* Search looks at the document frequencies for each term
* If a term has a document frequency less than 0.1% (0.001) then it is considered “low frequency”. Otherwise, it is moved to a secondary “high frequency” list

I have taken the threshold probability of 0.001

The total no of tokens in the corpus are 3571418.

Starting with the unigram term frequency data structure, it checks for each term that if has enough occurrence in the corpus to be considered as a stopword.

In this case, if the probability of occurrence of that term is more than 0.001, it’s a stopword.

With this rule, I have total 87 stopwords.

Most of them are the common words, whereas words like “energy”, “renewable” comes in almost each document. Which means it doesn’t contain any substantial value with it. As, with respect to these terms every document is the same.