Stop List

The following list of words can be used as stop list. These words were generated using the unigram data.

1. the:75128
2. of:44835
3. and:32428
4. a:25641
5. in:25024
6. to:19534
7. e:11735
8. is:9758
9. on:9509
10. for:8091
11. hurricane:7674
12. t:7595
13. as:7289
14. by:7160
15. from:7099
16. s:6514
17. was:6212
18. r:6109
19. with:5868
20. i:5861
21. n:5841
22. o:5288

We can set the cut-off values as words with frequency 5200 and above. This is because the word ‘Tropical’ occurs 5166 times and if ‘Tropical’ is added to the list of stop words, we will be losing the main context from the users queries.

This is because the corpus is mainly built on topic ‘Tropical Cyclone’ and removing the word tropical from the query would be a bad choice since most of the queries from users will contain the word ‘Tropical’, which is the main context of the corpus. Hence the cut-off value can be set to 5200 and above.

The above words were selected as list of stop words based on their frequency of occurrence. Based on this list we can remove these words from the query and the main context of the query will not be lost. By identifying and removing stop words we can improve the efficiency of search engine. This is because the number of words that are posted to the inverted index will reduce if we remove stop words from the query.

Stop words are mainly most common English words such as the, or, and, with, etc. These words do not contribute towards the context of the main content. So the presence or absence of these words will not affect the context of the document.