**The LNM Institute of Information Technology**

**Mid Semester Exam-2017**

**Information Retrieval**

**Max Marks: 20 Time: 1 Hr 30 Min**

**Note: Each question carry equal marks**

**Q1. Consider a collection with the documents:**

**document1**: the way to the school is long and hard when walking in the rain

**document2**: the rain has not stopped in days and the school has closed

**query**: school closed rain

**Answer the following questions using the documents and query above.**

1. What would be the similarity score of the query with each document given above using Jaccard coefficient, if You assume there are no stop words, for this part of the question.
2. The list of stop words is (the, to, is, and, in, has, not )
3. The query is converted to a unit vector using tf-idf weighting [wt,d = tft,d \* log10(N/dft)] and Euclidean normalization. The documents are converted to unit vectors using raw tf weighting and Euclidean normalization. Compute the cosine similarity of the query with each document.

**Q2. What is document preprocessing? How do stopping and stemming reduce the size of an inverted index?**

**Q3. The following 20 ranked results have been returned as a response to a query. Since we are experts in the area, we find results 2, 6, 8, 9, 10, 12, 15, 16, 17, 18, and 20 relevant.**

**We know that there are in total 20 relevant documents in the collection.**

**(a) Draw the precision‐recall curve for the first 20 results.**

**(b) What is *precision at 10* for this list of results?**

**Q4. Recent research has suggested that Wikipedia is useful for query expansion and there are many different ways of integrating Wikipedia into query expansion systems. In your opinion, how would you use Wikipedia to improve the current query expansion systems?**

**Q5. The “Vector Space Model” is one of the most commonly used information retrieval models. Explain how it works and your answers should include how the documents and queries are represented, how the weights are calculated, and how the similarity between a query and a document is calculated.**