RESEARCH QUESTIONS

# Can I use PCA in order to identify characteristics of artists? Could I separate out rappers and rockstars

This would require heavier processing of the song data, to remove stop words, and stemming, then I would turn each song into a file, then use perhaps the average of every song of an artist, then use dimensionality reduction on that.

# Can I see the evolution of an artist by their album using similar analysis to before

# Finding the largest phrase in common between two artists

In an attempt to make cool meme videos.

# Which artist uses the X word the most, what are the top X words used by an artist?

Will require moving of stop words

Mechanicms to answer research questions

Song comparison analysis

Simplifying an album or artists songs into a vector with words encoding

Looking at basic counts of all words as a dictionary

TODO

Little things

* Error checking if the artist already exists? Meh, not important for the usecase
* Deciding how I want to actually use the data – start with remaking the comparison program?
* Need a function to read in all the data files from a folder
  + Want to end up with a list of lists, where each list is an album
  + Later this kind of list (a list of paths to separate song files) will be sent to a function
  + If a later function is to ignore albums as being separate can just unfold this list structure,

COMPARISON PROGRAM

* Eventually can make a simple UI like I did with the vocal program!
* Just compare two artists and return the longest , can mostly borrow code for this I think
* Need to **remove the punctuation** though

NEW NOTES/STRUCTURE

The idea of the program is first to be able to scrape an artists songs lyrics into a folder which has album subfolders inside (just for later ideas).

# Data processing

This stage will depend on the intended uses for the data but could include:

* Removing stop words which are common – this doesn’t have to be done until the data is being analysed, there is no need to save the data with this bias
* Stemming – again, can be sorted out when the data is processed
* Lemmitization or something?
* **Removing repeated words, sometimes songs are like ‘aright alright alright’**
* Remove anything between brackets of any kind

Ways to summarise

* Dictionary of occurrences per word
* Somehow looking at which word is most likely to follow a sentence , becomes some kind of question answering tooll which has been trained by the lyrics

# Possible ideas with the data

* Given a dataset of songs, write a new song which is in the style
* Comparing different artists in terms of their full datasets
  + Finding matching phrases as I had at the moment
  + Using the data summaries to make simple comparisons e.g who swears more, who has more words per song
* Use an artists lyrics to answer a question, or finish a song.

OLD NOTES

Notes and ideas about the lyric program

CAN DO NOW EZ

* Function to compile lyrics and return dictionary of most used phrases

FUTURE POTENTIALS

* Using the album name too? Somehow getting that in there as well. The album labes are in bold tags so it would be relatively easy to get them, but not sure if this is an essential feature. Just makes the files tidier. Would need to split the html file based on the bold album tags
* When getting the lyrics, being smart and taking things like “driftin’ “ and turning them into ‘drifting’ so that they is a higher potential for matches.

Ideas or potential:

Another function which takes the artis folder as input and compiles all the lyrics into one list and then counts the occurrences of each word, returning a dictionary which c an then be plotted or presented.

FUNCTIONS TO MAKE:

Lyric text file to histogram/dictionary of data

Song selector – takes a base folder, and then an artist input then a song , NOTE NECESSARY~?

Simple component =