Homework #5 for the Information Retrieval Course

Deadline: December 8, 2020

General Guideline

This homework is basically an individual-oriented work. Each student has to do it by himself or herself. The final score will be evaluated from the analysis and demonstration.

Homework Overview

Build up a biomedical abstract (text document) database and construct an index set for search. It includes the tasks of writing an agent software to retrieve medical documents from the PubMed file server periodically, building a local data base for those documents, and then making a reasonable MeSH-based index set as control vocabulary. You have to implement both BSBI and SPIMI schemes. After the indexing, you need to rank the documents based on similarity, in which you have to choose one similarity computation method. Then implement the search function on the documents.

System Description

1. This homework utilizes the PubMed medical documents Data, which can be obtained from **http://www.pubmed.gov/** for information retrieval purpose. All documents were marked up with PMID.
2. The final search should provide "MeSH-based" keyword for search. (http:// **www.nlm.nih.gov/mesh/**)
3. Note: the database schema design could either based on the MySQL or Oracle DBs.