**TASK 3**

**Stoplist:**

Algorithm:

* Take the first 100 terms from the corpus whose frequency is highest after sorting the term frequencies table.
* Compare these term with the stopword list provided by nltk package.
* Find these terms in the document frequency table.
* If the number of documents it occurs in is more than 500 (half of the total number of documents, in the corpus) then add it to the stoplist, else do not add.
* The combination of the most frequent terms that occur in most of the documents will yield the stopwords list.

Stoplist:-

1. the
2. of
3. and
4. in
5. to
6. a
7. is
8. for
9. as
10. by
11. that
12. energy
13. are
14. with
15. on
16. from
17. be
18. or
19. was
20. power
21. it
22. an
23. at
24. ‘s
25. this
26. can
27. has
28. also
29. have
30. its

There are approximately 1 lakh terms in the corpus. So the top 30 of the terms whose frequency is the greatest and that occur in many documents can be taken as stopwords for this corpus.

The above mentioned words are the most frequent words in the corpus. Hence, they can be used as stopwords because they occur in almost every document in the corpus. The stoplist contains words like “the”, “of”, “and”, “is”, “a”, etc which occur a lot in the English language. Keeping that in mind, it would be a good choice to consider frequent terms like these as stopwords, as they occur is most of the documents and do not provide much meaning to the documents. However, if these terms are considered as stopwords in the corpus, they should also be filtered from the query.