**Assignment 1**

**Question:**

Implement a search engine by choosing a relevant corpus.

The search engine should include implementation of inverted index.

The implementation should include i)Boolean Model ii) retrieval based on similarity

The text content should be pre-processed before the creation of inverted index.

Implementation should support phrase queries (combination of Bi-word index and positional indexing).

The retrieved documents should appear in a ranked manner.

Precision and Recall should be computed.

**Analysis Report:**

In addition to implementation, each group should submit a detailed Analysis report that should include

1. Justification of pre-processing techniques used
2. justification for choosing an appropriate data structure

iii) justification for choosing a typical similarity scoring scheme

**Group size of 4**

**Evaluated for 20 and reduced to 10 Marks**

**Evaluation Parameters:**

1. Pre-processing of raw data

2. Generate Inverted Index (variation in data structures)

3. Handling wild card and phase queries

4. Retrieve relevant text using similarity index **(2 marks)**

5. Ranking of retrieved documents

6. Advanced search: relevance feedback, semantic matching, re-ranking of results, finding out query intention **(2 marks)**

7. Demonstration and Analysis Report

**Datasets (not exhaustive):**

1. Tweets

2. IMDB

3. RCV1

4. TREC datasets

5. Collection of Wikipedia pages

etc...