CSE508: Information Retrieval

Assignment 1

Max Marks: 40

Instructions-

- The assignment is to be attempted in groups (same group as your project)
- Language allowed: Python
- For plagiarism, institute policy will be followed
- You need to submit README.pdf and code files. The code should be well commented.
- You are allowed to use libraries such as NLTK for data preprocessing.
- Mention methodology, preprocessing steps, and assumptions you may have in README.pdf.
- You will be required to use Github for code management.
    - Each group will create a GitHub repository with the name IR2021_A1_GroupNo (Eg - IR2021_A1_1 for Group No-1).
    - Each group would add the assigned TA as a collaborator to the GitHub repository. TAs' GitHub handles would be shared shortly.
    - While uploading on Classroom, each group would need to upload a link of the GitHub repository. Only one member needs to submit.
- You will have 10 days to complete the assignment.

Question-

Download the stories dataset from the given link: http://archives.textfiles.com/stories.zip

[Data Size is approximately 15MB and 467 files]

a. [8 points] Carry out the suitable preprocessing steps on the given dataset
b. [8 points] Implement the unigram inverted index data structure.
c. [1+1+2+2 = 6 points] Provide support for the following queries-
    i. [1 point] x OR y
    ii. [1 point] x AND y
    iii. [2 points] x AND NOT y
    iv. [2 points] x OR NOT y
d. [18 points] During the demo, your system would be evaluated against some queries in the format mentioned below. Marks would be awarded based on the correctness of the output.

Where x and y would be taken as input from the user.
Your query output should include:

- The number of documents retrieved
- The minimum number of total comparisons done (if any)(only in merging algorithm)
- The list of document names retrieved

Note-

- Try to write generalized code where the number of words in the query can be variable. The queries can be of more than 2 words of the form: "x OP1 y OP2 z" where OP1, OP2 = AND, OR, NOT.
- Perform preprocessing on the input query as well.
- The number of operations specified for a query would be under the assumption that the suitable preprocessing steps have been applied.

Input format:
The first line contains the number of queries, N.
The next 2N lines would represent the queries.
Each query would consist of two lines:

a) line 1: Input sentence
b) line 2: Input operation sequence

Some example queries-

1. **Input query:** lion stood thoughtfully for a moment
   Input operation sequence: [OR, OR , OR]

   Expected query after preprocessing: lion OR stood OR thoughtfully OR moment

   Output-
   Number of documents matched**:** 270
   No. of comparisons required: 671

2. **Input query:** telephone,paved, roads
   Input operation sequence: [OR NOT, AND NOT]

   Expected query after preprocessing: telephone OR NOT paved AND NOT roads

   Output-
   Number of documents matched**:** 466
   No. of comparisons required: 739