Question 2:

The dataset we will be using is a subset of 2005 TREC Public Spam Corpus. It contains a `training set` and a `test set`. Both files use the same format: each line represents the space-delimited properties of an email, with the first one being the email ID, the second one being whether it is a spam or ham (non-spam), and the rest are words and their occurrence numbers in this email. In preprocessing, non-word characters have been removed, and features selected similar to what Mehran Sahami did in his original paper using Naive Bayes to classify spams.

Dataset

The data set can be downloaded from here.

Your Task:

Code

Implement the Naive Bayes algorithm to classify spam.

Report

Use your algorithm to learn from the training set and report accuracy on the test set.