T1:

|  |  |
| --- | --- |
| **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* |

Perform logistic regression for input message, each word is a feature:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | f(buy) | f(bicycles) | f(free) | f(ride) | f(some) | f(messages) | f(motorbikes) | f(spam) |
| **spam** | 0.2 | 0.5 | 1 | 1 |  | 0.4 | 0.3 | 1 |
| **normal** | -0.8 |  | -1 | 2 |  | 1 | 0.2 | 1 |

T2: Install NLTK, Perform tutorial

Create list of 20 kazakh names, and find how much of them was found correctly

(<http://textminingonline.com/dive-into-nltk-part-vii-a-preliminary-study-on-text-classification> )

T3: Implement Naive Bayes and MaxEnt classifier for this task:

Try to predict genre, and origin of movie by **plot**

<https://www.kaggle.com/jrobischon/wikipedia-movie-plots>

Compare by precision, recall and accuracy