DEPARTMENT OF INFORMATION TECHNOLOGY, NITK SURATHKAL
IT458: INFORMATION RETRIEVAL (Jul – Dec 2021)

ASSIGNMENT 1

Due date: 8th Sept 2021, 11.59PM

Construct the inverted index representation for a corpus of any 100 Google News articles across different categories like Health, Sports, Entertainment, Politics, Finance etc (use the text from each article as one document).

a. Observe and report the effect of different preprocessing techniques applied to the corpus, and the changes in the vocabulary size w.r.t. final index terms, when compared to the number of initial tokens (show each step in the process clearly)

b. What type of data structure may be most optimal in storing this index? Compare and provide a detailed analysis w.r.t. the different data structure choices and give the cost analysis from storage and retrieval/insertion/deletion perspectives.

c. Using the constructed inverted index, perform some sample Boolean queries of the pattern shown below. What is the time complexity of finding the results assuming that there are N postings lists in the inverted index? Analyze and explain in detail.

    i. *term1* AND *term2* AND *term3*

    ii. *term1* OR *term2* AND NOT *term3*

Note:

1. **Submit a detailed report on your observations and analysis, supported by the necessary code snippets w.r.t your program and results observed.**
2. **Upload your report and code (well documented) on Moodle to the folder provided before the deadline of 11.59PM on 8th Sept 2021.**

EVALUATION RUBRICS:

Inverted Index construction	- 10 marks
Part a: Code and Analysis	- 10 marks
Part b: Cost analysis and observations	- 10 marks
Part c: Query complexity analysis	- 5 * 2 = 10 marks
Detailed report on observations	- 10 marks