Ref No.: Ex/CSE/T/423B/2018

B.C.S.E. 4th Year 2nd Semester Examination, 2018

Natural Language Processing

Time – 3 hours **Full Marks - 100**

Answer any five questions

1.

a. Write a shell script to normalize case, tokenize and show the tokens ending with "*ing*" in a corpus in decreasing order of frequency. Explain your answer. *5*

b. Find out the edit distance and alignment between the two strings "*processing*" and "*impression*" considering an equal cost for all the edit operations. *10*

c. Define the following terms with suitable examples - type, token, vocabulary. *3*

d. What is case folding? Describe some situations where case folding is undesirable. *2*

2.

a. Define the base condition and the recurrence relation for defining the minimum edit distance. *4*

b. Discuss the time and space complexity of the Levenshtein Edit Distance algorithm and the Backtrace. What are the best-case and worst-case time complexities of the Backtrace algorithm? *2+2*

c. Why you might require weighted edit distance? *2*

d. Write the Needleman-Wunsch Algorithm. What changes you need to make to it to arrive at its overlap detection variant? *4+1*

e. Define and deduce perplexity. *3*

f. Discuss how language models could be evaluated extrinsically. *2*

3.

a. Discuss the essence of back-off smoothing. *2*

b. Discuss how to deal with web-scale language models? What smoothing technique is used for web-scale language models? *3+2*

c. State and explain Zipf's law. *3*

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

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

4.

a. Discuss the Naïve Bayes classifier for the text classification task. What is "naïve" in Naïve Bayes? *6+1*

b. Discuss how multi-way classifiers can be evaluated. *3*

c. Discuss some positive and negative aspects of the Naïve Bayes classifier with regard to performance issues. *4*

d. Given the following training documents, compute which class the test document belongs to. *6*

{ Turn over

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

5.

a. Discuss the difference(s) among term-document incidence matrix, term-document count matrix and tf-idf matrix. *2*

b. Why document frequency is preferred over collection frequency in ranked IR? Does idf have any effect on ranking for single-term queries? Justify your answer. *2+2*

c. What are the main disadvantages of Boolean information retrieval? *2*

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

e. 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 5000, 100000, 10000, 25000, 20,000 and 7500 respectively, and the document collection size is 1,000,000. *7*

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

Query: *best DSLR camera*

6.

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

b. 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	boil	data	result	fry
Fish	0	2	0	0	2
Potato	0	2	0	0	1
Digital	2	0	3	2	0
information	1	0	6	4	0

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

d. Define synonym, homonym, hyponym and meronym. Discuss the properties of hyponymy. *4+2*