INTRODUCTION:

GOAL:

The goal is to implement a retrieval system both from scratch and by using Lucene. For the first retrieval system, we used three retrieval models namely BM25, tf-idf and Query Likelihood Model with Jelinek Mercer Smoothing. We ran these retrieval models both with and without stopping and stemming of the corpus. We also performed query enrichment on [model name] and retrieved the ranked documents. The top 100 ranked documents of the cacm test collection were displayed along with snippets using [ ]snippet generation technique and the query terms were highlighted as well. To assess the effectiveness of each of these retrieval system runs evaluation techniques of Precision and Recall Mean Average Precision, Mean Reciprocal Rank and Precision at Values at K were used.

The project was divided into the following phases:

Phase 1: Indexing and Retrieval

Task 1 : Building retrieval system with Lucene,bm25,tf-idf and Query Likelihood Model

Task 2: Performing query enrichment on []model run

Task 3 :

1. Performing stopping on the corpus and running all the retrieval system models.
2. Running all the retrieval system models on the stemmed corpus and analyzing the results of the 3 interesting queries.

Phase 2: Implementing snippet generation and query term highlighting to display the results

Phase 3: Evaluation using MAP,MRR,P@K,Precision and Recall and plotting a Recall and Precision curve.

MEMBER CONTRIBUTION:

Phase 1: Task 1 - Sanhitha Murthy

Task2 - Sanhitha Murthy

Task 3: Monisha Carol Karise

Phase 2: Dipanjan Haldar

Phase 3: Monisha Carol Karise

Extra credits: Dipanjan Haldar