|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **National University of Computer and Emerging Sciences, Lahore Campus** | | | | |
| C:\Users\saif\AppData\Local\Microsoft\Windows\Temporary Internet Files\Content.Word\final design.jpg | **Course:** | **Information Retrieval** | **Course Code:** | **CS317** |
| **Program:** | **BS(Computer Science)** | **Semester:** | **Spring 2019** |
| **Duration:** | **3600 seconds** | **Total Marks:** | **22** |
| **Paper Date:** | **12-04-19** | **Weight** | **15%** |
| **Section:** | **BS-CS** | **Page(s):** | **2** |
| **Exam:** | **Mid II** | **Reg. No** |  |
| **Instruction/Notes:** | Show all your working  Neatly write your answers on answer sheet, mention question number on top of your each solution on answer sheet  Cross out any rough work | | | |

**Question 1: [2 points]**

What is the difference between stemming and lemmatization?

**Question 2: [5 Points]**

Find the BM25 between Query Q and Document D1, the formulas are given below and the statistics of Q and D are also given

Query Q: “information retrieval”

No relevance information present i.e. S= s=0

Total document N=500,000

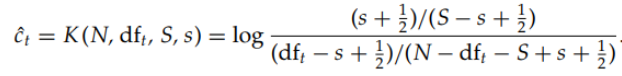
Term **information** occurs in 40,000 in all documents, and 15 times in Document D1

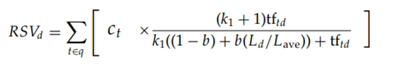
Term **retrieval** occurs in 300 in all documents, and 25 times in Document D1

Ld (length of document D) = 900

Lave (average length of documents) = 1151

The parameter values we use are k1 = 1.2, b = 0.75



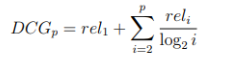


**Question 3: [5 points]**

Consider 4 document D1, D2, D3 and D4. There relevance score (***rel***) with query is 3,3,2,1 respectively. If the order in which these are returned by ranking function is

D4, D1, D3, D2

What will be the Discounted Cumulative gain of this ranking at each position p? The formula is as follow



**Question 4: [5 points]**

What will be the Probability of being in node A i.e. P(A), at time t=2, if at time t=0 P(A),P(B),P(C),P(D) is [0,0,0,1]

You have to use page rank algorithm to find P(A) at time t=2. The teleportation probability is 0.1.

You can either use the Markov’s chain method as given in Manning or iterative method as given in Bruce Croft.

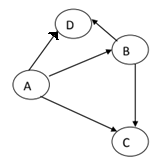


Figure 1 Graph for question 4

**Question 5: Learning to rank** **[5 points]**

Figure 1 shows the results of query “Information retrieval”, the results are ranked according to some base line algorithm, the links box are the ones on which user clicked.

From these clickthroughts, what example you will add in your training data if you are trying learning to rank using clickthroughs? Your answer should be based on the paper *Optimizing Search Engines using Clickthrough Data* by *Thorsten Joachims*, as presented in class

