About:

Our inspiration came from our interest in music and shmoods. These two interests culminated in us wanting to develop something that allows us, and all users to generate a playlist of songs based on our shmood. We wanted our music to be diverse, yet filled with quality songs. This is how we landed on our specific dataset.

We discovered our BillBoard Hot 100 dataset on Kaggle. Credit to Kaylin Pavlik (walkerkq on GitHub) who generated the dataset by scraping Wikipedia’s entry for each Billboard Year-End Hot 100 songs and then using that list to scrape lyrics form sites that had predictable URL strings. For example, about 78.9% of the lyrics were scraped from metrolyics.com, 15.7% from songlyrics.com, 1.8% from lyricsmode.com. About 3.6% (187/5100) were unavailable. She points out that the scraping included errors that have not been corrected (i.e. spelling errors or derivatives like “nite” instead of “night”).

We went through the dataset and manually assigned 428 songs to a shmood. The 5, all encompassing shmoods we used were: Love, Dance, Sad, Angry and Saucy. The shmoods assigned were based strictly on lyrics and did not include any sounds such as rhythm or beat. We used these songs as our training set and set the testing set to the remainder of the data. We started at a 9% accuracy rating, but after some cleaning and renaming we were able to achieve a 43% accuracy rating.