## CSE508 : Information Retrieval Assignment 1

Max Marks: 50

## Instructions

- The assignment is to be attempted individually
- Language allowed: Python
- For Plagiarism, institute policy will be followed
- You need to submit README.pdf, Code files and Analysis.pdf
- You are allowed to use libraries such as NLTK for data preprocessing.
- Mention methodology, preprocessing steps and assumptions you may have in README.pdf.
- Mention your outputs, analysis done (if any) in Analysis.pdf
- Submit code, readme and analysis files in ZIP format with the following name:
   A1 <roll no>.zip

## Questions

Use the 20newsgroups dataset to build a unigram inverted index.

1. Provide support for the following commands:

(25 Marks)

- x OR y
- x AND y
- x AND NOT y
- x OR NOT y

Where x and y would be taken as input from the user.

Your query output should be:

- the number of docs retrieved
- the minimum number of total comparisons done (if any)
- the list of documents retrieved.
- \* Note that the queries can be of more than 2 words of the form: "x OP1 y OP2 z" where OP1, OP2 = AND, OR, NOT. Try to write generalized code where the number of words in query can be variable.
- 2. Provide support for searching for phrase queries using Positional Indexes. (For this question, build index only on comp.graphics and rec.motorcycles)
   You may assume phrase query length to be of length less than equal to 5.