RQ 1.1: Why do we remove common stop words and lemmatise the text? Give an example of a scenario where, in addition to common stopwords, there are also application-specific stop words.

Answer:

In the lab assignment, we need to find the nearest documents for the query. In order to do this, we have to get the df-idf matrix. The documents we can find through this way should contain some special information that are similar to query. Although there might be a lot of stop words that are in common in both, these common terms are less meaningless like 'a','the'. In other words, we cannot conclude two documents are similar just because they have some stopwords. Besides, removing these meaningless words can reduce the calculation burden.

Lemmatising words can ensure that words with different forms can be regarded as the same words. Otherwise, even there are some common meaningful words in document and query, it is possible the corresponding vectors are not similar.

In Medical Text Analysis, "mg" (milligram) can be regarded as one of the stopwords. As a unit, this is common in most of document and meaningless in terms of differentiation.

RQ 1.2: In Problem 2, what do the dimensions of the matrix X correspond to? What information from the original data is preserved, what information is lost in this matrix?

Answer:

The number of rows corresponds to the number of documents and the the number of columns corresponds to the number of words that occur in all of the documents.

As for the information, we cannot read out the original context from this matrix. However, we can still know the number of original documents.

RQ 1.3: What does it mean that a term has a high/low idf value? Based on this, how can we use idf to automatically identify stop words? Why do you think is idf not used as a term weighting on its own, but always in connection with term frequency (tf–idf)?

Answer:

The high tf-idf value mean the proportion of files containing the term is low, which means this term carries specific information of certain files and vice versa.

Based on this, we can indeed use idf values to identify stop words, since these words have low idf values.

This is because we should use weights to calculate the score (the similarity between two documents). In this case, if one files contains have more common words with the query than another file, even though the difference is only about amount these words instead of types, we can easily conclude that the first file should be similar to the query. Based on this, we have to consider the amount of terms in files, and thus only idf value is not enough.