CSE508: Information Retrieval
Assignment 1
Group5

Pre-processing:

- Tokenization:
  The sentences in the data are split into words using the word_tokenize package of the NLTK library.

- Converting to lowercase:
  All the tokenized words are converted into lowercase to bring consistency in the data so as to map different forms of the same word to a single word. This normalizes the text.

- Removing punctuations:
  The characters are matched to the following and replaced with an empty string:
  \w : Matches Unicode word characters
  \s : Matches Unicode whitespace characters (which includes [\t\n\r\f\v], and also many other characters

- Removing stopwords:
  Some words which are commonly used are removed from the text since they provide less information.

- Stemming:
  Words are transformed into their root forms and this helps in reducing the size of the index.

Creation of Inverted Index:

- createDictionary(): Removed special characters from sentences and created terms to docID mapping.

Query Processing:

Function Process the given query

- AND: Intersection of the given 2 posting lists
  Simple merge algorithm, which compares the docID of both lists and adds them to a third list, if it exists in both.

- OR: Union of the given 2 posting lists
  Simple merge algorithm, which compares the docID of both lists and adds them to a third list, if it exists in either.
- notAND: Uses doclist as a universal set and calculates negation of 2nd posting list and then uses AND function on 1st posting list and the 2nd.
- notOR: Uses doclist as a universal set and calculates negation of 2nd posting list and then uses OR function on 1st posting list and the 2nd.
- process: Takes the 2 posting list and boolean expression needed, and applies the appropriate function.

User gives a query and expression as input, sets the first word and extracts its posting list, and loops over the input query and expression. Evaluate the posting lists of each input and evaluate the expression, respectively.

Assumption :

- Preprocessed file is saved to RT file and from that RT file comparisons have been done.
- Stemming is used, hence number of matches high
- Punctuation, tokenization and stop words are used during pre processing
- Two coding files are made
  - Preprocessing file
  - Query File

https://github.com/CaptainPramil/IR2021_Group5_Assignment1

Output Screenshot