Data Mining Project Proposal

Group

Angel Dhungana

Avram Twitchell

Data

We intend to apply data mining on the Million Song dataset. Specifically, we plan to data mine the lyrics, musical characteristics (e.g. tempo), artist, year, and genres data. This data, and associated sub-datasets, can be found at https://labrosa.ee.columbia.edu/millionsong/.

Structure

In musical criticism, there are commonly accepted narratives as to how certain artists or genres influence each other. We want to see if there exists a quantifiable structure to these influences. We will do this by examining similarities in lyrics and musical characteristics such as tempo, to see if relationships exist between artists, genres, and the year, that coincides with these commonly accepted narratives.

Value

This problem is interesting on a few different fronts, but we will be focusing on two.

First, this examination can potentially give insight on how human beings interact, collaborate, borrow, steal, or draw inspiration from each other.

Second, it may offer insights on the evolution of music throughout the years. Using the time data, we could potentially see how these things shift.

Instructor Value

I think that this may be an interesting application of finding "structure" in an organic, human network. I also think that this may be a bit more of an advanced investigation of clustering, and potentially including multiple ways that the network can be compared (artist vs genre vs year) may provide interesting insights.