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).

Have a script which does SOLELY this. Problems to solve:

* Finding the correct artist name , ‘I couldn’t find x’ try y? Or just being able to find the right name by searching it or something – but this is minor since you could just feed in the artist page
* Getting ip banned by the website for scraping

# Data processing

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

* Removing stop words which are common
* Stemming
* 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 =