Information Retrieval and Text Mining (CS567)

Programming Assignment No. 2

Submission Date: November 04, 2014

IR with Lucene & Terrier using Vector Space Model

In this assignment you need to get familiarize yourself with two IR tools (i) Lucene and (ii) Terrier. You need to implement Vector Space Model to rank a set of documents that are given to you using cosine similarity between a given query and document. You are required to use either Lucene or Terrier for indexing your documents for vector space.

Lucene

Lucene is an extremely rich and powerful full-text search library written in Java. You can use Lucene to provide full-text indexing across both database objects and documents in various formats. It is supporting full-text search using Lucene requires two steps: (1) creating a lucence index on the documents and/or database objects and (2) parsing the user query and looking up the prebuilt index to answer the query. It is widely used for Text/NLP application across the globe. Download from:

Terrier

Terrier is a highly flexible, efficient, and effective open source search engine, readily deployable on large-scale collections of documents. Terrier implements state-of-the-art indexing and retrieval functionalities, and provides an ideal platform for the rapid development and evaluation of large-scale retrieval applications. Download from:
http://terrier.org/

Submission

For each query you need to provide top 10 documents, with their DocID(file name of each document). Using vector space model and cosine similarity score as a rank value.

Files Provided with this Assignment:

1. Corpus as a zipped file contains (50 documents)
2. Stop-words list as a single file
3. Queries in a single file. (3 queries)