Data

The lyrical data we will be analysing has been sourced from two different locations. Each “document” of lyrics is very similar, containing metacharacters and non-lyrical components in square brackets. We also believe that we have selected sufficiently popular songs to accurately reflect lyrical trends over time.

The first is via web scraping using the Spotify and Genius lyrics APIs. By choosing a Spotify playlist (we’ve focused on the top hits of each year/decade playlists curated by Spotify themselves), our code (attached) produces a csv with song information as well as the lyrics. Unfortunately, this was not a flawless process as not all songs returned lyrics and sometimes lyrics contained non-alphanumeric characters but has given us a total of XX songs to work with.

The second location is this Kaggle dataset which can be accessed here. This dataset contains lyrics for approximately 25000 popular songs ranging from the 1950s to the 2010s. However, the dataset only contains music from 150 artists so we have randomly sampled the dataset and used XX songs as to not outweigh our primary source.

Experimental Design