MASTER OF COMPUTER SCIENCE AND ENGINEERING
First Year, Second Semester Examination, 2018
Text Analytics

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

Answer Question No. 1 and Any Three Questions from the rest

1. Answer all the questions 10 X 4=40
   a. Why do we use Cosine measure instead of Euclidian distance in case of calculating vector space proximity? Justify your answer with respect to query and document.
   b. Where and how do we use Mean Reciprocal Rank (MRR)?
   c. What are the differences between Natural Language Generation and Natural Language Understanding?
   d. 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"?

   e. Does idf have any effect on ranking for single-term queries? Justify your answer.
   f. Why accuracy is not a good evaluative measure for search engine?
   g. Define Micro and Macro average precisions.
   h. Why do we use template for training using CRF? Give an example to justify your answer.
   i. How do you relate Singular Value Decomposition (SVD) with Latent Semantic Indexing?
   j. What is the role of proximity matches in case of query feature identification?

2. a. State different unsupervised content selection techniques for text summarization. What are the differences between BLEU and ROUGE metrics? The following are three reference summaries along with a system generated summary. What are the scores of ROUGE-2 and ROUGE-3 evaluation schemes?

- 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. Write down the basic architecture of a modern factoid based Question-Answering (QA) system.

(6+2+5)+7=20

3. a. What are the differences between Information Extraction and Information Retrieval? Write the names of different supervised and unsupervised classification approaches used for Information Extraction?

b. 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?

(3+6) + (3+3+5) = 20

4. 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. 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

(3+5+2)+10=20

5. a. Why we use Bag of Words (BOW) model and write with an example how Naïve Bayes Classifier is employed for text classification task.

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

(3+7)+10=20

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	?

6. a. What do you mean by Subjectivity and/or Sentiment analysis? Write down six challenges of Sentiment Analysis? What are the different components required for identifying Sentiments?

   b. What do you mean by an anchor text? Write down a Page Rank algorithm and explain it with a suitable example. What do you mean by inverted index?

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