Assignment 1: *Content Based Document Retrieval System*

Theme: Understanding basics of text mining and retrieval systems

The goal of the project is to build a text document retrieval system by exploiting different weighting measures, and comparing their performances. While building the system, you are also expected to explore and get familiar with the related text mining basics. Your system should not only retrieve the documents, but also provide some analytical observations. Your system should provide the following outputs.

Retrieval:

1. Web based interface. Host the engine on your machine and it should be accessible from other machines through Web Browser.
2. Given a query (free form unstructured query), the task is to retrieve the set of documents relevant to the submitted query. Use TF, TFIDF and BM25 term weighting measures to retrieve the relevant documents. In separate panels, output of these three measures should be displayed. Their Precision, Recall and NDCG should be compared.
3. For each of the results, appropriate snippet should be displayed.
4. Your query model should support AND, OR and NOT Boolean operations as well as phrasal queries.

Analytics:

1. Amount of reduction in index size with and without stop-word removal.
2. Performance comparison of the system with and without term stemming.
3. Latency time for queries of different length.
4. Entropy, Zifp's and Heap's law analysis of the distribution.

Logging:

1. You should maintain a log file capturing the following information; <time of submitting query, query, clicked result, position of the clicked results>
2. You are free to capture more information

Dataset:
Each team will be using different Dataset

Evaluation Criteria: to be announced soon

Methods:

1. **You may use any of the open source search engine tools. Few of the open source source tools are**
    1. **Whoosh (Python)**
    2. **Apache Lucene (Java)**
2. **If some of the team wish to implement indexing module of your own, you are also encouraged to do so. There will be bonus mark for such group.**
3. **You are also encouraged to explore Hadoop based Map-Reduced framework to account for scalability.**

Due Date: 24th Hours, 04/09/2015 Sharp.