A BRIEF NOTE ON STOP WORDS FOR TEXT MINING AND RETRIEVAL

Kavita Ganesan
ganesan.kavita@gmail.com

WHAT ARE STOP WORDS?

Stop words are basically a set of commonly used words in any language, not just English. The reason why stop words is 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. So, 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 more closely bring up pages that are really of interest. This is just the basic intuition for using stop words. Stop words can be used in a whole range of tasks and these are just a few:

- **Supervised machine learning** – removing stop words from the feature space
- **Clustering** – removing stop words prior to generating clusters
- **Information retrieval** – preventing stop words from being indexed
- **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.

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 to constructing a stop word list for your specific data set (e.g. tweets, clinical texts, etc)

REFERENCES

[1] http://snowball.tartarus.org/

[2] Ounis, Iadh, et al. "Terrier: A high performance and scalable information retrieval platform." *Proceedings of the OSIR Workshop*. 2006.