Jack Barry

Chris Cordaro

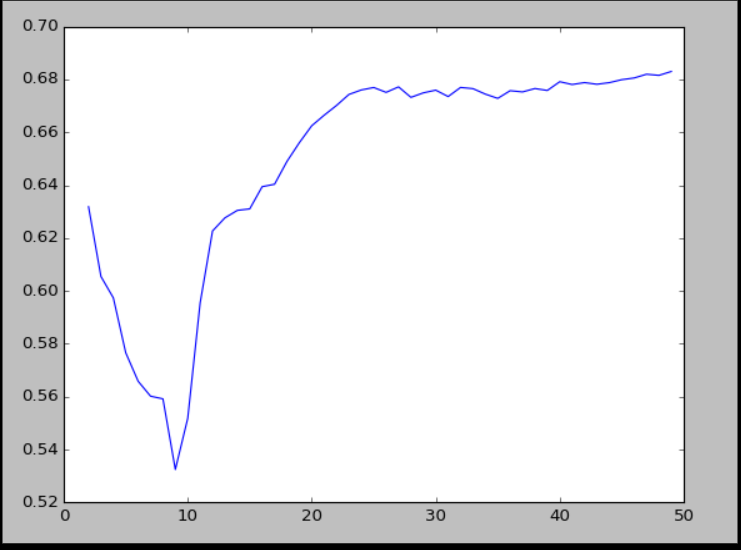
AI final project write up

For our final project we decided to use a data set from Columbia university titled “The Million Song Dataset”. The data set is formatted in an HDF5 file, and having never worked with the HDF5 format, our first challenge was figuring out how to unpack the data. We found a library that a Columbia professor made that would open the HDF5 file into a JSON object, which is then iterated over in order to unpack the file into a CSV file. To scale the data to our machines’ capacity we used a subset of the data that was 10,000 rows. With a more robust IT capacity our code could be easily updated to scale to the full dataset.

We applied PCA in order to discover what features were significant in our analysis of the data. PCA returned 10 statistically significant features, which included Song Duration, Key Signature, Key Signature Confidence, Loudness, Tempo, Time Signature, Time Signature Confidence, Album Name, Artist Name, and Title. Our first step was to remove any non numerical data including Album Name, Artist Name, and Title. For our analysis these string values had no analytic context related to what genre a song would below to and each value was unique values which could cause complications with overfitting.

Next, we wanted to get the number of features lower in order to maintain the greatest accuracy possible while analyzing our data, so we decided to remove Song Duration, Time Signature, and Time Signature Confidence. Our logic behind this decision was that a songs duration would obviously be independent from the rest of the songs features, so PCA would decide that it is a feature of importance and one that needs to be kept. However, a song’s duration does nothing to define how a song sounds, or what genre it may fit into. Additionally, Time Signature and Time Signature confidence follow that they do not depend on other features in our data set, but they fail to classify a song’s genre or provide any information regarding a song’s sonic nature.

For our analysis, we decided to implement Kmeans in order to cluster songs according to our data features. We analyzed how accurate our model was as the number of clusters changes, while at the same time not overfitting. Below is a visualization of our model’s accuracy compared to the number of clusters.



It is clear that as 10 clusters is approached there is a drop in the accuracy of our model. The optimal solution is produced somewhere around 20 clusters, suggesting that there are around 20 genres of music existing in our dataset. This seems reasonable if we are to consider more abstract genres like pop punk. As the number of clusters increased beyond its relative maxima at about 26 clusters, the accuracy slowly increased and as it plateaued it seemed as though the model was overfitting the data.

In conclusion we found this project to be a challenging and rewarding experience. One of our biggest challenges was the conversion of the HDF5 file into something we could easily work with. This dataset has been heavily used in research and academic projects, so to our advantage we were able to leverage existing documentation on how to access the data. Our results were suggestive and given more time we would like to build a framework to assign songs to certain genres and possibly sync this functionality to an application that learns a user’s song preferences and can recommend new music to listeners based on their recorded preferences. This would benefit both the listeners and new artists trying to break into the music scene such as Poughkeepsie based pop punk band, “Not Freshmen”. @notfreshmen