**Information Retrieval and Text Mining (CS567)**

Programming Assignment No. 1

**Submission Date: October 01, 2014**

**Information**

In this assignment, all you need to implement an information retrieval model called Boolean Information Retrieval Model with some simplified assumptions. You will be provided a bunch of overly simplified documents collection (corpus), you need to implement a simplified Boolean users queries that can only be formed by joining three terms (t1, t2 and t3) with ( AND, OR and NOT) Boolean operators. For example a user query may be of the form (t1 AND t2 AND t3).

Basic Assumption for Boolean Retrieval Model

1. An index term (word) is either present (1) or absent (0) in the document. A dictionary contains all index terms.
2. All index terms provide equal evidence with respect to information needs. ( No frequency count necessary, but in next assignment it can be)
3. Queries are Boolean combinations of index terms at max 3.
4. Boolean Operators (AND, OR and NOT) are allowed. For examples:

X AND Y: represents doc that contains both X and Y

X OR Y: represents doc that contains either X or Y

NOT X: represents the doc that do not contain X

As we discussed during the lectures, we will implement a Boolean Model by create a posting list of all the terms present in the documents. You are free to implement a posting list with your choice of data structures; you are only allowed to preprocess the text from the documents in term of tokenization in which you can do case folding and stop-words removal but no-stemming. The stop word list is also provided to you in assignments files. You query processing routine must address a query parsing, evaluation of the cost, and through executing it to fetch the required list of documents. A command line interface is simply required to demonstrate the working model. You are also provided by a set of 10 queries, for evaluating your implementation.

Files Provided with this Assignment:

1. Corpus as a zipped file contains (10 documents)
2. Stop-words list as a single file
3. Queries in a single file.