Progress Report

In the past few weeks, we have explored several possible project topics around song classification and recommendation. We have identified strong interest within our group to analyze the language component of songs, i.e. lyrics, to group songs into relatable concepts such as moods, occasions, and themes using techniques including natural language processing. A direct application would be to produce playlists associated with different emotions or serve specific purposes (after break-up songs, holiday music, party mix, et cetera).

As discussed in the meeting last week, our first step of this project would be to use existing song tags as targets and train the model to be able to associate song lyrics with different tags in the Million Song Database (MSD). A more advanced version would allow the user to retrieve a playlist related to a specific emotion or adjective of his/her choice. The ultimate goal is to help create an interactive and highly personalized music experience for the users, leveraging mostly the rich emotional content encrypted in song lyrics.

Pavlos pointed out some foreseeable difficulties in the project and suggested perhaps involving other elements of songs for better classification, which is something we might explore if lyrical information is not enough or if time permits for additional modeling enhancements on top of lyrics analysis.

Current stage:

In this past week, we planned out the scope of the entire project and decided on the final goal to construct an automatic tagging/recommendation system which can be used to produce playlists related to different emotions and purposes.

With this general direction in mind, we spent most of our time this past week on data processing and visualization. We obtained song tags information from Last.fm Database and stemmed lyrics data from MusiXmatch Database, and joined the two datasets together using their MSD ID. After pre-processing, we were able to compute word counts of the 5000 most frequently appeared words by song, and group songs into 4 mood categories (happy, sad, energetic and calm). We also visualized how stem words were distributed among songs with different tags. For instance, 'alon' and 'befor' were found to be closely related to sad songs, whereas 'better' and 'danc' were more represented in happy songs.

Another area we worked on this week was to research and experiment with potentially applicable model. Under Issac's guidance, we looked into many natural language processing and classification models, including SVM, Naive Bayes, LDA, pLSA and word2vec. The first two in the list are basic classification models, and the latter three are more advanced models that are more widely used in the field of NLP.

Our plan for next week:

Next week, we will dive deeper into model research and model selection. We also plan to review literatures and previous work on lyrics processing and learn about potential problems we might face during the modeling stage.

We plan to build a baseline model to have a ballpark accuracy we will be able to achieve with our current data. If the current inputs are not sufficient enough to achieve a reasonable baseline performance, we will consider scraping more data from other resources, such as the raw lyrics of the songs, novels or poems, to generate better feature inputs.