Ref No.: Ex/PG/MTCT/T/221B/2024

M. Tech. in Computer Technology, 2nd Year 2nd Semester Examination, 2024

Natural Language Processing

Time – 3 hours **Full Marks - 100**

Answer any five questions

1.

a. Find out the edit distance and alignment between the two strings "*pressure*" and "*supersede*", considering the following costs for the edit operations. *10*

```
insertion cost = deletion cost = 1
substitution cost = (your exam roll number % 2) + 1
```

b. Write a regular expression to identify all instances of the English word "`the`". *2*

c. Write a shell script to normalize case, tokenize and show the tokens ending with "`ing`" that could potentially be verbs in a corpus in decreasing order of frequency. Explain your answer. *6*

d. What is case normalization? Describe some cases where case normalization is desirable and where it is not. *2*

2.

a. Define surface form, lemma, morpheme, stem and affix with suitable examples. *4*

b. Derive the trigram language model using chain rule, Markov assumption, maximum likelihood estimation and add-1 smoothing. *6*

c. Define and deduce perplexity. Discuss the notion of perplexity as a branching factor. *4*

d. Discuss the Kneser-Ney smoothing technique. *6*

3.

a. Given the following training documents, compute which class the test documents belong to. *10*

	Doc_ID	Words	Class
Training	1	`wicket wicket run pitch`	`C (Cricket)`
	2	`wicket run bat ball`	`C`
	3	`score boundary ground`	`C`
	4	`score goal foul penalty`	`F (Football)`
	5	`referee freekick score pitch`	`F`
Test	6	`score ball goal penalty`	`?`
	7	`wicket score run pitch`	`?`

[Turn over

b. Discuss some positive and negative aspects of the Naïve Bayes Classifier. *4*

c. What is a confusion matrix in the context of spelling correction? Describe the four confusion matrices and how they are used in estimating the likelihood probability in the Noisy Channel model for spelling correction. *6*

4.

a. Define homonym, homograph, homophone with examples. *3*

b. Compare multivalue classification and multinomial classification. *2*

c. Given the following confusion matrix, compute the class-wise precision and recall. Also compute macroaveraged and microaveraged precision values. *5*

Actual / System	A	B	C
A	100	20	80
B	60	200	40
C	50	50	400

d. Compute the alignment probabilities and the translation probabilities according to the `EM algorithm` assuming `no NULL token` and `only 1-to-1 alignments` for the following toy parallel training corpus. Show the first 3 iterations. *10*

Translation pair id	Source Language	Target Language
1	`big house`	`casa grande`
2	`the house`	`la casa`

5.

a. What are the main disadvantages of Boolean information retrieval? *3*

b. Discuss the inverted index construction process. *5*

c. Discuss how phrase queries are handled in Information Retrieval. *4*

d. Compute the score assigned to the following query-document pair by the tf-idf model using the lnc.ltc weighing scheme. Assume that the document frequencies of the terms “sensor”, “best”, “electric”, “bike”, “battery” and “motor” are 5,000, 50,000, 10,000, 1,000, 25,000 and 40,000 respectively, and the document collection size is 1,000,000. *8*

Document: `bike electric bike sensor bike battery motor`

Query: `best electric bike`

6.

a. Define Positive Pointwise Mutual Information (PPMI). What does it measure? *2*

b. Define hyponym and hypernym. Discuss the properties of hyponymy. *3*

c. Discuss the Resnik method and Lin method of measuring semantic similarity between two words in terms of information content. *5*

d. Given the following term-context matrix, compute which of the following word pairs - [`data, information`] and [`lemon, orange`], is more similar according to distributional similarity using `add-`*`n`* `smoothing`, where *`n`* `= (your exam roll number % 2)+1`. *10*

context / term	computer	digital	pinch	sugar	program
data	2	2	0	0	1
information	1	6	0	0	4
lemon	0	0	1	1	0
orange	0	0	1	2	0

7. Write short notes on any four: *4*5*

a. Vector space model for IR.

b. Forward algorithm in HMM.

c. Viterbi algorithm.

d. BLEU MT evaluation metric and its performance issues.

e. Good-Turing smoothing.

f. tf-idf model for ranked information retrieval.