**Quest 4**

Python, although very universal and intelligent, cannot take raw data and understand it. The data has to be prepared using algorithms in order for python to be able to read it. The first step is setting up the algorithm to be able to access the data. My preferred method is to use pandas in order to read the file containing the data. Having this data means nothing to python as it is perceived as just jumbled letters. To make everything universally understandable to python, the words need to be split up, also known as parsing, and dropped to lowercase. This can be done in one live using the .split() method, and the .lower() method. This puts the individual words into a list. While python can now read the data properly, it doesn’t quite understand what it means. There will also be quite a lot of punctuation in these words. Because of this we need to import string and run the list through a function that removes letters if they are found in the string.punctuation library. From here we would want to count each word and store it in a dictionary so that python can find the main topic in the data. This is done by using an algorithm that puts each new word it comes across into the dictionary and has its value set to one, and then incrementing the dictionary description by one each time it comes across each word. Even though python can now see the most common words, it’s going to be full of words like “a”, “the”, etc. to get rid of this there needs to be a blacklisted list of words that the algorithm will search through. This is done during the counting process where if the word is blacklisted, it is not put into the dictionary. This information will have many forms of each word, we want to simplify it by combining different forms of the same word, into its base form. You can either use the process of stemming or lemmatizing. Stemming will attempt to find the common base to a word. While this is very fast, when it comes to data analysis, it’s not very effective given that you can end up with words that are not actually words. Lemmatizing is much slower, but will be much more accurate. Lemmatizing finds the dictionary root of a word, and can differentiate between noun, verbs, and other parts of speech. This would be done before the information is put into the counting dictionary. This data is essentially ready for python analysis. The last thing that needs to happen is spam detection. An algorithm will be used to take the text data and compare it to data already analyzed and determine if the data lines up with what would be considered spam. This is done by comparing the top words to the top words of flagged spam data. While all this data prep seems overwhelming, in the end it is relatively straightforward.

