MASTER OF COMPUTER SCIENCE AND ENGINEERING

First Year, Second Semester Examination, 2019

Text Analytics

Time- Three Hours **Full Marks-100**

Answer Any Five Questions

1. a. What do you mean by answer type taxonomy? Where and how do we use Mean Reciprocal Rank (MRR)? How do you select query words from the following sentence in case of developing a QA system? Show with steps.

"Who can identify the term "LOVE" in his poem "BHALOBASAR NIBEDON"?

b. What are the differences between Natural Language Generation and Natural Language Understanding? What do you know about Jaccard Coefficient?

(3+3+7) + (3+4) = 20

2. a. State Log-Likely hood Ratio (LLR), the unsupervised content selection techniques for text summarization. Compare unsupervised and supervised content selection methods? The following are three reference summaries along with a system generated summary. What are the scores of ROUGE-2 evaluation scheme?

- Human 1: We are the great Indian citizen who can devote for the country.
- Human 2: You are the great Indian citizen who can identify the proper value for the country.
- Human 3: We are really proud to be the great Indian citizen who can devote for the nation.
- *System answer: We are the great Indian citizen who can sacrifice their lives for the country.*

b. How do you use Singular Value Decomposition (SVD) in Latent Semantic Indexing?

(5+2+6)+7=20

3 a. What is relevance feedback query? State and explain Rocchio SMART algorithm for calculating a relevance feedback query using VSM. What is the difference between the original and modified queries?

b. Consider the following two tables which show the results of two classes, A and B. What are the Macro-average and Micro-average Precision values? Is Micro-averaged score dominated by score on common classes?

Class A	Truth: yes	Truth: no
Classifier: yes	20	20
Classifier: no	10	40

Class B	Truth: yes	Truth: no
Classifier: yes	20	20
Classifier: no	20	60

(2+6+2) + (8+2) = 20

4 a. What are the pros and cons of the Vector Space Model (VSM)? What are the differences between document frequency and inverse document frequency (idf)? What are the roles of a tf-idf model for ranked information retrieval?

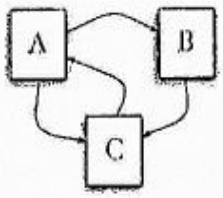
b. Suppose, three documents are taken from class "X" and two documents from class "Y" and employed them into two Naïve Bayes classifiers (Normal and Boolean) as training set along with their sentence level constituents. See the following Table. Calculate the probabilities of the test documents (id6 and id7) to be assigned into a particular class for two classifiers separately. Show each of the steps and compare the results of the classifiers with justification.

Table 1	**Doc**	**Sentences**	**Class**
Training	id1	A B A	X
	id2	A A C	X
	id3	A D	X
	id4	A P Q	Y
	id5	P C Q	Y
Test	id6	A A A P Q	?
	id7	P Q C Q	?

(3+3+4) + (5+5) = 20

5 a. What are the different components required for identifying Sentiments? What do you mean by BOW model for sentiment analysis? How do we calculate PMI?

b. What do you mean by an anchor text? What is the role of proximity matches in case of query phrase identification? Using Page Rank algorithm, in different iterations, identify the ranks of the webpages A, B and C as shown in the following figure.



(2+3+3) + (2+3+7) = 20

6 a. How do we modify Naïve Bayes for underflow prevention? Write down the basic architecture of a modern factoid based Question-Answering (QA) system.

b. Define Kappa measure and state its use. Calculate Kappa for the A and B Classes from the following Table

Class		Agreement Values	
		Yes	No
Class A	Yes	9	11
	No	8	10
Class B	Yes	27	16
	No	14	20

(2+6) + (2+10) = 20