National University of Computer & Emerging Sciences
FAST-Karachi Campus
Information Retrieval (CS317)
Class Activity # 2

Dated: April 29, 2020 Submission: May02 ,2020 6PM
Time: 45 min. + Chapter Reading 13 and 14
Std-ID: ___________________ Marks: 30

Problem No. 1

Consider the following examples for the task of text classification

Dataset	**DocID**	**Features- Words in documents**	**Class Fruit=Yes/No**
Training set	1	Orange, Orange, Lemon, Red	No
	2	Orange, Red, Blue, Yellow	No
	3	Apricot, Apple, Mango	Yes
	4	Apple, Banana , Orange	Yes
	5	Blue, Orange, Yellow	No
Test set	6	Orange, Mango, Melon	?
	7	Orange, Red, Lemon, Yellow	?

a. Using the training data calculate the class prior probabilities?
b. Using Multinomial Naïve Bayes to estimate the probabilities of each term (feature) that that are given in the problem.
c. Predict the class labels for the two instances in test set?

Problem No. 2

a. What is concept drift? Which version of Naïve Bayesian is robust against it? Justify your answer.
b. Why Naïve Bayesian Classifiers (NB) called Naïve? Consider the statement "Correct estimation implies accurate prediction, but accurate prediction does not imply correct estimation." Why it is true for NB?
c. Suggest one strategy for NB Classifiers to handle out of training set vocabulary for test instances? How the NB guarantee that it will do it best attempt for classification?
d. Consider the following set of overly simplified documents:
   D1: w1 w3 w1 w4
   D2: w1 w4 w1 w3
   D3: w4 w1 w3
   Which of the documents in above table have identical and different bag of words representations for (i) the Bernoulli model (ii) the multinomial model? If there are differences, describe them.

Problem No. 3

Consider the following examples for the task of text classification

Dataset	**DocID**	**Features- Words in documents**	**Class Fruit=Yes/No**
Training set	1	Orange, Orange, Lemon, Red	No
	2	Orange, Red, Blue, Yellow	No
	3	Apricot, Apple, Mango	Yes
	4	Apple, Banana , Orange	Yes
	5	Blue, Orange, Yellow	No
Test set	6	Orange, Mango, Melon	?
	7	Orange, Red, Lemon, Yellow	?

a. Shows the tf-idf vector representations of the five documents given in this problem, using the formula term frequency = $(1+\log_2 tf_{t,d})$ and Inverse document frequency as $\log_2(5/df_t)$.

b. Show all the computation for computing centroids for each class. $\mu_{Fruit=Yes}$ and $\mu_{Fruit=No}$

c. Classify the documents 6 and 7 using Rocchio’s approach.

d. Use the vectors computed in part (a) to classify the documents 6 and 7 using KNN approach using k=3