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 the trends that we will try and find.

The first 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. Unfortunately, only the year of release is provided and these values are in a separate csv file so some table joining is required.

The second is via web scraping using the Spotify and Genius lyrics APIs. By choosing a Spotify playlist (we’ve focused on decade specific 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 some songs returned non-alphanumeric characters and other oddities, which we will aim to ignore in our analysis.