Set
   1. kaggle\_songs.txt
   2. kaggle\_users.txt
   3. kaggle\_visible\_evaluation\_triplets.txt
3. Project idea. (This should be approximately two paragraphs)
   1. The goal for this project is to predict which songs a user will listen to based on user’s listening history. From beginning, we have three data sets. Kaggle\_songs.txt contains 386,213 records of song ID, kaggle\_users.txt contains 110,000 unique user ID, the kaggle\_visible\_evaluation\_triplets.txt contains 1,450,933 records of listening history for 79,451 users. The data set is made up by user ID, song ID, and the corresponding times of listening.
   2. We plan to separate the kaggle\_visible\_evaluation\_triplets.txt into training (70%) and testing data (30%), use the testing data to do model performance measurement. After that, we will use the model to predict which songs the left 30,549 users will listen to in kaggle\_users.txt.
4. Software/Algorithms you will need to implement or use.
   1. Collaborative filtering
5. Teammate: will you have a teammate? If so, whom? Maximum team size is  
   three students. We expect projects done in a group to be more substantial than  
   projects done individually.
   1. Total two people in the group: Rujun Dai and Yang Men
6. If working with a team, you should also describe what portion of the project each  
   partner will be doing.
   1. Yang Men
      1. Clean and input the source data for algorithm
      2. Design and finalize the model
   2. Rujun Dai
      1. Test the model
      2. Improve model performance
7. List three detailed big milestones
   1. Mar 4, Sat: Prepare the data for algorithm, ex, cleaning, organizing, etc
   2. Apr 1, Sat: Design and finalize the most accurate prediction model
   3. Apr 15, Sat: Use the finalized model to predict users in kaggle\_users.txt; Finish report writing