Following is the list of stop words that I have chosen from my unigram output

|  |
| --- |
| **List of Stop Words** |
| **the** |
| **of** |
| **and** |
| **in** |
| **to** |
| **a** |
| **is** |
| **for** |
| **as** |
| **by** |
| **from** |
| **on** |
| **with** |
| **retrieved** |
| **s** |
| **that** |
| **are** |
| **or** |
| **be** |
| **edit** |
| **at** |
| **an** |
| **it** |
| **this** |
| **which** |
| **has** |
| **also** |
| **1** |
| **2** |

I have chosen top 50 most frequent terms in my corpus. Among them I have checked which all terms are not frequent in all the documents. Because terms which are frequent in corpus and frequent in every document are considered as stop word.

I have taken 895 as the threshold of document frequency. That is words that are frequent and have document frequency as 895 or more , will be considered as stop word.

So the threshold of term frequency is 7751 and doc frequency is 895 because going beyond 7751 is giving my terms whose document frequency is less than 895 so this is why I am using this threshold.