Project Progress Report

Progress

So far, we have managed to set-up our data retrieval pipeline. We first mine billboard's website to scrape song names from their top 100 charts for a range of years. We use these song names and query Genius.com's API to obtain the lyrics for the song. We then cleaned these lyrics for processing by removing stop words and non-useful information. This way the data is in a consistent universal format throughout.

We've combined all the meta-data we have on these songs with the lyrics and consolidated it into a dataframe. We are currently working on our sentiment analysis model to assign each song a sentiment value.

Remaining Tasks

We have part of the sentiment analysis on the lyrics remaining. We will then add this sentiment value as an additional feature in our data frame for each song.

Once we are done with this, we need to create the search/recommendation engine to recommend songs based on the user's query.

Challenges Faced

One of the biggest challenges we faced while collecting the lyrics was that we weren't able to find any ready datasets online for lyrics. Most were bag-of-word representations or very limited. We decided to scrape the top 100 billboard website to get all top 100 songs and then we used those song names to query the "Genius" API to retrieve lyrics for each of the songs. This way we achieved more flexibility and customization in our retrieval.