3.

Report any decisions made, such as how you decided to handle punctuation. It is fine to do only the minimal preprocessing needed. If you use a package, describe

We have decided to use NLTKpackage for this test classification task. We used word_tokinize to tokenize, wordNetLemmatizer to lemmatize and isalpha to eliminate punctuation, stopwords method to eliminate stop words from the corpus.

Report the accuracy on the test set.

Accuracy on the training set = 0.9980676328502416

Accuracy of the test set = 0.9874396135265701

How do you account for different prior probabilities for spam and ham?

We use Naive Bayes classifier. This classifier tries to choose the most probable class or label among the classes spam and ham. The algorithm looks at the prior probabilities of each word and chooses the most probable class.

$$c = \mathit{argmax}_{c \in \{\text{spam, ham}\}}\ P(c \mid \text{words})$$

Probabilities P(spam) and P(ham) show the distribution of spam and ham classes in the training set.

The algorithm then evaluate:

$$P(\text{spam}) \text{ x } \prod_{\text{wordnet} \in \text{text}} P(\text{word} \mid \text{spam}) > P(\text{ham}) \text{ x } \prod_{\text{wordnet} \in \text{text}} P(\text{word} \mid \text{ham})$$

If the statement above is true then classify email as spam, otherwise classify as ham.

What are the most discriminative words based on the learned probabilities?

Most Informative Features

```
beck = 1            ham : spam   =   309.2 : 1.0
sally = 1           ham : spam   =   158.0 : 1.0
meeting = 2         ham : spam   =    99.6 : 1.0
causey = 1          ham : spam   =    98.2 : 1.0
cc = 1              ham : spam   =    82.6 : 1.0
kevin = 1           ham : spam   =    69.0 : 1.0
creative = 1        spam : ham   =    59.2 : 1.0
```

How does the performance of the classifier change when Laplace smoothing is added?

The accuracy has slightly reduced to:

Accuracy on the training set = 0.9920289855072464

Accuracy of the test set = 0.9826086956521739