# Song Lyrics Dataset Analysis

Bossa Nova Dataset: https://www.kaggle.com/datasets/mcarujo/bossa-nova-lyrics
Lyrics: https://www.kaggle.com/datasets/karnikakapoor/lyrics

## Functionality:

The goal of this project is to find relevant and interesting statistics and relationships within a set of song lyrics. The project goes about this by creating a graph based on the number of lyrics each song in the dataset has in common with each other song. With this graph, the analysis computes relevant statistics such as the following:

- Number of edges (links from songs that shared lyrics with each other) in the graph compared to the maximum possible number of edges.
- The two songs in the dataset which shared the most lyrics and how many lyrics they shared.
- The song that shares lyrics with the most other songs and the number of songs it is connected to.
- The minimum spanning tree of the graph
- The normalized closeness centrality value of the graph

## Takeaways:

There are a few aspects of the graphs which are surprising. One of those aspects is the number of total edges in the graph. It is very close to the maximum possible number of edges, implying that nearly every song in the dataset shared at least one lyric with another song in the dataset.

A question that arises from this is how this variable is affected by the sample size (number of songs in the data set). I speculate that there is a curious correlation between the two variables.

An even further surprising aspect comes from the graph of the bossa nova songs. There are a variety of languages that bossa nova songs are written in. From the *Kaggle* page of the bossa nova songs dataset, there are songs from three different primary languages: Portuguese (67%), English(12%), Spanish, French, and Italian (Last three make up the remaining 21%). Despite the variety of languages these songs are written in, the number of edges of the graph were nearly the maximum possible number of edges.

## How to Use the Project:

Use one of the two read ___ csv to edges functions with the associated csv file's path, and put these edges into the vector of the datasets variable.

Upon running the project (with cargo r –release), the relevant statistics listed in *Functionality* will be legibly printed in the terminal.

If an external csv file is used:

- The user must create a new read ___ csv to edges function in the file_reading module under file_reading.rs. This is to customize the Struct the csv reader is reading to. Start off by copying the read_bossa_csv_to_edges function.
- Create a new struct in the file_reading module under file_reading.rs with #[derive(Debug, Deserialize)]. The arguments should be named after the headers in the csv file and should be of type String literal.
- Then change bossa_record to whatever name you'd like.
- Then change the elements of current_record to the associated names in your new struct.
- Then you're good to go!