Assignment 1

September 27, 2016 (Deadline: October 18, 2016)

IEMS Assignment 1: Wikipedia topic classification

1 Problem

Your task is to classify the topic of a sentence extracted from Wikipedia using a **Naïve Bayes Classifier** trained from the training data. There are 3 topics in total: Biology, Computer and Music. Sentences with Biology topic should be marked "B"; with Computer topic should be marked "C" and Music topic "M". The accuracy of your classifier (classifying the training data) should be given. The classifier is used to classify the testing data and accuracy should be given as well. Any words in the sentence can be used as the classifying feature. The classification result should be stored in the file **result.txt** which is acutally the **test.txt** file with the classification result appended to the end of each sentence, for example:

`result.txt`

There is strong evidence from genetics that all organisms have a common ancestor. B Natural bacterial transformation is considered to be a primitive sexual process and occurs in both bacteria and archaea. B An organism is any contiguous living system. C At present there is no theoretical model for how adaptation occurs that is close to being complete. C ⋮ Melodies also often contain notes from the chords used in the song. M

Here the first two sentences are correctly classified as "Biology". The third and fourth sentences are wrongly classified as "Computer" by the classifier. The last sentence is correctly classified as "Music". Please notice that there is a *space* between the sentence and the label.

2 Input

We have 4 files, including **biology.txt**, **computer.txt**, **music.txt** and **test.txt**. The first three files store their respective training data in the specific topic. Each of them include 45 training sentences. No other input files or corpora are allowed. The **test.txt** file stores 15 testing sentences. The first 5 sentences are with topic Biology. Next 5 with topic Computer, and the last 5 with the topic Music.

3 Submission

A runnable Python program file should be submitted (assignment1.py). NLTK is the only 3rd party library allowed in this assignment and it is optional, but you are encouraged to use it. The output txt (result.txt) should be submitted as well.

4 Output

You need to print 2 results: the accuracy of your classifier using training data and testing data respectively. After runing the python program, the screen should display: (actual number varies)

```
>>>classifier's accuracy on training data:  0.854
>>>classifier's accuracy on testing data:  0.662
```

5 Grading Scheme

50%	Complete runnable Python program
20%	Correct format of **result.txt** and program output
30%	Accuracy of the classification

$$Accuracy = \frac{no.\ of\ correctly\ classified\ sentences}{total\ no.\ of\ test\ sentences\ (=15)}$$

You should try several different classifying features as the classification performance is counted towards the final 30% of marks. The scoring of the accuracy part is relative. The submission with the best performance gets full mark. Others get a relative proportion of the mark.