Intro. to Text Processing and Information Retrieval (CS/DS 823)

Assignment - 1

Prof.G.Srinivasaraghavan

Date Posted: Feb 1, 2018	**Submit By**: Feb 15, 2018, Midnight	**Max. Marks**: 20

1. Pick up the Reuters R52 dataset from `https://www.cs.umb.edu/~smimarog/textmining/datasets/`. Work on this dataset using PyLucene (`http://lucene.apache.org/pylucene/`) or Lucene (`https://lucene.apache.org/core/`) directly. The objectives of the this exercise are as follows:

   (a) Implement a document retrieval system based on word/phrase queries on the given dataset, using Lucene. Figure out how to customize the query parser, and query score capabilities of Lucene. Propose customizations that work well for this dataset.

   (b) Build a simple document clustering algorithm (use any simple clustering algorithm like say k-means) based a similarity measure that you come up with. Again try customizing the extensive set of options already in Lucene.

   (c) Pick one aspect of document search/retrieval and research on (you may have to potentially look at the source code for this part) how Lucene implements it. Submit a short summary of your findings.