Semester: 6th
Subject :- NLP (IT-3035)
Branch (s): - All

SPRING MID SEMESTER EXAMINATION-2024

School of Computer Engineering
Kalinga Institute of Industrial Technology, Deemed to be University
Subject: Natural Language Processing
Subject Code: IT 3035

Time: 1 1/2 Hours **Full Mark: 20**

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

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

a) Lemmatize the following words: "running", "mice", "going", "happiness".

b) For the below table, given a training corpus with total number of words is 1000. Determine the correction word for the mistyped word "acress" based on the Noisy Channel Model.

Correction	Correct Letter	Error Letter	x|w	P(x|w)	Word Frequency in Corpus
actress	t	-	c|ct	0.00016	80
across	o	e	e|o	0.0000091	40

c) What is the Markov Assumption? Explain how it is useful in NLP?
d) What is Entropy? Write the corresponding equation form.
e) Explain the differences between *inflextional* and *derivational* morphology.

2. Consider the following corpus consisting of four sentences: [5 Marks]

<s> The cat sat on the mat. </s>
<s> The dog barked at the cat. </s>
<s> The cat chased the mouse. </s>
<s> The dog slept on the mat. </s>

Calculate the Probability of the sentence S: The cat slept on the mat. assuming a bigram language model. <s> represents the start of a sentence, and </s> represents the end of a sentence. Show your calculations and explain the steps you take to compute the probability of the given sentence.

3. Answer the following questions: [2.5 Marks x 2]

(a) Define TF-IDF. Write the formulas to calculate TF and IDF individually. Compute the TF-IDF score after Data Pre-processing for the following corpus having 3 documents.

D1: Inflation has increased unemployment.
D2: The company has increased its sales
D3: Fear increased his pulse

(b) Calculate the Levenshtein distance between the words "kitten" and "sitting". Assume the cost of insertions, deletions, and substitutions 1. Create the distance matrix to calculate the minimum distance.

Semester: 6th
Subject :- NLP (IT-3035)
Branch (s): - All

4. A team of researchers presented their finding to a panel of judges to receive funding. The comments of the panel and their associated sentiments is given below. The head of the panel gives the comment as "**Poor talk but good research**". Apply Naïve Bayes method to classify the given sentence into "Positive" or "Negative" class. Show all required data processing steps and Probability calculations. For smoothing, the smoothing parameter value = 1.

[5 Marks]

Text	**sentiment**
I liked the talk	Positive
It is a good talk. Excellent Outline	Positive
Nice research but poor application	Negative
Approach is bad but the team has worked hard	Positive
Sorry, no good results	Negative

5. What is Natural Language Processing and what are its applications? Explain at least four challenges in NLP with suitable examples. [5 Marks]

*** Best of Luck ***