Subject Code: Ex/CSE/PE/B/T/414C/2024

B.C.S.E. 4th Year 1st Semester Examination - 2024

Natural Language Processing

Time: 3 Hours

Full Marks: 100

CO1: Mathematical Models

Answer any 3 questions *3*10=30*

1. *10*

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

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

2. *10*

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.

context term	computer	digital	ripe	result	juicy
data	2	2	0	1	0
information	1	6	0	4	0
lemon	0	0	1	0	1
orange	0	0	1	0	2

3. *(7+3)=10*

a. 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 *n* * 1,000,000, where *n* = (your exam roll number % 2)+1.

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

b. If P(0)=0.19 and P(any other digit)=0.09, compute the perplexity of your 12 digit class roll number according to the unigram language model.

4. *10*

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.

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

[Turn over

CO2: Algorithms

Answer any 3 questions *3*10=30*

1. *(4+4+2)=10*

a. Discuss the Smith-Waterman algorithm for best local alignment between two strings.

b. Describe the four confusion matrices and how they are used in estimating the likelihood probability in the Noisy Channel model of spelling correction.

c. Explain the intuition behind Good-Turing smoothing.

2. *(7+3)=10*

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

b. Discuss about the performance issues of the Naïve Bayes Classifier.

3. *(7+3)=10*

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

b. Discuss how hypothesis recombination can be used to reduce the search space in SMT decoding.

4. *(4+6)=10*

a. Discuss how phrase queries are handled in information retrieval.

b. Discuss the vector space model for ranked information retrieval.

CO3: Linguistics

1. Answer any 5 questions: *2*5=10*

a. Why are stop-words not considered in Information Retrieval?

b. NLP is more difficult for Indian languages than for English. Justify this.

c. Define homonym, homograph and homophone, with suitable examples.

d. Discuss the properties of hyponymy.

e. “Perfect synonymy is rare”. Explain this.

f. Discuss how the probability of a concept (or sense) can be measured.

g. Mention some cases where case folding is useful and where it is not.

CO4: Generalization and Analysis

Answer any 3 questions *3*10=30*

1. *(5+3+2)=10*

a. 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.

b. Compare Needleman-Wunsch algorithm and the Levenshtein Edit Distance algorithm.

c. What are the best-case and worst-case time complexities of the Backtrace algorithm? Mention the cases where they occur.

2. *(8+2)=10*

   a. Discuss and formulate the trigram HMM model for POS tagging.

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

3. *(5+3+2)=10*

   a. Discuss the formulation of BLEU MT evaluation metric. Mention some performance issues of the BLEU MT evaluation metric.

   b. Explain why Lin similarity provides better semantic similarity than Resnik similarity.

   c. What is a term-context matrix and how it is computed?

4. *(4+4+2)=10*

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

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

   b. Discuss some efficiency related practical issues to deal with web-scale language models.

   c. Compare multivalue classification and multinomial classification.