ASSIGNMENT 1

Due date: 6th Sept 2020, 11.59PM

Marks: 10

Construct the inverted index representation for a corpus of any 100 Wikipedia articles (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?

    i. t*erm1* 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.
2. Upload your report and code (well documented) on Moodle to the folder provided before the deadline of **11.59PM on 6th Sept 2020.**