Text Retrieval & Search Engine (CP423)

Assignment 1

Max Marks: 100

Instructions:

- This assignment requires group work with **2 members** per group. Only one member needs to submit the work.
- Please use Python as the programming language. If you are unfamiliar with Python, this is an opportunity to learn it. Please refer to the Python programming language resources provided in MyLS.
- Submit the code files with proper commenting.
- You are allowed to utilize libraries such as NumPy and NLTK for data preprocessing.
- Download the dataset, which is approximately 15MB in size and consists of 467 files.

Question 1: Text Preprocessing [25 points]

Perform the following preprocessing steps on the given dataset:

- Convert all text to lowercase.
- Tokenize the text using NLTK.
- Remove stop words using NLTK.
- Exclude special characters except alphanumeric characters.
- Eliminate singly occurring characters.
- Create a set of all the words.

Question 2: Inverted Index Implementation [25 points]

Implement an inverted index data structure for the preprocessed dataset.

Question 3: Query Support [25 points]

Support the following queries, where x and y would be taken as input from the user.:

1. x OR y
2. x AND y
3. x AND NOT y
4. x OR NOT y

Note: Aim to write generalized code that can handle queries with a variable number of words in the format "x OP1 y OP2 z," where OP1 and OP2 can be AND, OR, or NOT.

Question 4: System Evaluation [25 points]

Evaluate your system against the set of provided queries. Marks will be awarded based on the accuracy of the output. Your output should include:

- The number of documents retrieved.
- The minimum number of total comparisons made (if applicable, only for the merging algorithm).
- The list of retrieved document names.

Apply preprocessing to the input queries as well.

Submission Format:

- **Source Code:**
  - Provide the complete source code for all practical implementations.
  - Ensure that the code is well-documented and properly formatted for readability.
- **Video Walkthrough:**
  - Submit a video walkthrough of your entire project.
  - In the video, demonstrate the functionality of your code and explain the logic behind your implementations.
  - For each question, briefly discuss your approach, the results obtained, and any insights or learnings.
  - Ensure the video clearly shows the code running and the outputs generated.
  - **Failure to submit this video will result in a zero grade for the assignment.**

Input format:
The first line contains the number of queries, N.
The next 2N lines represent the queries. Each query consists of two lines:
a) Line 1: Input sentence
b) Line 2: Input operation sequence

Example queries:

1. **Query #1:**
   Input sentence: “lion stood thoughtfully for a moment”
   Input operation sequence: [OR, OR, OR]

   Expected preprocessed query: “lion **OR** stood **OR** thoughtfully **OR** moment”

   Output:
   Number of matched documents: 270
   Minimum number of comparisons required: 671
   List of retrieved document names

2. Query #2:
   Input sentence: “telephone, paved, roads”
   Input operation sequence: [OR NOT, AND NOT]

   Expected preprocessed query: telephone **OR NOT** paved **AND NOT** roads

   Output:
   Number of matched documents: 466
   Minimum number of comparisons required: 739
   List of retrieved document names