Pre-processing

Stop words were removed using the NLTK packages. Punctuations were removed with white spaces. All words with less than two characters and digits were removed. Every word was converted into lower case. A test was done based on stemming and lemmatizing just to extract the tokens but then it did not give right tokens due to which those steps were removed. These pre-processing steps were followed in both the question.

Methodology

1. A dictionary was created which contained words as key and document id as values for the first question. This was stored in a file Dictionary_word.pkl just to make sure so that next time when we run the code, we don't need to do the same pre-processing again. A generalised solution with multiple function has been provided to perform the task for n operands with n-1 operators.
2. A dictionary of dictionary was created in which key for the outer dictionary was words, whereas key for the inner dictionary was document id and value of that was position where the token is present. This was stored in a file Dictionary.pkl just to make sure so that next time when we run the code, we don't need to do the same pre-processing again. Positional index was created through this pkl file. The phrasal query was done based on words present next to each other and it could exceed to any word limit with a generalized solution.