Information Retrieval (CS4051)
Programming Assignment No. 2
Spring 2024

Submission Date: April 14, 2024

Assignment Objective

This assignment focuses on Vector Space Model(VSM) for information retrieval. You will be implementing and testing a set of queries using VSM for information retrieval. You need to build a vector space of features using some specified feature selection techniques. The dimension of the space will be R^n, the query is also represented in the same feature space. Cosine similarity is used to compute the similarity between documents and queries on normalized vectors. A given threshold can be used to filter the results for a given query. The threshold should be fixed for a given set for queries (say alpha).

Datasets

You are given a collection of ResearchPapers (File name: ResearchPapers.zip) for implementing inverted index and positional index. This zip file contains 20 papers from some research domains. These are all English language documents extracted from PDF. You need to implement a pre-processing pipeline, to get the meaning full features. It is recommended to first review the given text file for indexing. You need to treat each research paper as a unique document(DocID). This observation offers you many clues for your pipeline implementation and feature extraction.

Query Processing

The query processing of VSM is quite tricky, you need of optimize every aspect of computation. The high-dimensional vector product and similarity values of query (q) and documents (d) need to optimized.

Basic Assumption for Vector Space Model (VSM) Retrieval Model

1.Simple model based on linear algebra. Terms are considered as features using a weighting scheme TF*IDF.
2.Allows partial matching of documents with the queries. Hence, able to produce good institutive scoring. Continuous scoring between queries and documents.
3.Ranking of documents are possible using relevance score between document and query.

As we discussed during the lectures, we will implement a VSM Model by selecting features from the document by specifying tf and idf values. You are free to implement a posting list with your choice of data structures; you are only allowed to preprocess the text from the documents in term of tokenization in which you can do case folding, stop-words removal and stemming. The stop word list is also provided to you with assignments files. Your query processing routine must address a query parsing, evaluation of the cost, and through executing it to fetch the required list of documents. The list of documents should be filtered with an alpha value say (α= 0.05), A command line interface is simply required to demonstrate the working model. You are also provided by a set of 10 queries, for evaluating your implementation.

Coding can be done in either Java, Python, C/C++ or C# programming language. There are additional marks for intuitive GUI for demonstrating the working Boolean Model along with phrase query search.

Files Provided with this Assignment:

1. ResearchPapers
2. Stop-words list as a single file
3. Queries Result-set (Gold Standard- 10 example queries)

Evaluation/ Grading Criteria

The grading will be done as per the scheme of implementations, query responses and matching with a gold standard (provided query set).

Grading Criteria:
Preprocessing (3 marks)
Formation of Index (1 mark for code complexity 1 mark for saving and loading the indexes)
Vector Space Model (2 marks)
Query processing (2 marks)
Code Clarity (1 mark)
Bonus: GUI (1 mark for making the GUI 1 mark for Good Looking GUI)
The proper clean and well commented code will get 05% more marks.

<The End>