Song sentiment analysis

The Overall objective of our project is to create a program that will use both features from the instrumental and features from the lyrics to create an overall sentiment analysis for a given song. However, for the current scope for our midterm report, we wanted to attempt to create an unsupervised model that will output an overall sentiment analysis for the lyrics only.

DATA

ACQURIING DATA
The data was acquired from Billboard Top 100 and through the Genius API. However, some of the data was either not generated or were not in English. The final sample consisted of 66 songs

CLEANING DATA
To clean the data, we got rid of stop words, removed punctuation, replaced shortened word forms, and skipped lines indication song strucutre

WORD2VEC
The words were then processed through Wrod2Vec to reduce the dimensions of the words and to gain a vectorized version of the words

K-MEANS

RESULTS
Overall, K-means shows poor precision and accuracy. Our model generally fails to make positive predictions accurately, and correct predictions are not likely significant. Our recall score is good, though. This indicates that the model returns most of the relevant results. Our F1 score is ok, but not indicative that our model was very successful

	T0	T1
C0	1041 (TP)	830 (FP)
C1	147 (FN)	146 (TN)

Precision = 0.5564
Recall = 0.8763

Accuracy = 0.5485
F1 Score: 0.6806

GMM

RESULTS
The evaluation metrics analyzed were the same as those analyzed for K-means. Precision, recall, accuracy, and f1 score are all low and lower than those for the K-means implementation. The final probabilities for all words in the current implementation are either <0.00001 or >0.99999, with no words in between, yielding a failed model.

	T0	T1
C0	539 (TP)	692 (FP)
C1	649 (FN)	284 (TN)

Precision = 0.4379
Recall = 0.4537

Accuracy = 0.3803
F1 Score: 0.4457

DBSCAN

RESULTS
After running DBSCAN with eps = 3 and minPts = 2, our result is 93 separate word clusters. Analyzing each cluster and all the words it contains manually would be time-consuming, and upon a first glance, words within a cluster don't necessarily correspond to one similar topic or subject. Increasing minPts allows for our cluster count to reduce (ex: minPts = 8 reduces to 5 clusters), but most points are assigned the same cluster.

LIMITATIONS\FUTURE ANALYSIS

LIMITATIONS
- current method for analyzing sentiment using clustering is ineffective for K-means
- unsupervised algorithms resulted in similarly insignificant results
- Small sample size

FUTURE ANALYSIS
- Make use of derived sentiment values with a trained model that uses these sentiment values as target variables
- Implement additional features by including Spotify characteristics of each song
- Include a larger sample size that only includes English