|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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:** | **Fall 2016** |
| **Duration:** | **60 Minutes** | **Total Marks:** | **22** |
| **Paper Date:** | **11-Nov-16** | **Weight** | **13%** |
| **Section:** | **ALL** | **Page(s):** | **5** |
| **Exam:** | **Midterm-2** | **Roll No:** |  |
| **Instruction/Notes: *Attempt the examination on this question paper.****. You can use extra sheets for rough work but do not attach extra sheets with this paper. Do not fill the table titled Question/marks* | | | | |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Question | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Total |
| Marks | **/** 2 | **/** 4 | **/** 2 | **/** 2 | **/** 1 | **/** 1 | **/** 6 | **/** 4 | **/22** |

**Q1)** Following is page rank formula taken from Wikipedia page. What is missing in this formula? What will happen to the page rank values if we use this formula for web. M(pi) is set of pages that have link towards pi, L(pj) is total number of outgoing links from pj. [2 points]

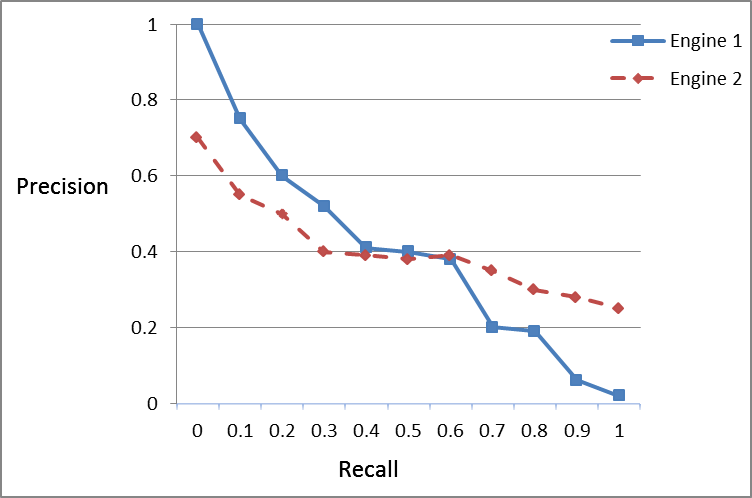
**PAGE RANK.png**

**Q2)** An IR system returns 20 results for a query. The results returned at ranks 1, 2, 4, 8, 16 are relevant; the results returned at all other ranks are irrelevant. We further know that there exist 10 relevant results for this specific query. [4 points]

(a) Compute Precision and Recall. [2 point]

(b) Compute Precision@5. [1 point]

(c) Compute Mean Average Precision. [1 point]



**Q3)** The figure above depicts interpolated precision-recall curves for two search engines that index research articles. There is no difference between the engines except in how they score documents. Imagine you’re a scientist looking for all published work on some topic. You don’t want to miss any citation. Which engine would you prefer and why? [2 Points]

**Q4)**Why do we use F measure instead of simply taking average of precision and recall? What problem will arise if we use average of precision and recall for evaluation? Illustrate with an example. [2 points]

**Q5)** Rocchio Classification Algorithm does not work well on what kind of data? [1 Point]

**Q6)** Which of the following are non-linear classifiers? [1 Point]

* Naive Bayes
* Rocchio
* KNN

**Q7)** Encircle correct option (True / False). [6 Points]

1. Bernoulli NB classifier works better for long text documents as compared to Multinomial NB classifier. True / False
2. We should select K that minimizes RSS (Residual Sum of Squares) for KMeans clustering algorithm (Select K which gives minimum value of RSS). True / False
3. We should select K that maximizes RSS (Residual Sum of Squares) for KMeans clustering algorithm. True / False
4. Good initial seeds should be close to each other for KMeans clustering algorithm. True / False
5. Time Complexity of KMeans is O(n2) where n is total number of documents. True / False
6. Macroaverage of an evaluation measure for multiclass classification problem gives equal weight to each class. True / False

**Q8)** Following figure shows training examples form three classes a, b, and c in two dimensions. Classify test document x with 1NN classification algorithm using

1. Euclidean Distance
2. Cosine Similarity

Show all calculations. The vectors are a = (0.5 1.5), x = (2 2), b = (4 4), and c = (8 6) [4 Points]

