INFORMATION RETRIEVAL

Assignment 4: Spam Detection (Text Classification)

Instructions:

- **This is an individual assignment.**
- **You are not allowed to use NLTK, scikit-learn or any other machine learning or language modeling toolkit.**
- Make a report in PDF format, clearly mentioning question/ part number against your answers.
- Plagiarism of any sort is not acceptable and will be severely punished.
- Late submission will not be accepted.
- **Submit your report, and code as one compressed file (.zip, .rar) on Google Classroom. The name of file should be your roll number i.e. <Roll#>.zip**
- Deadline to submit this assignment is: **Friday, June 5, 2020, 11:59 pm.**

Problem:

The purpose of this assignment is to get you familiar with text classification and more specifically spam email detection. By the end of this assignment you will have your very own "Spam Detector". You are given with Enron Email Dataset that contains separate labelled train and test set. Your task is to train a Naïve Bayes classifier on train set and report several evaluation metrics on test set. The directory structure of dataset is as follow:

train /

    spam

    ham

test /

    spam

    ham

Ham is an alternate name for not-spam.

Implement Naïve Bayes keeping in view all the assumptions discussed in class. Refer to the lecture on Text Classification. You are required to fill a confusion matrix (like the one below) with values obtained by running your Naïve Bayes classifier on test set. Also report Precision, Recall, Accuracy and F1 score.

		gold labels	
		spam	ham
system/ classifier output	spam	t_p	f_p
	ham	f_n	t_n