Search Engine Design

CS6200: Information Retrieval

Fall 2016, Prof. Nada Naji

Team Members:

Partha Sarathi Jena

Rahul Pyne

Tridiv Nandi

# **1. Introduction:**

## 1.1 Overview:

The goal of the project is to design and build our own information retrieval systems, evaluate and compare their performance levels in terms of retrieval effectiveness. This project designs search engines using four distinct retrieval models as mentioned below:

* BM25 retrieval model
* tf-idf retrieval model
* Cosine Similarity retrieval model
* Lucene Systems

Along with these four baseline runs, we have three additional runs using the following techniques:

* *Cosine Similarity* retrieval model along with *Pseudo relevance feedback* query expansion technique
* *Cosine Similarity* retrieval model along with *Stopping* technique
* *BM25* retrieval model along with *Stopping* technique

Now we assess the performance of these seven distinct runs in terms of the following retrieval effectiveness measures:

* Mean Average Precision (MAP)
* Mean Reciprocal Rank (MRR)
* P@K measure where K=5 and K=20
* Precision & Recall
* Moreover, we perform another run using *Cosine Similarity* retrieval model on the stemmed version of the corpus *‘cacm\_stem.txt’* using *‘cacm\_stem.query’*.

## 1.2 Contribution of team members:

All the team members were responsible for taking the design decisions regarding implementations of the project. The detailed contribution of each individual member is elucidated below:

* Partha Sarathi Jena: Partha was responsible for implementing the query expansion technique using *Pseudo relevance feedback*. Additionally, he has developed the search engine which incorporates the stemming and stopping techniques. He was also responsible for the corresponding documentation and query-by-query analysis for stemmed and non-stemmed runs.
* Rahul Pyne: Rahul was responsible for implementing the four baseline runs using tf-idf measure, BM25 Model, Cosine Similarity Model and Lucene Systems. Moreover, he has developed the business logic for the seventh run where we have implemented BM25 along with stopping technique. He was also responsible for documenting his design choices and rationale behind his decisions.
* Tridiv Nandi: Tridiv was responsible for implementing the various evaluation measures like MAP, MRR, P@K, Precision & Recall for the seven distinct runs. Moreover, he made notable contributions in documenting the conclusions and outlooks from the findings, observations and analyses of the results.

# **2. Literature and resources:**

## 2.1 Overview:

The overview of the techniques used for implementing various aspects of the project are outlined below:

* tf-idf measure: For calculating tf-idf for each document, we are using the raw value of term frequency for the document multiplied by its inverse document frequency.
* Cosine Similarity: For calculating cosine similarity, we have used tf\*idf as the term weights for the query and document terms.
* BM25 Model: For BM25 model, we have used ‘*cacm.rel*’ file as the set of relevant documents. The values of ‘*K*’, ‘*k1*’ and *‘k2’* are chosen as per TREC results.
* Lucene: We have used the standard Lucene open source library with minor modifications to perform indexing and search operations.
* Query Expansion: The query expansion for our project has been done using Pseudo Relevance Feedback technique
* Stopping: The standard stop list – ‘common\_words.txt’ has been used to perform stopping. Any word appearing in the above-mentioned stop list has not been indexed.
* Stemming: We have performed stemming using the Cosine Similarity model with the help of the stemmed version of the corpus *‘cacm\_stem.query’* and query *‘cacm\_stem.query’*.
* Precision & Recall: The formulae used for calculations of precision and recall are:
  + - Precision = |Relevant ∩ Retrieved| / |Retrieved|
    - Recall = |Relevant ∩ Retrieved| / |Relevant|
* MAP: The formulae used for calculations of Mean Average Precision is:
  + - MAP = Σ Average Precision / Number of Queries
* MRR: *Reciprocal rank* is reciprocal of the rank at which the first relevant document is retrieved. Mean Reciprocal Rank is the average of the reciprocal ranks over a set of queries.
* P@K: P@K is calculated as the number of relevant documents in the *top K* retrieved documents. We have calculated for K = 5 and K = 20.

## 2.2 Third Party Tools:

The following third party tools have been used in moderation as and when required:

* Beautiful SOUP: Beautiful SOUP has been used for parsing purpose to process both the documents in the corpus and the queries.
* Lucene Libraries: In order to do the local setup for Lucene and run the code, we need 3 following referenced libraries:
* lucene-core-VERSION.jar
* lucene-queryparser-VERSION.jar
* lucene-analyzers-common-VERSION.jar

## 2.3 Research Article References:

The following articles were consulted for the various evaluation measures:

* <http://web.stanford.edu/class/cs276/handouts/EvaluationNew-handout-6-per.pdf>
* <https://ils.unc.edu/courses/2013_spring/inls509_001/lectures/10-EvaluationMetrics.pdf>

The following articles were consulted for Query Expansion Technique using Pseudo Relevance Feedback:

The following articles were consulted for designing the four baseline runs:

# **3.Implementation and Discussion:**