Pattern Recognition –HW#6

About the Assignment

The main aim of the assignment is to learn text classification and statistical feature extraction. Contributions of this lab are;

- Ability to analyze the statistical features.
- Ability to analyze the training a model on text data.
- Understanding idea of text representation with TF*IDF.
- Understanding idea of text representation with Information Gain (IG).

Step1:

TF: Term Frequencies and **IDF:** Inverse Document Frequencies

Data	Document	Words	Class
Training Data	d1	free, free, free, buy, discount, combo, pleasure,	S
	d2	free, free, free, discount, pleasure, smile, smile, smile	S
	d3	cat, mouse	N
	d4	cat, cat, dog, dog, dog, dog	N
	d5	mouse	N
Test Data	d6	dog, cat, mouse, cat	???
	d7	Free, free, smile	???

You are given the two classes, S and N, with related training data.

Step1. Choose the best two features (words) by using Mutual Information (MI) as shown in Eq. (3). **Hint.** You will compute the MI score of each word. Sort the MI in descending order and select the highest two features (words).

In probability theory, the MI score of two random variables is a quantity that measures the mutual dependency of them. The MI formula is given in Eq. (2-3) [3].

$$\mathrm{MI}(W) = \sum_{(w\in\{0,1\},c\in\{S,N\})} P(W=w, C=c)\, x \log \frac{P(W=w, C=c)}{P(W=w,)P(C=c)} \quad (2)$$

$$\begin{aligned} MI\ (W) = {} & P(W=0, C=S)\ x\ log_2(P(W=0,\ C=S)/(P(W=0)\ x\ P(C=S))) + \\ & P(W=1, C=S)\ x\ log_2(P(W=1,\ C=S)/(P(W=1)\ x\ P(C=S))) + \\ & P(W=0, C=N)\ x\ log_2(P(W=0,\ C=N)/(P(W=0)\ x\ P(C=N))) + \\ & P(W=1, C=N)\ x\ log_2(P(W=1,\ C=N)/(P(W=1)\ x\ P(C=N))). \end{aligned} \quad (3)$$

In (2) and (3),

S and N refer to the spam and normal emails, respectively.

P (W=0, C=S): the probability of the word not to be included in S.

P (W=1, C=S): the probability of the word included in S.

P (W=0, C=N): the probability of the word not to be included in N.

P (W=1, C=N): the probability of the word included in N.

Step2. You are expected to compute the TF*IDF score of selected two features (words).

Step3. Represent the each document with these selected two features, called TF*IDF values of two selected features (words). After Step3, A matrix (5x2) will be formed.

Step4. Calculate the TF*IDF values of selected two features (words) for d6. A vector (1x2) will be formed.

Step5. Calculate the TF*IDF values of selected two features (words) for d7. A vector (1x2) will be formed.

Step6. Predict the class label of d6 by using the KNN algorithm..

Step7. Predict the class label of d7 by using the KNN algorithm..

Submit the Assignment

Ex: No_Name_Surname_HW#.zip

Hint

You can look at the implementations available in internet.