**Assignment 2 - Report**

**Naive Bayes –**

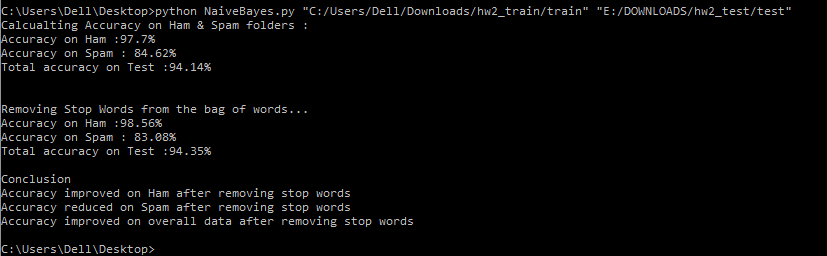
It is an algorithm that uses Bayes' theorem to classify the dataset. Naive Bayes classifiers assume naïve & independence between features of data points. Naive Bayes classifiers popularly used for spam filters, text analysis etc. Naive Bayes is also known as simple Bayes or independence Bayes.

For this assignment -

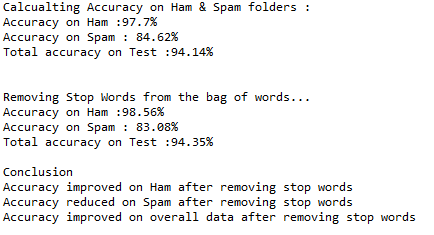
* Code written in NaiveBayes.py,
* Accuracy 94.35% (combined in both ham and spam)
* Filtered the words after removing all the punctuations and other characters like +,- etc.
* Used Regex [A-Za-z0-9\’]+ to extract words from documents

**Command executed** in cmd – python NaiveBayes.py "C:/Users/Dell/Downloads/hw2\_train/train" "E:/DOWNLOADS/hw2\_test/test"

Screenshot



**Output in Spyder –**



**Description on accuracy improvement after removing stop words from the set-**

* Accuracy improved on Ham after removing stop words however accuracy on spam reduced after removing stop words
* Overall accuracy has been increased by nearly 0.2% after removing stop words
* Reason why accuracy changed on removing stop words is that these words are general words and it is always expected that their count could be high or less in any of the mails. So, agenda behind removing stop words is to get the more accurate results on predictions.

**Logistic Regression**

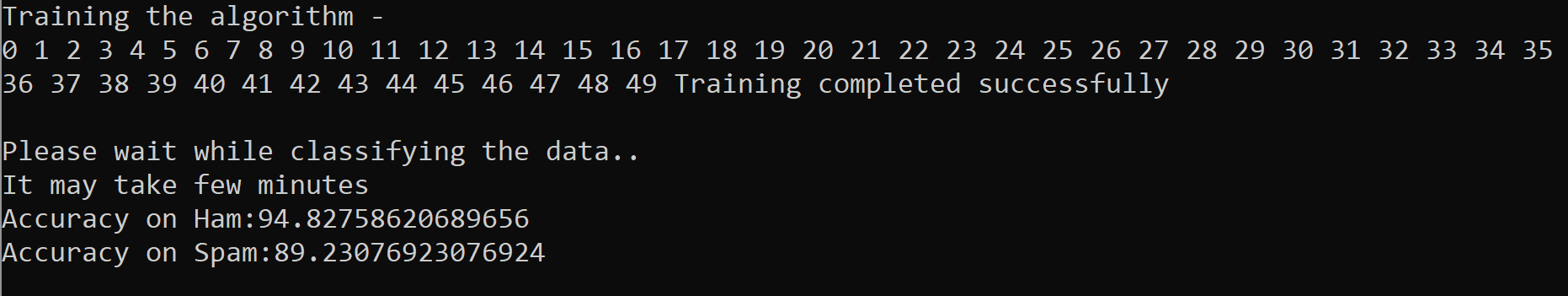
Logistic regression is a predictive analysis algorithm. It is used to describe data in the form of the vectors and making predictions with the use of sigmoid function.

For this assignment -

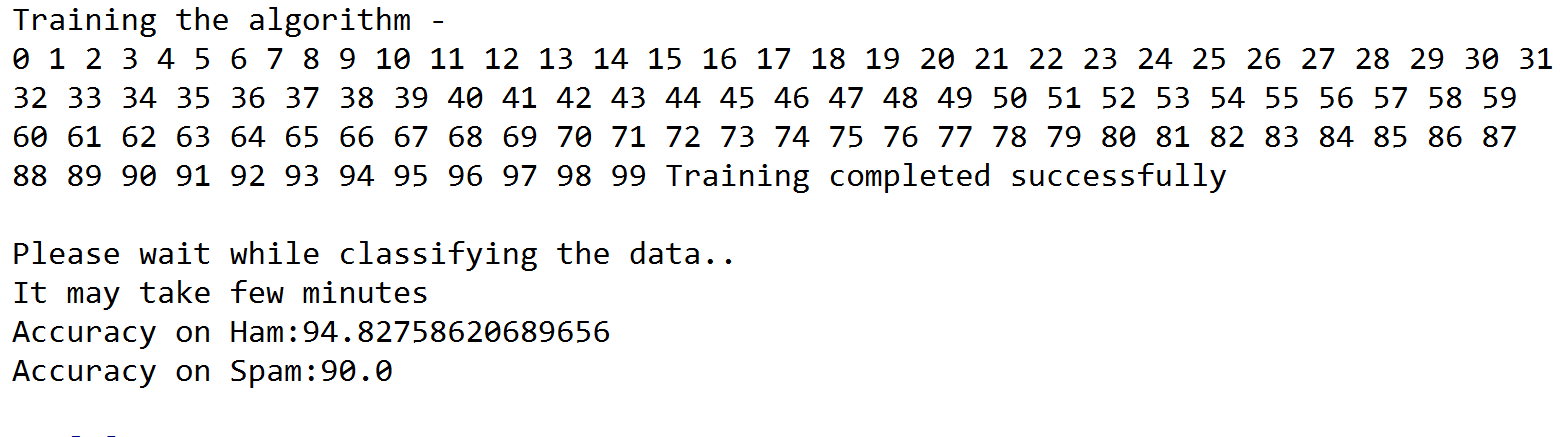
* Code written in LR.py
* Accuracy varies in LR depending on the different value of weight initialized, lambda, stop words with yes or no combined in both Ham & Spam
* Filtered the words after removing all the punctuations and other characters like comma, full stop, +, - etc.
* Used Regex [A-Za-z0-9\’]+ for above
* Command executed – python LR.py " C:/Users/Dell/Downloads/hw2\_train/train " " E:/DOWNLOADS/hw2\_test/test " "yes" "3" "100"
* Please note that on my system it took nearly 65 minutes for 100 iterations and making prediction.

Outputs for different values of λ, with/without removing stop words -

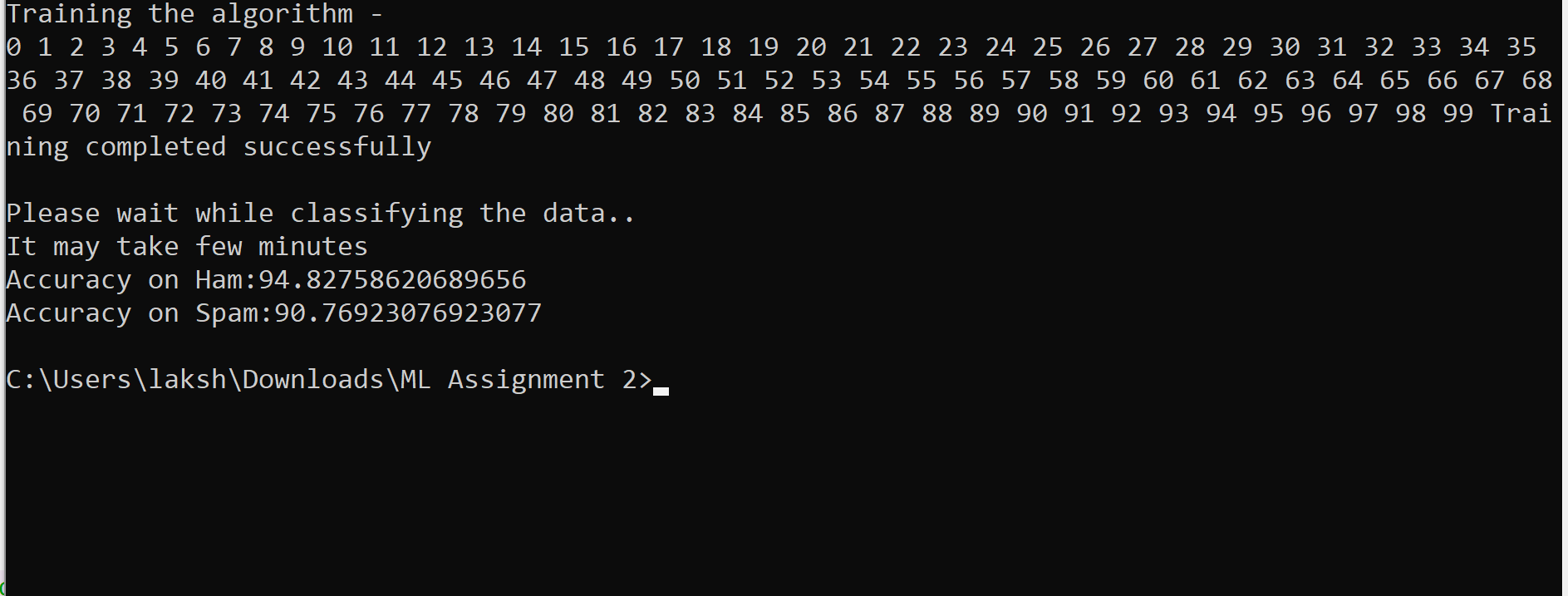
Lambda = 0.1, iteration 50, stop words – Yes



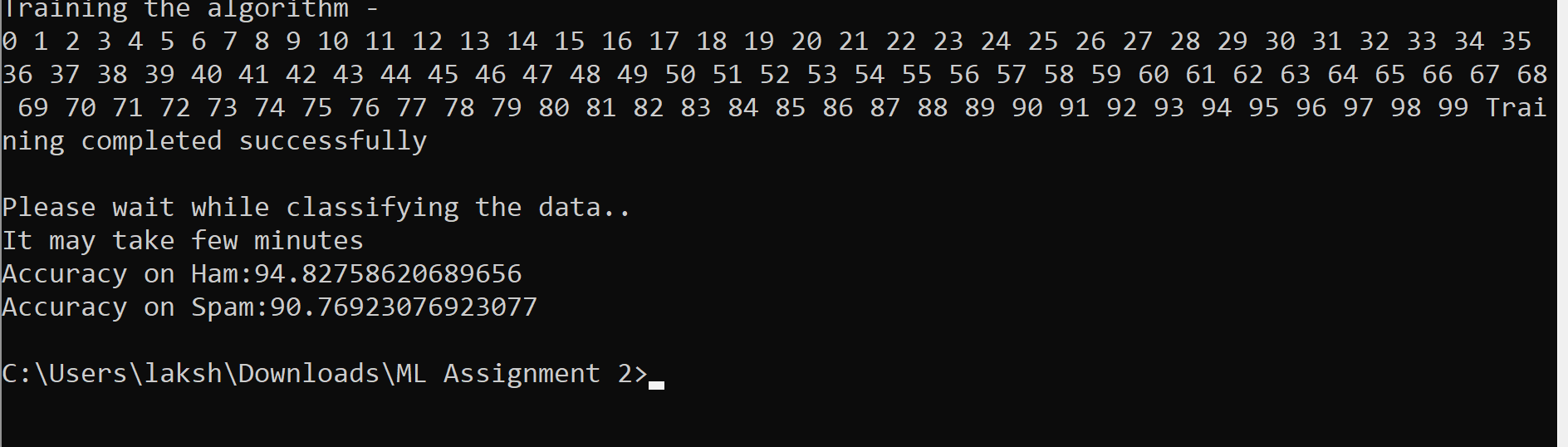
Lambda = 0.1, iteration 100, stop words – yes



Lambda = 3, Iteration 100, Stop Words – Yes



Lambda = 3, Iteration 100, Stop Words – no



**Description on accuracy improvement after removing stop words from the set-**

With change in value of the number of iterations the accuracy is affected. I observed that the more the number of iterations the more amount of accuracy is obtained on the data given. Stop words are the list of generic words hence removing them could impact the accuracy, however is not mandatory that accuracy will improve on removal of stop words from the pool of words. Also, in case of logistic regression, the accuracy will also be dependent upon Lambda and number of iterations.