Following are types to be followed to run the code :

AUTHOR :

- PANKTI BHALANI

- DARSHAN PANSE

- AMUL MEHTA

1. Clean the Corpus :

- python clean-cacm-corpus.py

- Output :- Folder : cacm\_corpus

- All the Files in the Folder with cleaned Corpus

2. Clean the Query :

- python query-clean1.py

- python query-clean2.py

- Output : query\_cacm.txt (Cleaned Query File)

All the runs File are Stored in the Folder named "Output". This Files are the Workbooks which have different worksheets and each worksheet has output per query i.e top 100 Documents for the given run.

3. Task 1 Compilation :

- Run TFIDF on Console :

- python create\_retrieval\_model.py 1

- Output : tfidf\_retrieval\_cacm.xlsx

- Run Cosine on Console :

- python create\_retrieval\_model.py 2

- Output : cosine\_retrieval\_cacm.xlsx

- Run BM25 on Console :

- python create\_retrieval\_model.py 3

- Output : bm25\_retrieval\_cacm.xlsx

- Run Lucene :

1. Open the Lucene Project in the Exclipse or any other Java-IDE

2. Run the Project

3. Enter the Index path for the Corpus : which should be the path of a foldr where the index will be created.

4. Question 2 :

Enter the path where your Corpus is : in my case it is in "cacm-corpus" folder so give the path for this foldr.

- cacm-corpus is a folder where my parsed files are saved according to the document-names

5. Query will be read from the file named query\_cacm.txt

6. Ouput will be in a file named "LuceneCACM\_Relevant.xls"

4. Task 2 Compilation :

- Run Task1 :

- python query-expansion.py

- Output : query-expansion-bm25-verison1.xlsx

5. Task 3 Compilation :

- Run Task3 part A :

- python stopwords-verison-task3.py

- Output : stop-query-expansion-versions1.xlsx

- Run Task3 part B :

- python create-stemmed-index.py

- python stemmed-verison-task3.py

- Output : stemmed-bm25.xlsx

6. Implementation – Phase 2 :: Evaluation :

- Run :

- python cosine-stopwords.py 2

- Output : cosine\_retrieval\_stop\_cacm.xlsx

- EVALUATION :

- python evaluate-system.py

- Output : 7 Excel Files displaying the Precision, Recall, P@5 and P@20 Values for each run, each excel file will contain worksheet per Query. Also, an Additional Excel File where the per System MAP and MRR are stored.

1. bm25-evaluation.xlsx

2. cosine-evaluation.xlsx

3. tf-idf-evaluation.xlsx

4. query-exp-stop-words-evaluation.xlsx (Query Expansion with Stop words)

5. query-exp-v1-evaluation.xlsx (QueryExpansion with Psuedo Relevance)

6. stopped-evaluation-25.xlsx (BM-25 with Stop-words)

7. lucene-evaluation.xlsx

8. System-wise-Evaluation.xlsx (MAP and MRR for all the given 7 runs)

9. cosine-stop-evaluation.xlsx

7. Snippet Generation (BONUS)

- Run :

- first go to the Folder "Snippet Generation" and run.

- python snippet.py

- OutPut will be a HTML file named : snippet.html