34

ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)

ORGANISATION OF ISLAMIC COOPERATION (OIC)

Department of Computer Science and Engineering (CSE)

MID SEMESTER EXAMINATION
DURATION: 1 HOUR 30 MINUTES

SUMMER SEMESTER, 2022-2023
FULL MARKS: 75

SWE 4841: Natural Language Processing

Programmable calculators are not allowed. Do not write anything on the question paper.
Answer **all 3 (three)** questions. Figures in the right margin indicate full marks of questions with corresponding COs and POs in parentheses.

---

1. a) Zipf's law, a statistical principle observed in natural language processing (NLP), states that the frequency of a word in an English corpus is inversely proportional to its rank. Consider a corpus of text where the most frequent word ("the") occurs 10,000 times. — 2 + 4 (CO1) (PO1)

   i. If Zipf's law holds true for this corpus, approximately how many times would the third most frequent word occur?

   ii. How would you design the experiment if you intend to investigate if Zipf's law can be applied in languages other than English?

   b) What is Regular Expression? Write regular expressions for the following language: — 7 (CO1) (PO1)

   i. the set of all lower case alphabetic strings ending in a ***b***

   ii. the set of all strings from the alphabet *a, b* such that each ***a*** is immediately preceded by one ***b*** and immediately followed by another ***b***.

   iii. Use positive lookahead to find all the prices of various products from a document. All price tags start with a ***$*** sign.

   c) How can measuring the edit distance help in the task of coreference? Consider the following two sentences: — 3 + 9 (CO1) (PO1)

   i. The sun rises in the east every morning

   ii. Every morning, the east witnesses the rising of the sun.

   Calculate the cosine similarity and Jaccard similarity between these two sentences. Based on these similarity metrics, which metric do you think is qualitatively better in terms of capturing the semantic similarity of documents?

2. a) What are n-gram language models? How do n-gram language models learn to assign probabilities? Explain the trade-offs for using longer and shorter n-grams in predicting the next word. — 6 (CO1) (PO1)

   b) Two n-gram language models have perplexity of 0.3 and 5.33 respectively. Which language model is qualitatively better? Justify your answer. — 5 (CO1) (PO1)

   c) What is smoothing/discounting? Discuss how the Kneser-Ney smoothing technique outperforms alternative smoothing techniques. — 5 (CO2) (PO1)

   d) Answer the following questions: — 9 (CO1) (PO1)

   i. Why is accuracy not reported in text classification tasks in contemporary research? What are the better alternatives?

   ii. Why is Naive Bayes considered a generative text classifier and logistic regression a discriminative classifier?

iii. Why is the gradient descent algorithm guaranteed to find the optimal weights and biases? Justify your answer.

3. a) Discuss the differences between the following concepts using appropriate examples: 6 (CO1) (PO1)

i. word similarity and word relatedness

ii. semantic field and semantic frame

iii. valence and arousal

b) Consider the following word-word co-occurrence matrix from a large corpus and answer the following questions: 12 (CO2) (PO1)

Table 1: Data for Question 3.b)

	education	meals	institution	shop
phd	119	2	84	7
fruitcake	3	223	12	134
research	92	0	194	5
price	19	294	1	483

i. What is the dimension used for each word vector in this representation? What can be an alternative dimension? Explain the distributional hypothesis behind these two representations.

ii. Determine which focus words are contextually similar from the above matrix.

iii. Which of the embedding can you apply on the above matrix: tf-idf or PMI? How do these embeddings eliminate the importance of ubiquitous words?

c) "Training a word2vec model is a lot like training logistic regression" - do you agree with this statement? Justify your answer. Why is negative sampling applied during word2vec training? How are the negative samples chosen? 7 (CO2) (PO1)