Ref No.: Ex/CSE/T/425D/2017 (Old)

B.C.S.E. 4th Year 2nd Semester Examination, 2017

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 of a corpus in decreasing order of frequency. Explain your answer. *5*

b. Find out the edit distance and alignment between the two strings "*sanskriti*" and "*sensitize*" considering an equal cost for insertion, deletion and substitution. *8*

c. Describe an algorithm for word segmentation that can be used for languages that do not have spaces between words. *2*

d. Build a decision tree for detecting end-of-sentence in English. Discuss some sophisticated features that you can use in machine learning based classifiers for this task. *3+2*

2.

a. Discuss the notion of perplexity as a branching factor. *2*

b. Discuss how the reconstituted counts (c^*) are calculated in Laplace smoothing. *2*

c. Discuss how interpolation can be used in Language models. Explain with an interpolated trigram model. *3*

d. What are the best-case and worst-case time complexities of the Backtrace algorithm? *2*

e. Discuss the Smith-Waterman algorithm for finding the best local alignment between two strings. *5*

f. 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

3.

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

b. 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*

c. Discuss some advanced features that are typically employed for spelling correction in state of the art systems. *3*

d. Derive the trigram language model using maximum likelihood estimation, chain rule and Markov assumption. *6*

e. Define mean reciprocal rank (MRR). Define mean average precision (MAP). Compute MAP_{10} for the following search results. *2+2+2*

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

4.

a. Compare multivalue classification and multinomial classification. *2*

b. Naïve Bayes has an important similarity to language modeling. Explain this. *3*

c. Define and discuss precision and recall in the context of text classification. What is F-measure? Discuss how you could compute class-specific (i.e. per class) precision, recall and accuracy when dealing with multiple classes? *2+2+3*

d. Given the following training documents, compute which class the test document belongs to. *5*

	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	?

e. Briefly discuss about the performance issues of the Naïve Bayes model for text classification. *3*

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. What is a positional index and why this is useful? *2*

c. What is the "bag of words" representation? How it is different from set? What are the limitations of the bag-of-words model? *1+1+1*

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

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 5,000, 50,000, 10,000, 1,000, 25,000 and 40,000 respectively, and the document collection size is 1,000,000. *6*

Document: *camera DSLR camera digital camera lense zoom*

Query: *best DSLR camera*

6.

a. The contexts can be weighted using Pointwise Mutual Information (PMI). Explain, giving formulae, how PMI is calculated and how individual probabilities are estimated from a text corpus. *5*

b. Discuss the Resnik's information content based method for measuring similarity between two words. How Lin similarity is different from Resnik similarity. *4+2*

c. Compare thesaurus based semantic similarity with distributional semantic similarity. *3*

d. Define Pointwise Mutual Information (PMI). What does it measure? Prove: PMI(x,y)=log(P(x|y)/P(x)). *1+1+2*

e. Discuss how syntax could play a role in determining word similarity. *2*

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

a. Kneser-Ney Smoothing.

b. Noisy channel model for non-word spelling correction.

c. Good-Turing smoothing.

d. Vector space model for IR.

e. Handling of phrase queries in IR.
