**Information Retrieval (CS60092)**

**Computer Science and Engineering, Indian Institute of Technology Kharagpur**

**Session: Autumn 2012 – 2013**

**Class Test 1**

**Time:** 1 hour

**Full Marks:** 20

*Attempt all questions.*

*Use of calculator is allowed.*

*State any assumptions made clearly.*

**Q. 1>** For the document collection:

*D*1*: catholic church in brisbane*

*D*2*: garden city church brisbane*

*D*3*: brisbane courier garden city*

*D*4*: where in brisbane catholic church*

**a.** Draw the term-document incidence matrix.

**b.** Draw the inverted index that would be built. **(1 + 1 = 2)**

**Q. 2>** What would be the best query processing order for the Boolean queries below, given the following term postings size:

*poison* 4133

*blue* 97002

*dart* 1079

*life* 27145

*frog* 466

*cycle* 3162

**a.** (poison OR blue) AND (dart OR frog) AND (life OR cycle)

**b.** (cycle OR blue) AND (poison OR frog) AND (dart OR life) **(1 + 1 = 2)**

**Q. 3>** What would be the permuterm vocabulary for “cat”? **(1)**

**Q. 4>** What is the likely effect of (a) Stemming and (b) Lemmatization on

**(i)** Vocabulary size: Increase, Decrease, Unpredictable?

**(ii)** Precision: Increase, Decrease, Unpredictable?

**(iii)** Recall: Increase, Decrease, Unpredictable? **(3)**

**Q. 5>** Let the relevance of top ten documents (leftmost = Rank 1) retrieved for a query be:

R, NR, R, R, NR, R, NR, R, NR, NR

where R = relevant and NR = non-relevant.

For this list, plot the (i) Precision-Recall curve and (ii) Interpolated Precision-Recall curve. **(3 + 3 = 6)**

**Q. 4>** Let the top ten documents (leftmost = Rank 1) returned by an IR system for three queries be graded for relevance as (6-point relevance scale, 0-5):

*q*1: 5, 5, 3, 3, 5, 4, 2, 1, 0, 0

*q*2: 4, 3, 0, 2, 2, 1, 5, 5, 5, 5

*q*3: 4, 4, 5, 5, 5, 2, 1, 1, 1, 1

nDCG@10 = DCG@10/IDCG@10. DCG@*p* of a graded ranked list of *p* documents is given by

where *p* = 10 in this case, *reli* is the relevance rating of document at Rank *i*.

Assume IDCG@*p* = DCG@*p* for a list of *p* documents where each document has the maximum rating (5 in this case).

nDCG = Normalized Discounted Cumulated Gain

DCG = Discounted Cumulated Gain

IDCG = Ideal Discounted Cumulated Gain

Find the average nDCG@10 of the system for this result set. Show each step of the computation. **(6)**