Continuous Assessment Part 1

(50% of CA mark; Strict deadline 22nd October)

Apache Lucene is a full-featured text search engine library written entirely in Java. It is open source and free to download . CA Part 1 is an individual assignment in which you will install and Configure Apache Lucene. Your assignment includes:

- Fully familiarize yourself with Lucene search engine

- Select appropriate Lucene Analyzers for content processing

  – tokeniser, stop-word removal, stemming, etc.

- Download and index the Cranfield Collection -

  – http://ir.dcs.gla.ac.uk/resources/test_collections/cran/

  – a collection of ~1400 documents (short abstracts)

- Implement and test different scoring approaches in Lucene

  – including at least the Vector Space Model and BM25

- Test your search engine using the 225 queries and graded (1-5) relevance judgements provided with the Cranfield Collection

  – Generate Mean Average Precision and Recall scores based upon the provided relevance judgements using TREC Eval

- Demonstrate your functioning search engine

- Write a short report (Max 2 pages not including a bibliography)

  – describes your implementation, explains your choice of analyzers

and scoring, reports the performance of your search engine

We'll be using the cloud computing platform, Microsoft Azure. See the Azure Guide on Blackboard for more details.