**Name : Bentic Sebastian**

**Project Title**: To create a Naive Bayes Classifier in Spark to track Twitter Stream for Amazon complaints

**Question:** How can I classify the nature of tweets connected to the company Amazon?

Methods: Bernoulli Naive Bayes( 0 or 1)

    Multinomial Naive Bayes( range of values)

**Introduction:**

I wanted to use Spark to analyse Twitter data.

I received access to the Twitter API, and wrote a script(GatheringDataFromTwitter.ipynb)

to generate a file with around 5000 tweets(result\_Tweets\_full).

I searched for posts directed to '@amazonHelp', and collected all resulting posts.

The tweets were interesting, since I could use a Naive Bayes Classifier to analyse the topics that were most frequently mentioned.

This method would be useful for Amazon to monitor which segments of their company need to be improved

Once I had the tweets, I wrote a script in PySpark (main.py) that manipulated the Twitter data.

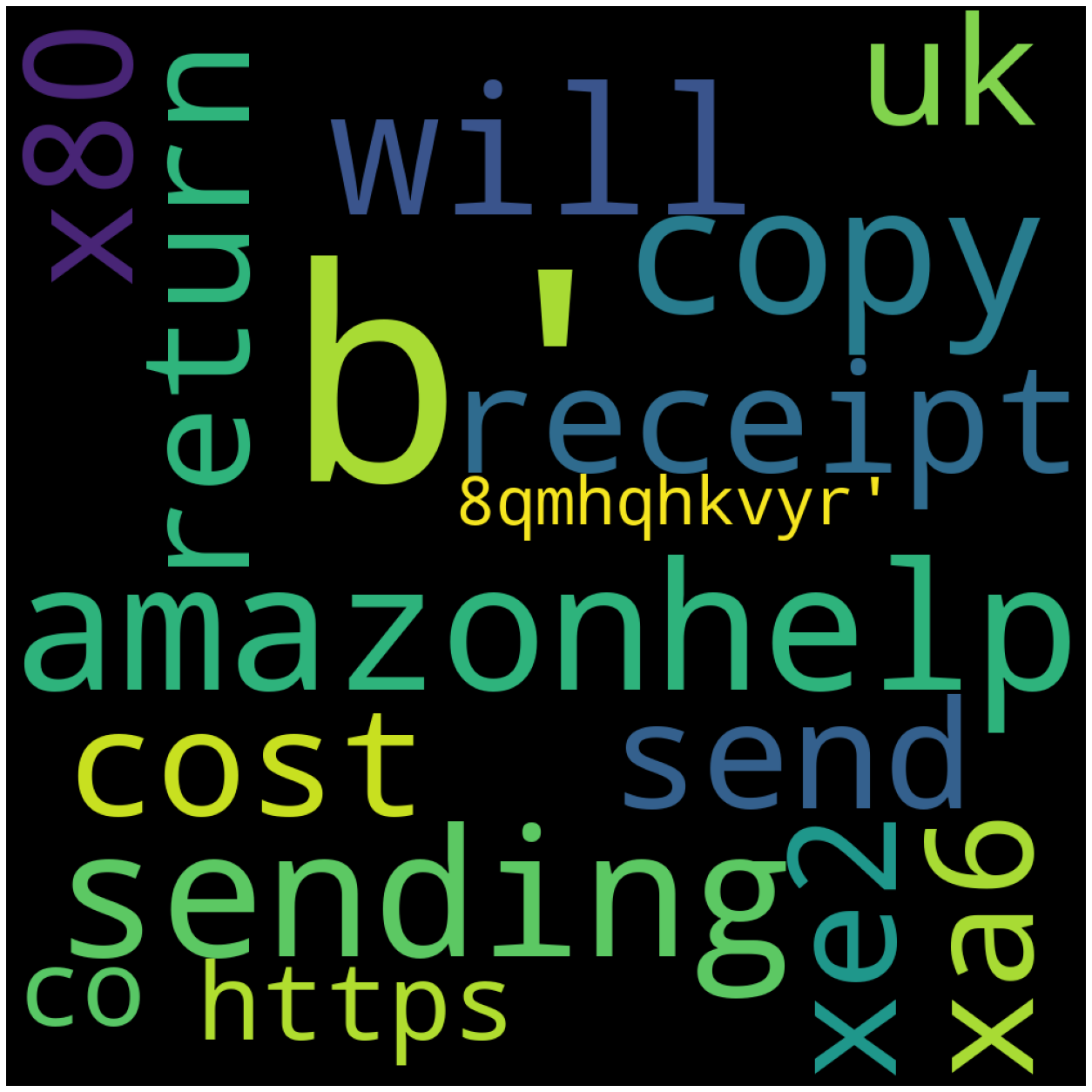
**Results:**

I have been able to collect the frequencies of words, after removing stop words etc.

The results are saved in 'wordcount\_16thDec.out'

The most common words are “sending”, and “return”. Surpringly, “Receipt” also shows up. This may suggest that Amazon should focus on making sure that customers are receiving the correct receipts for their purchases, either electronically or in the packages.

Although I wanted to create a complete Naive Bayes Classifier , I was not able to do so.



**Future work:**

I would like to continue to implement this classifier, by calculating the Idf values and use the join command to find the tf-idf values.

The Tf-Idf values are useful in vectorizing the data,

in order to find the priors and use it for the Bayes Classifier.

Another future work I would like to be involved in is in creating a Sentiment Analysis algorithm, to predict which tweets are positive and negative.