What are Stop Words?

By Kavita Ganesan / Text Mining Concepts

When working with text mining applications, we often hear of the term “stop words” or “stop word list” or even “stop list”. Stop words are basically a set of commonly used words in any language, not just English.

The reason why stop words are critical to many applications is that, if we remove the words that are very commonly used in a given language, we can focus on the important words instead. For example, in the context of a search engine, if your search query is *“how to develop information retrieval applications”*, If the search engine tries to find web pages that contained the terms “how”, “to” “develop”, “information”, ”retrieval”, “applications” the search engine is going to find a lot more pages that contain the terms “how”, “to” than pages that contain information about developing information retrieval applications because the terms “how” and “to” are so commonly used in the English language. If we disregard these two terms, the search engine can actually focus on retrieving pages that contain the keywords: “develop” “information” “retrieval” “applications” – which would bring up pages that are actually of interest. This is just the basic intuition for using stop words.
Stop words can be used in a whole range of tasks and here are a few:

1. Supervised machine learning – removing stop words from the feature space
2. Clustering – removing stop words prior to generating clusters
3. Information retrieval – preventing stop words from being indexed
4. Text summarization- excluding stop words from contributing to summarization scores & removing stop words when computing ROUGE scores

Types of Stop Words

Stop words are generally thought to be a “single set of words”. It really can mean different things to different applications. For example, in some applications removing all stop words right from determiners (e.g. the, a, an) to prepositions (e.g. above, across, before) to some adjectives (e.g. good, nice) can be an appropriate stop word list. To some applications however, this can be detrimental. For instance, in sentiment analysis removing adjective terms such as ‘good’ and ‘nice’ as well as negations such as ‘not’ can throw algorithms off their tracks. In such cases, one can choose to use a minimal stop list consisting of just determiners or determiners with prepositions or just coordinating conjunctions depending on the needs of the application.Examples of minimal stop word lists that you can use:

- **Determiners** – Determiners tend to mark nouns where a determiner usually will be followed by a noun
  examples: the, a, an, another
- **Coordinating conjunctions** – Coordinating conjunctions connect words, phrases, and clauses
  examples: for, an, nor, but, or, yet, so
- **Prepositions** – Prepositions express temporal or spatial relations
  examples: in, under, towards, before

In some domain specific cases, such as clinical texts, we may want a whole different set of stop words. For example, terms like “mcg” “dr” and “patient” may have less discriminating power in building intelligent applications compared to terms such as ‘heart’ ‘failure’ and ‘diabetes’.

In such cases, we can also construct domain specific stop words as opposed to using a published stop word list.

What About Stop Phrases?

Stop phrases are just like stop words just that instead of removing individual words, you exclude phrases. For example, if the phrase "good item" appears very frequently in your text but has a very low discriminating power or results in unwanted behavior in your results, one may choose to add such phrases as stop phrases. It is certainly possible to construct "stop phrases" the same way you construct stop words. For example, you can treat phrases with very low occurrence in your corpora as stop phrases. Similarly, you can consider phrases that occur in almost every document in your corpora as a stop phrase.

Published Stop Word Lists

If you want to use stop words lists that have been published here are a few that you could use:

- Snowball stop word list – this stop word list is published with the Snowball Stemmer
- Terrier stop word list – this is a pretty comprehensive stop word list published with the Terrier package.
- Minimal stop word list – this is a stop word list that I compiled consisting of determiners, coordinating conjunctions and prepositions
- Construct your own stop word list – this article basically outlines an automatic method for constructing a stop word list for your specific data set (e.g. tweets, clinical texts, etc)

Constructing Domain Specific Stop Word Lists

While it is fairly easy to use a published set of stop words, in many cases, using such stop words is completely insufficient for certain applications. For example, in clinical texts, terms like "mcg" "dr." and "patient" occur almost in every document that you come across. So, these terms may be regarded as potential stop words for clinical text mining and retrieval. Similarly, for tweets, terms like "#" "RT", "@username" can be potentially regarded as stop words. The common language specific stop word list generally DOES NOT cover such domain specific terms. Here is an article that I wrote that talks about how to construct domain specific stop word lists.

Tips for Constructing Custom Stop Word Lists

By Kavita Ganesan / Tips

Stop words are a set of commonly used words in any language. For example, in English, “the”, “is” and “and”, would easily qualify as stop words. In NLP and text mining applications, stop words are used to eliminate unimportant words, allowing applications to focus on the important words instead.

While it is fairly easy to use a published set of stop words, in many cases, using such stop words is completely insufficient for certain applications. For example, in clinical texts, terms like “mcg” “dr.” and “patient” occur almost in every document that you come across. So, these terms may be regarded as potential stop words for clinical text mining and retrieval. Similarly, for tweets, terms like “#” “RT”, “@username” can be potentially regarded as stop words. The common language specific stop word list generally DOES NOT cover such domain specific terms.

The good news is that it is actually fairly easy to construct your own domain specific stop word list. Here are a few ways of doing it assuming you have a large corpus of text from the domain of interest, you can do one or more of the following to figure out your stop words:

1. Most frequent terms as stop words

Sum the term frequencies of each unique word, w across all documents in your collection. Sort the terms in descending order of raw term frequency. You can take the top N terms to be your stop words. You can also eliminate common English words (using a publish stop list) prior to sorting so that you are sure that you target the domain specific stop words. Another option is to treat words occurring in more X% of your documents as stop words. I have personally found eliminating words that appear in 85% of documents to be effective in several text mining tasks. The benefit of this approach is that it is really easy implement, the downside however is if you have a particularly long document, the raw term frequency from just a few documents can dominate and cause the term to be at the top. One way to resolve this is to normalize the raw term frequency using a normalizer such as the document length (i.e. number of words in a given document).

2. Least frequent terms as stop words

Just as terms that are extremely frequent could be distracting terms rather than discriminating terms, terms that are extremely infrequent may also not be useful for text mining and retrieval. For example the username “@username” that occurs only once in a collection of tweets, may not be very useful. Other terms like “yoMateZ!” which could be just made-up terms by people again may not be useful for text mining applications. Note that certain terms like “yaaaaayy!!” can often be normalized to standard forms such as “yay”. However, despite all the normalization if terms still have a term frequency count of one you could remove it. This could significantly reduce your overall feature space.

3. Low IDF terms as stop words

Inverse document frequency (IDF) basically refers to the inverse fraction of documents in your collection that contains a specific term ti. Let us say you have N documents. And term ti occurred in M of the N documents. The IDF of ti is thus computed as:

$$IDF(ti)=Log\ N/M$$

So the more documents ti appears in, the lower the IDF score. This means terms that appear in each and every document will have an IDF score of 0. If you rank each ti in your collection by its IDF score in descending order, you can treat the bottom K terms with the lowest IDF scores to be your stop words. Again, you can also eliminate common English words (using a published stop list) prior to sorting so that you are sure that you target the domain specific low IDF words. This is not necessary really if your K is large enough such that it will prune both general stop words as well as domain specific stop words. You will find more information about IDFs here.

So, would stop words help my task?

So how would you know if removing domain specific stop words would be helpful in your case? Easy, test it on a subset of your data. See if whatever measure of accuracy and performance improves, stays constant or degrades. If it degrades, needless to say, don’t do it unless the degradation is negligible and you see gains in other forms such as decrease in size of model, ability to process things in memory, and etc.