Ref No.: Ex/PG/MTCT/T/221B/2018

M. Tech. in Computer Technology, 2nd Year 2ndSemester Examination, 2018

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 “calculator” and “computer” considering an equal cost for all the edit operations. *8*

b. What is the time and space complexity of the Edit Distance algorithm and the Backtrace algorithm. What are the best-case and worst-case time complexities of the Backtrace algorithm? Why weighted edit distance is useful? *2+2+2*

c. Given the following similarity matrix, find the alignment between the two small DNA sequences CATGC and GATTCA. *6*

	A	G	C	T
A	1	-1	-1	-1
G	-1	2	-1	-1
C	-1	-1	3	-1
T	-1	-1	-1	4

2.

a. Derive the trigram language model using maximum likelihood estimation, chain rule and Markov assumption. *5*

b. Discuss how interpolation can be used in Language models. Explain with an interpolated trigram model. *4*

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

d. Define surface form, lemma, morpheme, stem and affix with suitable examples. *5*

3.

a. How candidates are generated for non-word spelling errors? *2*

b. What is a confusion matrix in the context of spelling correction? Describe the four confusion matrices and how these are used in estimating the likelihood probability in the Noisy Channel model for spelling correction. *8*

c. What are real-word errors? How real word errors can be detected and corrected? *1+3*

d. What is the simplification assumption that is often made to reduce the search space in dealing with real word spelling errors and how much it is able to reduce the search space? *1+2*

e. Mention some advanced features that are used in state-of-the-art spelling correction models. *3*

4.

a. Define and discuss precision, recall and F-measure in the context of text classification. Discuss how multi-way classifiers can be evaluated. *3+2*

b. Given the following training data, compute which class the test document belongs to. *7*

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

c. Discuss some practical issues in text classification. *4*

d. Distinguish between macroaveraging and microaveraging. Given the following confusion matrices for two classes, compute macroaveraged and microaveraged precision. *2+2*

Class 1

Actual System	yes	no
yes	10	10
no	10	970

Class 2

Actual System	yes	no
yes	80	20
no	10	890

5.

a. Explain the inverted index data structure. Why it is called 'inverted' index? How queries are processed with an inverted index. 3+1+3

b. Discuss the difference(s) among term-document incidence matrix, term-document count matrix and tf-idf matrix. *3*

c. What does the "lnc.ltc" weighting scheme mean for a search engine? *2*

d. What are the main disadvantages of boolean information retrieval? *2*

e. Given the following term-document count matrix, find out which document pair is the most similar according to the vector space model. *6*

Novel Term	Document A	Document B	Document C
computer	110	50	2
information	15	8	10
medicine	2	0	12
disease	0	1	24

6.

a. Differentiate between word similarity and word relatedness. *2*

b. Define synset, meronym and holonym with examples. *3*

c. What is a term-context matrix and how it is used to measure word similarity? *4*

d. Define Positive Pointwise Mutual Information (PPMI). What does it measure? *2+1*

e. Given the following term-context matrix, compute the distributional word similarity between each term -context word pair using add-2 smoothing. *8*

context term	computer	digital	pinch	result	sugar
Data	2	2	0	1	0
Information	1	6	0	4	0
Lemon	0	0	1	0	1
Orange	0	0	1	0	2

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

a. Resnik method and Lin method of measuring semantic similarity.

b. Good-Turing smoothing.

c. Naïve Bayes classifier for text classification.

d. tf-idf model for ranked information retrieval.

e. WordNet.

f. Handling phrase queries in Information Retrieval.