StopList

The words that can be used as stop words for the given corpus are

1. The
2. Of
3. And
4. In
5. To
6. A
7. Is
8. For
9. As
10. By
11. That
12. With
13. On
14. From
15. Or
16. Be
17. Was
18. S
19. It
20. An
21. At
22. This
23. Has
24. N
25. Can
26. Its
27. Not
28. But
29. They
30. All
31. Up
32. Had
33. If
34. No
35. So

The above list can be used as stop words as they occur in high frequency. And can be safely removed, the sentence will not be changed to a degree where it will affect the context in which the sentence is used.

Choosing the right stop word is important. I.e choosing a word just based on the frequency will not provide an appropriate stop word.

Eg. 'Power' occurs 12134 times in the corpus.

'an' occurs 11974 times in the corpus.

'it' occurs 12566 times in the corpus.

Choosing Power as a stop word will not be appropriate as power gives more meaning to a sentence than the words 'an' and 'it'.

We can have a cutoff frequency of around 10000. And cross check the output list with a common stop word list.

This will allow us to avoid having word like 'power' in our stoplist. And generate a more meaningful stoplist.