The stop list generated using unigrams is given below:

STOP LIST:

the:222965

of:127642

and:105142

in:89474

to:79651

a:69243

is:45417

for:33818

as:30026

by:25801

that:23161

~~energy:22659~~

are:22042

with:21677

from:20063

on:19534

be:18333

or:18200

~~power:15914~~

was:14700

it:13826

s:13708

at:13521

an:13464

As we can observe the stop words above are the frequent English occurring words which add no meaning (or little importance) to the context. So, a general idea would be to generate a term frequency table and sort it in decreasing order and pick a threshold till which words occur very frequently.

This will save disk space as they are more frequent and need a lot of space to save data (for ex: “the” occurs 222965 times)

Removing too many stop words would effect on the meaning of the sentence too, so the threshold should be chosen optimally.

One more interesting thing in the above list is the words “energy” and “power” which are related to the topic and has to occur frequently. In this case we should not include them as stop words, so we can cross verify with the keyword list and then add a word to stop list or manually filter these kind of words.