Information Retrieval System

Aditya Bhati - 16343086

March 4, 2018

1 Abstract

This project aims in development of an Information Retrieval system using open source Apache Lucene 7.2.1 and Java 1.8. This program performs splitting, indexing and searching on cran.all.1400 documents and shows relevance results from the given queries.

2 Program Structure

App.java: This is the main program which calls all the functions in the other class. Indexer is the object used to call functions such as parser, index files and scorer etc.

IndexFiles.java: In order to achieve relevant results we first need to clean the document by removing punctuations, stop word etc. regex was used to achieve this. In this class we split the two documents cran.qry and cran.all.1400 and store them in the crandoc directory. Then Indexing is performed in which we give path of the results directory. After this searching and querying is performed in which same path is given as in Indexing. Results were obtained from two types of analyser.

3 Types of Analyzers used

• Standard Analyzer : This analyser provides grammar based tokenization, It consist of standard tokenizer, token filter, stemming etc. converts the string in lists format for analysing the document. Helps in removing punctuation and unwanted characters and also transforms token stream.

• Simple Analyzer: This analyser filters letter tokenizer with lower case filter.

4 Scoring/Ranking:

It's the procedure in which we get top relevant documents when given a query to the search engine.Two types of scoring method are being performed in this program Vector Space and BM25.

- Vector Space: It converts text documents into vectors of identifiers. Relevance is calculated using document similarity theory by comparing deviation in between document vectors.
- BM25: It is also known as bag-of-word, based on probabilistic retrieval framework, this matched documents according to their relevance for a given search query, It was observed that BM25 performed better in comparison with vector space model

5 Evaluation

Trec evaluation is used which is a standard NSIT evaluation procedure for query processing. Evaluation is established on the basis two files cranqrel(query relevance) and results generated by the retrieval system.

- Cranqrel: This file consist list of documents that are relevant for the queries.
- Results: This file contains automatically generated ranking results for each query.

6 Observation

Evaluation Metrics			
Standard Analyzer	MAP	bref	Recall1000
Combined VectorSpace-BM25	0.1801	0.5220	0.8178
Simple Analyzer	MAP	bref	Recall1000
Combined VectorSpace-BM25	0.1769	0.5455	0.8621

After evaluating we observe that standard analyzer performs better than simple analyzer. MAP for standard analyzer is 0.1801 and for simple analyzer is 0.1769 which is greater for standard analyzer.

7 Output and Execution

Public DNS (IPv4): ec2-52-30-37-124.eu-west-1.compute.amazonaws.com
Username - [redacted] : Password - [redacted]
System login : ssh [redacted]@ec2-52-30-37-124.eu-west-1.compute.amazonaws.com

```
cd InformationalRetrieval
chmod +x evaluation.sh
./evaluation.sh
```