National University of Computer and Emerging Sciences, Lahore Campus

	Course Name:	Artificial Intelligence	Course Code:	CS 401
	Program:	BS(CS)	Semester:	Spring 2018
	Duration:	20 Minutes	Total Points:	10
	Paper Date:	Monday, May 7 2018	Weight	3%
	Section:	D,E	Page(s):	2
	Exam Type:	Quiz 2		

Student : Name:_____ Roll No._____ Section:_____

Problem [Naive Bayes]   [4 + 6 Points]

To classify an email as SPAM or HAM It has been decided to use the Naive Bayes along with unigram features (presence of a word in the email). Further, it has been decided that only the words with length greater than 3 will be used for classification

Hence, if an email with **K** key words **w$_1$, w$_2$, ... ,w$_K$** having length greater than 3 then its label is computed using

$$\text{class}= \arg\max_{c \in \mathbb{C}} P(c|d) = \arg\max_{c \in \mathbb{C}} P(c) \prod_{i=1}^{K} P(w_i | c).$$

where **arg max** means that the class with maximum probability will be the predicted class.

The probabilities **P(wi |c)** have already been estimated and given in following table.

W	P(W|SPAM)	P(W|HAM)
note	1/6	1/8
self	1/4	1/4
become	1/4	1/12
perfect	1/8	1/12

Part a) For each of the following cases, compute the label of the email given below

                    EMAIL TEXT : note to become perfect

if P(Y = SPAM) is 0.3

if P(Y = HAM) is 0.4

Given the following five emails as a training set:

Training Set

(SPAM) dear sir, I write to you in hope of recovering my gold watch. sir please return my watch.

(SPAM) dear customer, please retry.

(HAM) hey, lunch at 12?

(HAM) fine, watch it tomorrow night

(HAM) dear baig, I am planning to join FAST-LAHORE.

Part b) Compute the estimates of following parameters needed by Naive Bayes. Use Laplace smoothing with K=2 in your estimates. Ignoring all words of length less than 3, and ignoring all non-alphabetical characters.

i) P(W = sir | Y = HAM)

ii) P(W = dear | Y = SPAM)

iii) P(Y = HAM)