Singer Identification
Peter Williams
8/10

Description: I want to see if we can identify the singer/band of a song based on the lyrics of their songs, and other features not related to the actual sound/style of their music. Many artists have their own individual style, and it seems like the actual sound of the music is only part of it. If we can do this well, it would also be interesting to adjust the model to predict the genre instead and see the spread of predicted genres for each individual singer, or see which singers are close.

Features: The simplest set of features for this problem would be raw/normalized frequency counts of each type/token in a song. There is a lot of potential for extracting various linguistic features beyond that, including type-token ratios, average dependency lengths, readability score, sentiment analysis scores, number of parts of speech per sentence, and whole-language lemma/token frequencies, among others. We could also run the text through a topic modeler first so that we can use topics as features instead of (or in addition to) frequency counts. We also could use features like song length, genre, gender, year the song came out, and potentially others.

Song name	Year	Length	Genre	Type-token ratio	Most frequent word	Readability score (Flesch-Kincaid)	Sentiment	Average lemma frequency	Singer (target)
Hey, Jude	1968	7:11	Rock-pop	0.57	hey	3.75	.38	843	The Beatles

Data Collection: To gather this data, we will need to collect lyrics, ideally from artists with many songs, and from a variety of artists and genres so we have enough data. There are several websites that gather lyrics, including Genius.com, lyrics.com, azlyrics.com. Some of these also have the song metadata like year, song length, and singer name that we would want. We would scrape these and then parse these with the Python packages requests and

BeautifulSoup. There may be some preprocessing of the text itself necessary to format for punctuation and sentence endings. We would then parse and tokenize the lyrics using the Python package spacy, which would then allow us to calculate the other linguistic features.