Assignment 4

Q2.4. Test your solution using different options in in the tokenize function, i.e. with or without lemmatization, with or without removing stop words, to see how these options may affect the accuracy.
Analyze why some option works the best (or worst). Write your analysis in a pdf file.

Ans. Accuracy is affected separately according to different options. Lemmatization improves the accuracy because the different forms of words are reduced to their base forms. The best option is the removal of stop words along with lemmatization because the words which do not contribute in helping analyze the similarity are also removed from the questions.

Q3.2. Analyze the following questions

If you change the similarity threshold from 0.1 to 0.9, how do precision and recall change?

Ans. As the similarity precision is changed from 0.1 to 0.9, precision becomes high while recall becomes very low.

Consider both precision and recall, do you think what options (i.e. lemmatization, removing stop words, similarity threshold) can give you the best performance?

Ans. Considering both precision and recall, lemmatization along with removing stop words combined with a similarity threshold in the mid-range e.g. 0.5 can give us the best performance.

What kind of duplicates can be easily found? What kind of ones can be difficult to find?

Ans. The duplicates with same type of words but different implications can be easily found while the ones with less words that are of same types can be difficult to find.

Do you think the TF-IDF approach is successful in finding duplicate questions?

Ans. Yes, TF-IDF approach is successful in finding duplicate questions.