Semester - 6th
Subject Name:- NLP& Code:- IT3035
Branch (s): - CSSE

AUTUMN MID SEMESTER EXAMINATION-2023

School of Computer Engineering
Kalinga Institute of Industrial Technology, Deemed to be University
Natural Language Processing
[IT3035]

Time: 1 1/2 Hours **Full Mark: 40**

Answer Any four Questions including Question No. 1 which is compulsory.
The figures in the margin indicate full marks. Candidates are required to give their answers in their own words as far as practicable and all parts of a question should be answered at one place only.

1. Answer all the questions. [2 x 5]

a) Define a random variable?

b) Write the phrase structure of the following sentences using Brown tags.
   a. The students looked at the whiteboard.
   b. My aunt's can-opener can open a drum.

c) Mutual information is actually just a measure of how far a joint distribution is from independence. True/False. Justify your answer.

d) What is the difference in the distribution of letters in P w.r.t Q given in the following table:

P:	tit for tat					
P(x)	P(t)=0.2	P(i)=0.3	P(f)=0.4	P(o)=0.4	P(r)=0.4	P(a)=0.4
Q:	sweet potato is good for health					
Q(x)	Q(t)=0.3	Q(i)=0.4	Q(f)=0.4	Q(o)=0.1	Q(r)=0.12	Q(a)=0.3

P(x) and Q(x) are the probabilities of different letters in the sentence P and Q respectively.

e) Define inflection and cliticization with suitable examples.

2. Google found that 20% of mails are spam. GMAIL filters spam mail before reaching the inbox. It's accuracy for detecting a spam mail is 98% and chances of tagging a non-spam mail as spam mail is 5%. If a certain mail is tagged as spam find the probability that it is not a spam mail.

[10 Marks]

3. Find the the minimum edit distance and operations required to edit the word :

“PIRFECT” to “ PERFACT “.

[10 Marks]

4. a. The joint entropy of a pair of discrete random variables X, Y is given as:

$$H\,(X, Y) =-\sum_{x\in X}\sum_{y\in Y} P(x, y)\log P(x, y)$$

and the conditional entropy is given as:

$$H\,(Y \mid X) =-\sum_{x\in X}\sum_{y\in Y} P(x, y)\log P(y|x)$$

Prove that: $H\,(X,Y) = H(X) + H(\,Y \mid X)$

b. Suppose a tribal language has 5 alphabets. The letters along with their frequencies are given below in the table:

[10 Marks]

Letters	Џ	Б	Ж	Њ	Ч
Frequency	$\frac{1}{8}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{6}$

Find out the average number of bits they will require to send a letter?

5. Write short notes:

a) POS tagging

b) Noisy Channel Model

c) KL divergence [3+3+4]

*** Best of Luck ***