**Information Retrieval Project Report**

i.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sr. No. | Project Members’ Names | Course Name | Semester | Instructor Name |
| 1 | Akshat Shukla | MSCS | 1 | Prof. Nada Naji |
| 2 | Parshva Shah |
| 3 | Virat Goradia |

**ii. Introduction**

The project is divided into 3 phases as explained below:

(Contribution of each Project member is mentioned in brackets)

1. **Phase 1: Indexing and Retrieval**

It comprises of 3 Tasks as below:

1. Task 1: This task comprises of 4 steps:
2. Step 1: This step takes Raw HTML files (given corpus) as input, and returns same number of files but with each input file tokenized by case-folding and removing punctuations
3. Step 2: This step takes output of Step 1 as input and generates an inverted index and a document which maps Document ID and its length.
4. Step 3: This step takes the given query file and cleans it to convert into a file containing a mapping between query id and corresponding query.
5. Step4: This step constitutes implementing the following models for score calculation and finding top 100 documents for each:

* BM25 (Considering no relevance)
* BM25 (Considering relevance)
* Lucene
* Query Likelihood Model
* Tf-idf

1. Task 2: In this task, Pseudo Relevance Model is used to re-compute scores calculated for BM25 Model which considered relevance. The output comprises of top 100 documents according to this model.
2. Task 3: This task comprises of 2 parts:
3. Part A: This part consists of performing stopping on the corpus as well as queries (using common\_words.txt) to perform scoring by performing 4 steps same as those in Task 1:
4. Step 1: Tokenization
5. Step 2: Inverted Index Generation
6. Step 3: Query generation
7. Step 4: Score calculation and finding top 100 documents using:

* BM25 (Considering no relevance)
* BM25 (Considering relevance)
* Query Likelihood Model
* Tf-idf

1. Part B: This part is divided into 3 steps:
2. Step 1: Tokenization of the given stemmed corpus (Using cacm\_stem.txt)
3. Step 2: Inverted List Creation
4. Step 3: Finding top 100 documents by calculation scores for each using the following models:

* BM25 (Considering no relevance)
* BM25 (Considering relevance)
* Lucene
* Query Likelihood Model
* Tf-idf

1. **Phase 2: Displaying Results:**

This phase consists of snippet generation for the top 100 documents found for each query for each of the following models:

* BM25 (Considering relevance)
* Lucene
* Query Likelihood Model
* Tf-idf

1. **Phase 3: Evaluation:**

This phase consists of evaluating the results found by applying the following models for calculating scores:

* BM25 (Considering no relevance)
* BM25 (Considering relevance)
* Lucene
* Query Likelihood Model
* Tf-idf

It consists of 2 steps:

1. Step 1: This step consists of forming encoded dictionaries which are used by the Step 2 for evaluating the results generated by the above-mentioned models.
2. Step 2: This step consists of evaluating the results by calculating different evaluation measures for each result set like precision-recall tables, MAP, MRR, P@K.

**iii) Literature and Resources:**

Following approaches were used while performing each of the below mentioned tasks:

1. Phase 1 Task 1 Step 1: