Ref No.: Ex/PG/CSE/T/129B/2024

M.E. Computer Science & Engineering, 1st Year, 2nd Semester Examination, 2024

Natural Language Processing

Time – 3 hours **Full Marks - 100**

<u>Answer any five questions</u>

1. a. Define the following terms with suitable examples: type, token, vocabulary, lemma and morpheme. *3*

   b. Discuss the difference between the Needleman-Wunsch algorithm and the Levenshtein Edit Distance algorithm. *4*

   c. Write some sophisticated features that you can use to build a classifier to decide whether a '.' (dot) in English text marks the end of a sentence or not. *3*

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

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

2. a. Define and deduce perplexity. Discuss the notion of perplexity as a branching factor. *3+2*

   b. Derive the trigram language model using maximum likelihood estimation, chain rule, Markov assumption and add-1 smoothing *5*

   c. What is MLE? Add-1 smoothing is a non MLE estimator. Explain this. *2+2*

   d. Discuss some efficiency related practical issues to deal with web-scale language models. *4*

   e. What is continuation probability of a word? How it is computed? *2*

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

   b. Explain the inverted index data structure and how it is constructed. *5*

   c. Discuss how phrase queries are handled in Information Retrieval. *5*

   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 '`digital`', '`best`', '`DSLR`', '`camera`', '`lense`' and '`zoom`' are `5,000`, `50,000`, `10,000`, `1,000`, `25,000` and `40,000` respectively, and the document collection size is `1,000,000`. *7*

```
Document:   camera DSLR camera digital camera DSLR lense zoom
Query:      best DSLR camera
```

4. a. Define homonym, homograph and homophone, with suitable examples. *2*

   b. Discuss the properties of hyponymy. *2*

   c. WordNet is much more than a thesaurus. *2*

   d. Discuss the main difference between Resnik similarity and Lin Similarity. *2*

   e. In semantic similarity, path-based methods have problems with recall, while distributional models have problems with precision. Explain this. *2*

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

context term	computer	digital	boil	result	fry
data	2	2	0	1	0
information	1	6	0	2	0
eggplant	0	0	1	0	1
potato	0	0	1	0	2

5. a. Naïve bayes classifier has an important similarity to language modeling. Explain this. *3*

b. Discuss some positive and negative aspects of the Naïve Bayes Classifier with regard to performance issues. *3*

c. How real word errors can be detected and corrected? *4*

d. Given the following training documents and their membership to the two class – Cricket (C) and Football (F), compute which class the test document belongs to. Consider add-1 smoothing. *6*

	Doc_ID	Words	Class
Training	1	wicket wicket run pitch	C (C)
	2	wicket run run bat ball	C
	3	score boundary ground	C
	4	score goal goal penalty	F
Test	5	score ball goal penalty	?

e. Define mean average precision (MAP). Compute Average Precision (AP_{10}) for the following search results. *2+2*

Rank	1	2	3	4	5	6	7	8	9	10
Relevant	Y	Y	N	Y	Y	N	N	N	N	Y

6. a. Briefly discuss about the BLEU MT evaluation metric and its performance issues. *5*

b. Write the Viterbi algorithm for finding the optimal sequence of tags for an observation sequence, given the model. Explain every step of the algorithm. *5*

c. 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	`blue shirt`	`camisa Azul`
2	`the shirt`	`la camisa`