# Homework 2

## Create Inverted Index:

The submission has mainly two scripts – one is for creating tokens and the other one is for creating the index and offset files. They are divided into four categories depending on whether the stopwords are eliminated while creating the index and whether the words are stemmed. These are stored in four different folders given below:

1. **IndexwithStopNoStemmed** : The files in this folder create inverted index that has all the stopwords. The words are not stemmed.
2. **IndexwithStopStemmed** : The files in this folder create inverted index that has all the stopwords. The words are stemmed.
3. **IndexNoStopNoStemmed** : The files in this folder create inverted index that does not have the stopwords. The words are not stemmed.
4. **IndexNoStopStemmed** : The files in this folder create inverted index that does not have the stopwords. The words are stemmed.

The sizes of the token and index files are:

| Stop words present? | Words Stemmed? | Token file size (MB) | Index file size (MB) |
|---|---|---|---|
| Yes | No | 254.1 | 293.1 |
| No | No | 153.3 | 180.8 |
| Yes | Yes | 249.5 | 286.6 |
| No | Yes | 149.8 | 175.9 |

## Document-Query Scores:

The scoring models have been implemented only on the index which has no stopwords and the words are stemmed. The output files are present in the ScoreFiles_NoStop_Stemmed folder. Below is the summary of the precision obtained by running the scoring scripts.

| **Model** | **Output filename** | **Average Precision** | **Exact Precision** |
|---|---|---|---|
| Okapi TF | okapi_output.txt | 0.1487 | 0.1751 |
| TF-IDF | tf_idf_output.txt | 0.2274 | 0.2618 |
| Okapi BM25 | okapi_bm_output.txt | 0.2219 | 0.2468 |
| Unigram LM with Laplace Smoothing | unigram_lm_output.txt | 0.1405 | 0.1765 |
| Unigram JM with Jelinek-Mercer Smoothing | unigram_jm_output.txt | 0.1894 | 0.2192 |
| Proximity Search | proximity_output.txt | 0.1147 | 0.1612 |