Capstone Project 1 Proposal

Are there common lyrical themes in successful popular music? If there are common themes, what are they? Are there common themes among lyrics over time periods? If there are themes among time periods can something be predicted from that? Popular music is often criticized as being predictable and repetitive in the musical sense. How well do the lyrical themes of popular music fit that stereotype? Do generational values in music differ drastically over time? Does music need to address certain themes in order to be successful? How common is it that a popular song is very different from the other songs from their generation? The answers to these questions could provide interesting insights in terms of what leads to a song being successful. They could also give historical insight into how language has changed in popular music.

The dataset that would be looked at in an attempt to answer these questions is located here (https://www.kaggle.com/rakannimer/billboard-lyrics). This dataset contains the title, artist information, lyrics, and rank in the Billboard Year End Top 100 song list of songs from 1965 through 2015. The Billboard Year End Top 100 is calculated by This dataset is mostly comprehensive, however some songs are missing lyrics, and artist information and lyrics are not perfectly scraped. For example, some lyrics may have misspelled words, or there may have been errors from when the song was scraped. I would like to add genre information on a by song basis, which would need to be scraped and added to the dataset. Songs with missing lyrics would obviously need to have their lyrics inserted unless they had none, and misspellings of words within lyrics would need to be addressed. There is also a problem with how the data is timed based on the traditional definition of decades, as this dataset is missing a portion of the sixties, and the rest of the twenty-tens have not actually occured yet.

I propose creating an algorithm which would attempt to predict what decade a song is from based off of its lyrical content and possible other information(it is possible information like runtime, artist name, song, name, and key could also contribute to this). This tool could help see what trends were common amongst decades and possibly identify songs that were ahead of or behind the times musically when they were released. The results of the generation of this model would be shared along with a slide deck and report on the creation of the algorithm.