CMI/Sep-Dec 2021/Assignment 2 Oct 2021

Chennai Mathematical Institute

INFORMATION RETRIEVAL DEADLINE: OCT 30, 2021. MAX MARKS: 10.

ROLL NO.: ____________________

NAME: ________________________

You may choose any one question for this assignment.

Question 1: Use Lucene to build a simple search system. Write the code to index at least 100 text documents. You may use any of the datasets publicly available over the web. **Show that indexing the data well can impact precision and recall**. When you turn-in your assignment, submit the code, data indexed and also a report giving the precision and recall of your search system (with and without smart indexing). Do not forget to include the queries used in arriving at the precision and recall.

(or)

Question 2: Note that the soundex algorithm as discussed in the class has few problems in dealing with Indian names. For example, Mani and Mony end up with same codes. Such problems have inspired the design of improvements such as in Cologne phonetics. However, Cologne is tuned for the German language. Can you improve the soundex algorithm so that it works better with Indian names? Give a brief (one page) description of your algorithm. With sufficient examples, compare soundex map with the result of your algorithm. There is no implementation expected for this question.