IR Assignment 1 README FILE

Question 1:

METHODOLOGY:

- Pre-processing is done for the dataset and the inverted index is created for it.
- The inverted index is created using a dictionary.
- Now when a query along with the operators is entered by the user the query is evaluated and the no. of documents which satisfy the Boolean query as well as the list of documents is displayed to the user.

ASSUMPTIONS:

- I have done the optimization to minimize the no. of comparisons only when the query consists of only AND operations.
- When the query consists of only OR, only OR NOT, only AND NOT. In all these 3 cases the expression is being evaluated from left to right.
- In case the query expression consists of more than one type of operators i.e. AND, OR, NOT then the precedence order followed is AND NOT > OR NOT > AND > OR

PRE-PROCESSING STEPS:

- All text has been converted to lower case.
- All email-ids have been removed from the data.
- All digits and punctuations have been removed.
- Words of the type wouldn't have been converted to wouldnt while creating the inverted index as well as while processing the query.
- All stop-words have been removed and lemmatization has been performed in the data.

All the above pre-processing steps have been performed on the dataset while creating the inverted index as well as on the query while processing the query. The query tokens are displayed to the user after applying pre-processing and then the user can enter the operators.

Question 2:

METHODOLOGY:

- Pre-processing is done for the dataset and the positional index is created for it.
- The positional index is created using a nested dictionary.
- Now when a phrase is entered by the user to be searched, first it is also pre-processed and then the no. of documents containing the phrase as well as the list of documents is displayed to the user.

ASSUMPTIONS:

- If the user enters a phrase which contains some numbers then I have removed the numbers.

PRE-PROCESSING STEPS:

- All text has been converted to lower case.
- All email-ids, digits and punctuation marks have been removed from the data.
- Words of the type wouldn’t have been converted to wouldnt while creating the positional index as well as while processing the query.