Stop words

Some search engines don't record extremely common words in order to save space or to speed up searches. These are known as "stop words." In computing, **stop words** are words which are filtered out prior to, or after, processing of natural language data (text). There is not one definite list of stop words which all tools use and such a filter is not always used. Some tools specifically avoid removing them to support phrase search.

Any group of words can be chosen as the stop words for a given purpose. For some search machines, these are some of the most common words, such as *the*, *is*, *at*, *which*, and *on*. In this case, stop words can cause problems when searching for phrases that include them, particularly in names such as 'The Who', 'The The', or 'Take That'. Other search engines remove some of the most common words including lexical words, such as "want"—from a query in order to improve performance.

There are two files attached with this mail

1. List of Stop Words (mystwords.txt)

2. 10 documents (docs.rar)

Write Java programs to:

Q1. Remove the stop words (listed in “mystwords.txt”) present in all the files in “docs.rar” and create files without stop words and save them as “ndocs.txt”.

Q2. Find the list of unique words in all doc files.

Q3. Find the frequency matrix of each unique word in files “ndocs.txt“.

Please Note: Time and space complexity will be the evaluation criteria and use of Collection framework.