Text Retrieval & Search Engine (CP423B)

Assignment 1

Max Marks: 100

Instructions:

- This assignment requires group work with 4 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 8MB in size and consists of 249 files.

Question:

a. [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.

b. [25 points] Implement the inverted index data structure.

c. [25 points] Support the following queries, where x and y would be taken as input from the user.:

i. x OR y
ii. x AND y
iii. x AND NOT y
iv. x OR NOT y

d. [25 points] Evaluate your system against a set of queries provided below. Marks will be awarded based on the accuracy of the output.

- The query 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.

Note:

- Aim to write generalized code that can handle queries with a variable number of words. Queries can consist of more than two words in the format: "x OP1 y OP2 z," where OP1 and OP2 can be AND, OR, or NOT.
- Apply preprocessing to the input query as well.
- The number of operations specified for a query assumes that suitable preprocessing steps have been applied.

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