Stop List

Below are the words that can be stopped “removed” from the file in a somewhat safe manner without having much effect on the information retrieval

StopList

These are the stop words that I found appropriate to remove.

* the--> 151485
* of--> 76997
* and--> 59563
* in--> 51142
* to--> 40354
* a--> 39170
* is--> 26790
* as--> 18834
* by--> 14825
* are--> 13559
* for--> 13434
* that--> 13044
* with--> 12727
* on--> 12708
* from--> 11448
* or--> 10293
* was--> 10019
* it--> 8808
* be--> 8184
* at--> 7992
* which --> 7649
* an --> 7515

From the Unigram Term Frequency

these are the words that could be stopped without removing much meaning from the text. (i.e., They can be safely removed without significantly changing the meaning of the sentence in which they are used.)

Stop List should be generated carefully, You cannot ideally remove the top “n” words from the frequency table and stop those words.

For this corpus Ideal sop list will have a frequency of 7500(and greater)

Any Frequency above 7500 will be top n words that can BE CONSIDERED for Stop List

But there are words like “edit” with higher frequency than 7500 (it has frequency of 9000)

Choosing edit to be in stop list will not be wise as “edit” is more important than “it” or “or” and its removal would be detrimental for the information retrieval.

EXPLANATION :

Consider the case you run a index on sports magazine the word “sport” or “player” might have higher frequency but it cannot be removed since it may appear more frequently in the corpus.

Best way to select the stop words are by examining the top n words in the term frequency and

**Manually looking for the words to be included to the stop list before using the top n words.**

In this case I would not involve the words like “edit, air, one” even though they might have higher frequencies.

Since they might contribute to the meaning of the sentence and their removal is not safe.