CMI/Aug-Sep 2019/Assignment 2 Aug 2019

Chennai Mathematical Institute

Information Retrieval Deadline: Sep 5, 2019. Max Marks: 10.

Roll No.: _____

Name: _____

You may choose any one question for this assignment.

Question 1: Use Solr or Lucene to build a simple search system. Write the code to index at least 500 text documents. There are several datasets available over the web. For instance, you may consider indexing the short descriptions of news articles from:

`https://www.kaggle.com/rmisra/news-category-dataset.`

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. You are required to give a demo of your system before the assignment deadline.

(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.

Note: There are several datasets available over the web for Indian names such as

`http://aicf.in/assets/ratings/FIDE%20Rating%20List%202017-02.csv.`