**Q1 (DUPLICATED TASK, do not need to do it):** Find the probability of message S1 being spam and not spam using Laplace smoothing:

|  |  |
| --- | --- |
| **Training set** | **Input message** |
| **Spam messages:**  M1: *Buy bicycles for free*  M2: *Bicycles and motorbikes for free*  M3: *Motorbikes rides easy and free*  **Normal messages:**  M4: *Let's go ride bicycles*  M5: *Last week I bought motorbikes and they are cool*  M6: *Some messages about bicycles and motorbikes, that are free, are spam messages* | *Cool bicycles and motorbikes* |

Use Naive Bayes to find the probability of message Message being spam.

**Q2:** Calculate Precision, recall, and F1 for example below

|  |  |  |  |
| --- | --- | --- | --- |
|  | Sports | Politics | Culture |
| Classified as sport | 200 | 20 | 40 |
| Classified as politics | 10 | 150 | 30 |
| Classified as culture | 20 | 40 | 100 |

**Q3:**

Use your own implementation

Implement News Classifier. Use database BBC from <http://mlg.ucd.ie/datasets/bbc.html>

Get some random news and try to classify it by using Naive Bayes and Laplace smoothing.

This task is the same as task #2.

You can remove some of the words that are called stop words.

from nltk.corpus import stopwords

stop = set(stopwords.words('english'))

sentence = "this is a foo bar sentence"

print [i for i in sentence.lower().split() if i not in stop]