Readme File

Methodology

1. Preprocessing the data
2. Implementation of unigram inverted index
3. Performing the operations (OR, AND, OR NOT, AND NOT)

Explanation

1. Preprocessing the Data-
   In preprocessing of data, the steps include are-
   - converting the data in lowercase
   - removing the stopwords
   - removing punctuation, apostrophes and other symbols
   - removing the one letter words
   - replacing numeric values with their corresponding text values
   - performing stemming on words in the data
2. Creating posting lists for the words in the data
3. Converting Dataframe postings to a dictionary format
4. Creating a set for all 467 documents
5. Evaluating the queries-
   In order to evaluate the input queries, the steps include:
   - Taking two words and their corresponding operation
   - Calling different functions for OR, AND, OR NOT and AND NOT respectively
   - Applying the suitable operation on the two input words
   - Storing the documents matched as a dataframe
   - Summing up the number of comparisons
   - Printing the number of documents matched as the length of the dataframe and printing the total number of comparisons that were required

Assumptions

1. Working for N queries with next line 1 as input and line 2 as the sequence of operations.
2. Operation sequence should be inside [ ] and comma-separated
3. Taking two words of the input sequence at a time and putting the operations in sequence.
   Eg:

   input seq: lion stood thoughtfully for a moment
   sequence of operations: [OR, OR , OR]
   Expected query: lion OR stood OR thoughtfully OR moment