IR Assignment 1 (README)

Akshaj Patil(MT19111)

1:PreProcessing Steps:

- Case folding (all terms are converted to lower case).
- Stop words removal.
- Removal of punctuations.
- Removal of numbers.
- Lemmatization.

Same pre-processing is done for data and query.

2:Assumptions:

- Meta data of the documents is considered as part of document.
- Query should be in Boolean form. Eg(A and B or not C..)
- All the documents are tagged as numbers starting from 1 to 19997.
- Query should be of at least 1 word.
- “and”, “or” and “not” should not appear as operand.

3:Methodology for Q1:

In question 1 I first pre-processed data and constructed inverted index on that data. All unique words are stored as dictionary keys and their document id’s in which they appear are stored as value of that particular key in list.

Query is pre-processed in similar fashion. If by mistake user enter consecutive two “and” or consecutive two “or” or consecutive two “not” then one of the operator is removed.

After processing operations are performed on query in “not”, “and” and “or” precedence”. For optimization I performed operations on values with minimum frequency posting list.

3:Methodology for Q2:

In Q2 I first pre-processed data and constructed inverted positional index on that data. Here dictionary of dictionary is made. All unique words are stored as keys and there values are also dictionary with document id as a key with list value of positions of particular word in that document.

Query is pre-processed. By Iterating two words and performing combination and again storing the result in this fashion operations are performed. Here while combining we have to see that two words are consecutive in that document, if yes then only combine.