# APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

**08 PALAKKAD CLUSTER**

**Q. P. code : 2B-17-1** **(pages:3)** **Name:**

**Reg No:**

**SECOND SEMESTER M.TECH. DEGREE EXAMINATION APRIL/MAY 2017**

**08CS 6022 Information Retrieval**

**Time:3 hours** **Max Marks: 60**

Answer all six questions. Part 'a' of each question is compulsory.

Answer either part 'b' or part 'c' of each question

| Q.no. | Module 1 | Marks |
|---|---|---|
| 1.a | An information retrieval evaluation measure is different from traditional algorithm evaluation.Why? | 3 |
| | **Answer b or c** | |
| b | Briefly explain the information retrieval process with a block diagram | 6 |
| c | Consider a document collection with 5 relevant documents. Find the recall and precision at each document retrieval and plot the Precision/Recall curve(use interpolation). | 6 |

| n | Doc id | Relevant |
|---|---|---|
| 1 | 223 | yes |
| 2 | 432 | no |
| 3 | 123 | yes |
| 4 | 765 | yes |
| 5 | 222 | no |
| 6 | 453 | yes |
| 7 | 876 | no |
| 8 | 220 | yes |

| Q.no. | Module 2 | Marks |
|---|---|---|
| 2.a | What is the difference between stemming and lemmatization | 3 |
| | **Answer b or c** | |

b Consider a very small collection that consists the following three documents with the given contents: 6

d1: “Kerala Technological University results”

d2: “Calicut University results”

d3: “Karnataka University results”

Rank the documents for the query “Kerala results” using vector model. Assume tf-idf ranking scheme and length normalization.

c What is a positional index? Explain the algorithm for proximity intersection of postings lists p1 and p2 . 6

**Q.no.** **Module 3** **Marks**

3.a What is pseudo relevance feedback? What is its disadvantage? 3

**Answer b or c**

b Suppose that a user’s initial query is ‘cheap CDs cheap DVDs extremely cheap CDs’. The user examines two documents, d1 and d2. He judges d1, with the content ‘CDs cheap software cheap CDs’ relevant and d2 with content ‘cheap thrills DVDs’non-relevant. Assume that we are using direct term frequency (with no scaling and no document frequency). There is no need to length-normalize vectors. Using Rocchio relevance feedback what would the revised query vector be after relevancefeedback? Assume $\alpha = 1$, $\beta = 0.75$, $\gamma = 0.25$. 6

c Consider a Web graph with three nodes 1, 2, and 3. The links are as follows: 1->2, 3->2, 2->1, 2->3. Write down the transition probability matrices for the surfer’s walk with teleporting, with the value of teleport probability $\alpha$=0.5. Derive the ranking of the nodes using PageRank algorithm. 6

**Q.no.** **Module 4** **Marks**

4.a What are spatial access methods? 3

**Answer b or c**

b Explain the generic multimedia indexing approach for multimedia IR 6

c Briefly explain the vector space model for information retrieval from XML documents 6

| Q.no. | Module 5 | Mark |
|---|---|---|
| 5.a | Discuss about the overfitting problem in building a classifier | 4 |
| | **Answer b or c** | |
| b | Predict the class label of the given test data using Naive Bayes Classifier<br>Two Classes: ham and spam<br>Training Data<br>*doc*1: “good.” Class :ham<br>*doc*2: “very good.” Class:ham<br>*doc*3: “bad.” Class: spam<br>*doc*4: “very bad.” Class:spam<br>*doc*5: “very bad, very bad.” Class:spam<br>Test Data<br>*doc*6: “good? bad! very bad!” Class: ? | 8 |
| c | Briefly explain the K-NN classification algorithm? | 8 |

| Q.no. | Module 6 | Marks |
|---|---|---|
| 6.a | Define clustering as an optimization problem | 4 |
| | **Answer b or c** | |
| b | Use the k-means algorithm and Euclidean distance to cluster the following 8 examples into 3 clusters: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9). | 8 |
| c | What is a recommender system? Explain the different types of recommender systems. | 8 |