Corpus Specific Stop Words to Improve the Textual Analysis in Scientometrics

Vicenç Parisi Baradad[1] and Alexis-Michel Mugabushaka

[1] *Vicenc.PARISI-BARADAD@ec.europa.eu*
European Research Council Executive Agency, COV 24/161, B-1049 Brussels (Belgium)

Abstract

With the availability of vast collection of research articles on internet, textual analysis is an increasingly important technique in scientometric analysis. While the context in which it is used and the specific algorithms implemented may vary, typically any textual analysis exercise involves intensive pre-processing of input text which includes removing topically uninteresting terms (stop words). In this paper we argue that corpus specific stop words, which take into account the specificities of a collection of texts, improve textual analysis in scientometrics. We describe two relatively simple techniques to generate corpus-specific stop words; stop words lists following a Poisson distribution and keyword adjacency stop words lists. In a case study to extract keywords from scientific abstracts of research project funded by the European Research Council in the domain of Life sciences, we show that a combination of those techniques gives better recall values than standard stop words or any of the two techniques alone. The method we propose can be implemented to obtain stop words lists in an automatic way by using author provided keywords for a set of abstracts. The stop words lists generated can be updated easily by adding new texts to the training corpus.

Conference Topic

Methods and techniques

Introduction

Textual analysis -also referred to as "lexical analysis"," text mining", "co-word analysis" or "linguistic network"- has a long tradition in scientometric analysis. Earlier references can be found in the pioneering work of Eugene Garfield and others (see Garfield, 1967) studying the potential of citation analysis in information retrieval as compared to methods based on terms frequencies. Callon et al. (1983, 1986) introduced the concept of co-word analysis in science and technology studies. This technique was further developed and popularized in scientometrics by the work of Leydesdorff (1989) and researchers at the Center for Science and Technology Studies (CWTS) at the Leiden University (Noyons & van Raan, 1998).

With the availability of vast collections of research articles and better and faster computer tools, which help text analysis, the technique has firmly established itself in scientometric analysis. Nowadays it is used in various contexts: to study the thematic proximity in a collection of documents; to map scientific papers based on concept maps; to detect dynamics and trends of research based, for example, on centrality of concepts or to characterise a particular research community, by identifying relationships between the terms it uses.

While textual analytical techniques differ in degree of complexities and approaches they take, virtually all of them require relatively intensive pre-processing of the input texts. Typically, the following steps are involved in the pre-processing: (1) tokenization, (2) converting to lower case, (3) stemming and (4) removing stop words. For this last step, researchers typically use standard stop words lists obtained from texts in many different domains.

In this paper we argue that using corpus specific stop words might help the textual analysis. The paper is divided in four parts. The next section reviews briefly existing work on stop words and describes in detail two, relatively simple methods, to extract corpus specific stop words. In the subsequent, third, section we present a case study to illustrate the benefits of corpus specific stop words over more general stop words. The concluding remarks discuss limitations and point to future directions.

Related Work

When researchers in scientometrics started using textual analysis, they were standing in long tradition of information retrieval research. Early studies of word frequencies in a text or collection of documents appeared in the last century, when George K. Zip formulated an empirical law that relates terms frequencies (tf) to rank in a frequency ordered word list (Zip, 1932). This frequency characterisation was used later by Hans Luhn to obtain statistical information of words in texts and to compute a relative measure of the significance of individual words and phrases (Luhn, 1958). Using this measure Luhn hypothesized that the most discriminant words are those appearing in the middle of the frequency rank. Salton went a step further by incorporating the document frequency (df) as a measure of the discriminatory capacity of the words (Salton & Young, 1973). They suggested that words can appear in a document collection either in a random manner or concentrated in a few exemplars and they proposed the product of the term frequency times the inverse document frequency (tf • idf) as a measure of the degree of significance: the words appearing in many documents (df high) or with a low presence (tf low) are considered stop words. Based on these frequency descriptions Christopher Fox elaborated in the 90's a list containing stop words (Fox, 1990) extracted from the Brown Corpus of English literature. Although these stop words can be considered the standard or classical list and they have been frequently used, we note two limitations: first they are quite outdated and second they may be too general to take into account the specificities of a collection of texts. They may not be suitable to filter out words belonging to specific research fields or words of recent apparition. As Makrehchi & Kamel (2008) suggest, specific stop words differ from one domain to another.

Several methodologies have been proposed recently to create new stop words lists, customized to particular corpus. Among them, two proposals attracted our attention due to their relative simplicity.

On one hand, an unsupervised method to compute stop words lists arises from the study of the statistical distribution of words, by Church, K. and Gale, W. (1995) and their hypothesis that common stop words follow a Poisson distribution. This has been used to create a stop word list for particular Polish texts (Jungiewicz & Lopuszyński, 2014). We call this approach the *Poisson stoplist.*

Under this hypothesis one assumes that the document frequency of words (df) in a corpus can be estimated (dfe) from their term frequency (tf) and the total number of documents (N) by using the probability theory:

$$\frac{dfe}{N} = 1 - P(0),$$

where $P(0)$ is the probability of not appearing the word. Assuming a Poisson distribution for stop words, the probability of k instances of a word is given by:

$$P(k,\mu) = \frac{e^{-\mu} * \mu^k}{k!},$$

where μ is the average number of instances per document:

$$\mu = \frac{tf}{N}$$

The relation dfe/df is supposed to be close to 1 for randomly distributed terms (stop words) and shows an increase for highly cluttered terms (keywords); although this depends on the corpus, as Jungiewicz and Lopuszyński found when computing their stop word lists for legal texts from the public procurement domain. They realised that their most common stop words had a high variability in their distribution and replaced the Poisson assumption with a negative binomial distribution, which allows a larger variance.

On the other hand, S. Rose et al. (2010) proposed an unsupervised, domain and language independent method to extract keywords from individual texts called RAKE (Rapid Automatic Keyword Extraction) and a supervised method to elaborate stop word lists based on the intuition that words adjacent to keywords tend to be stop words.

RAKE uses stop words to parse the text and extract candidate key phrases (consisting in one or more words). The key phrases are then scored by computing word co-ocurrences and using a metric that favours words belonging to long key phrases. The top T candidates are chosen as keywords (key phrases).

The method proposed by S. Rose to extract stop words from a corpus resorts on accumulating for each word its 'adjacency frequency' (af) and 'keyword frequency' (kf), together with the term frequency (tf) and document frequency (df). Then, given a selection threshold n, the most frequent words with af > kf are chosen as stop words. This method is called by the author *keyword adjacency stoplist* (because it includes primarily words that are adjacent to and not within keywords: Rose et al. 2010, p. 14). We refer to this method as *RAKE stoplist* in this paper.

Case Study: stop list for a collection of abstracts of funded projects

To study the suitability of the above described methodologies and create our own stop words list we applied them to a corpus from abstracts of projects, funded by the European Research Council, in the Life Sciences domain. This corpus consists of 1579 projects covering diverse research areas. The table 1, shows the number of project abstracts by each research area (which corresponds to the scientific panel in which the project was evaluated).

Table 1. Overview of the corpus of abstracts used in the case study

Scientific areas	abstracts	%
Molecular and structural biology and biochemistry	176	11.1
Genetics, genomics, bioinformatics and systems biology	178	11.1
Cellular and developmental biology	164	10.4
Physiology, pathophysiology and endocrinology	176	11.15
Neurosciences and neural disorders	217	13.7
Immunity and infection	168	10.6
Diagnostic tools, therapies and public health	209	13.2
Evolutionary, population and environmental biology	168	10.6
Applied life sciences and biotechnology	115	7.3

Creating stop words

We randomly chose 80% of the abstracts as a training set and the other 20% as a test set.

Following the algorithms outlined in Rose et al. 2010, we wrote a program in Python to create a table (which we call Frequency table) with all the words (12621 in total) of the training set that contains the words, term frequencies (tf), document frequencies (df), keyword frequencies (kf) and adjacent frequencies (af).

This table was used to create both the *Poisson stoplist* and the *RAKE stoplist*. For the later, we set various thresholds to obtain the top n words with the highest term frequency.

Evaluating the stopwords

To evaluate if the corpus-specific stop words improve textual analysis, we use them in extracting keywords. We compare the keywords extracted using those stop words with author-provided keywords. The idea is that, depending on the stop words used, the keywords extracted will match more or less the ones provided by the authors and the higher the share of matched keywords the better the stop words list.

It should be noted that author-provided keywords do not necessary contain words which also appears in the abstracts. In our corpus, out of 7845 keywords given by the authors only 3494 (44.5 %) where encountered in the abstracts. This means that the precision and F-measure need to be taken into account with care and thus we have not used them for the evaluation of the quality of the stop words list, resorting only to the recall measure, computed as the relation between the total number of correct extracted keywords and the total number of keywords given by the authors, that appear in the abstracts.

We compared the keywords provided by authors with the keywords extracted using the following lists of stop words

1. Standard *Fox stop words list*
2. Stop words list created using the Poisson distribution hypothesis (*Poisson stoplist*)
3. Stop words list computed using keyword adjacency (*RAKE stoplist*)
4. Stop words lists computed using combinations of *Fox,Poisson* and *RAKE*

For keywords extraction we used a Python implementation of the RAKE algorithm (https://github.com/aneesha/RAKE)

1. *Fox stoplist*

This list serves as a baseline for our work and the computation of the recall of the keywords extracted using RAKE algorithm does not need to tune any parameter. The recall obtained is 56.42%.

2. *Poisson stoplist*

To extract the stop words using this approach we need first to set the threshold for the relation dfe/df. To do that we computed the mean and standard deviation of the dfe/df for all the *Fox* stop words that appear in the training set. Figure 1 shows the plot of these values, where the mean (dfe/df) + std(dfe/df) is 1.55. There are only 14 Fox stop words excluded from the list and apart from the words (*ordering, right and small*) their term frequency is very low. We have used this threshold to obtain the stop words from our training data appearing in at least 10 documents (df>10) and we have obtained a list of 2008 words that gives a recall of 58.25% in our test set, which is better than the *Fox stoplist*.

3. *RAKE stoplist*

To use the RAKE approach we extracted all the words from the training set with af>kf and created an ordered table, sorted in descending order of word occurrence (tf). This table consisted in a list of 2045 candidate stop words. To choose the top best frequency rank we tested subsets of these lists and computed their recall values. The result obtained using all the words in the list was 45.42 % of recall and the results improved by removing words from the list, having a peak at a 53.31% of recall, when using the first 185 words of the rank.

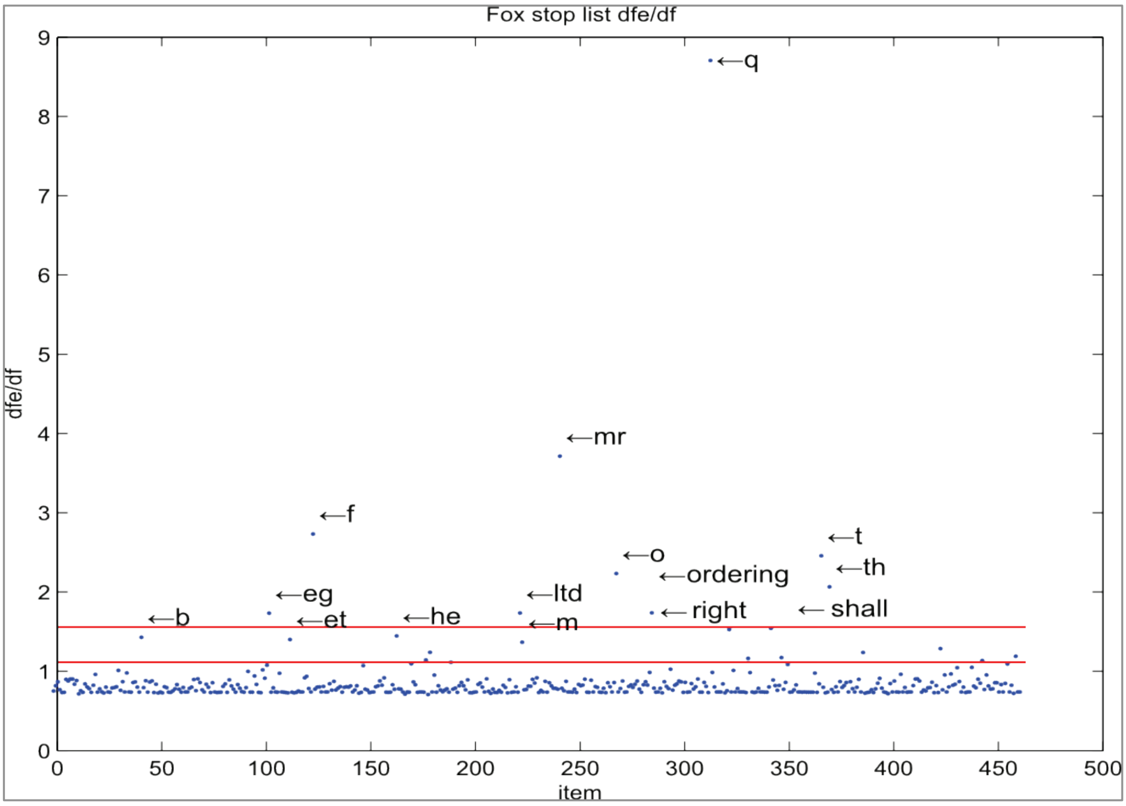

Figure 1. Fox stoplist dfe/df values found in the training corpus. Just a few words are above the standard deviation limit, and they are rarely found (tf very low).

4. *Combinations Poisson and RAKE*

Since the *RAKE stoplist* gave us worse results than the *Fox stoplist*, we tried to combine them with the Poisson approach (*RAKE-Poisson*) and we extracted the words with df>10,af>kf and dfe/df<1.55. This improved the previous results, giving a recall of 62.34%. Note that the condition dfe/df> r can also be seen as an adaptive threshold on tf, since, under the Poisson distribution, it can also be expressed as:

$$tf > N * \ln\left(\left(\frac{df}{N}\right)r - 1\right),$$

and instead of choosing a minimum common tf for all the words, we adapt the tf to each word's df. In Figure 2 we have plotted the df of the RAKE stop words (tf>0) , together with the Fox stop words found in the *RAKE stoplist*. Also we plotted the dfe/1.55 curve which shows the limit above which the words belong to *RAKE-Poisson stoplist*.

After inspecting the frequencies of the RAKE-Poisson stop words we found words expected to appear in Life Sciences texts and we questioned ourselves if their removal from the stoplist would improve the recall results. To check it we removed them by hand and the recall increased to 64.56 %. A more detailed inspection of the stoplist frequencies allowed us to see that just a few words (6 in total) belong to the life sciences domain (genetic, disease, protein, molecular, gene, cell), all them with kf>60 had a 1.1<dfe/df <1.55. In all them the af/kf relation was less than 5 (af/kf<5). This data gave us the intuition that we needed to decrease the dfe/df threshold and also to be more strict on the af/kf condition, so we tested a stoplist consisting in the RAKE intersection with Poisson stoplist (df>10 and af>5*kf and dfe/df<1.2) which gave a recall of 68.69%, being this the best result. We call it the *RAKEm-Poisson stoplist*.

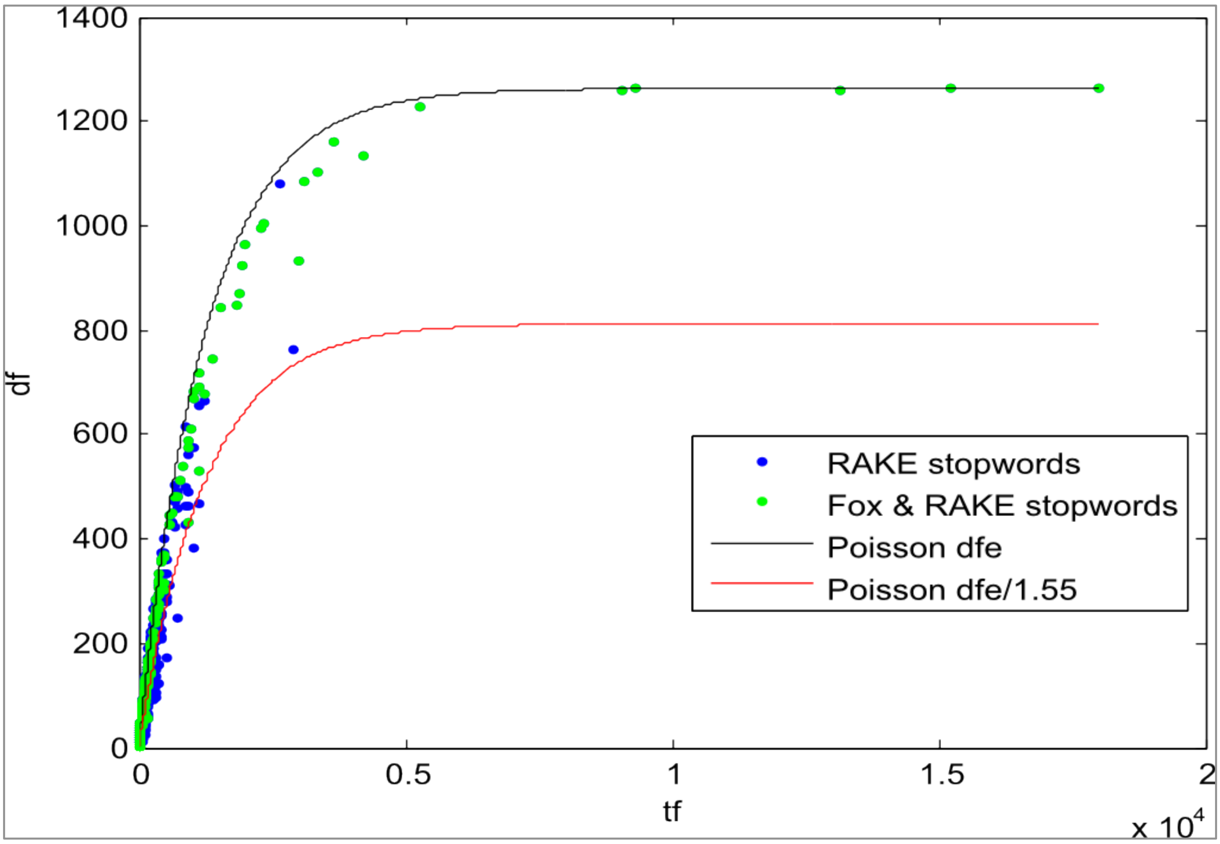

Figure 2. RAKE and Fox stop words. We can see that the Fox stop words follow the Poisson distribution better than the RAKE stop words, which appear more concentrated at low df values.

Conclusion

Our aim is to obtain stop words that help to provide meaningful and significant keywords that summarize the texts; the validation of the stoplists we did was based using the author given key phrases which most of the times had fewer words than the ones obtained using RAKE. We think that this circumstance is favouring standard stoplists since they will still produce single word keywords given by authors and end up yielding overall recall values similar to specific domain stoplists. Therefore we plan as a future work to use measures that evaluate semantic value of the key phrases.

We would like to remark that the RAKE-Poisson stoplist can be obtained from the word frequencies and the author keywords, without further human intervention. Our future work involves also the automatization of the computation of the best af/kf and dfe/df thresholds to generate the *RAKEm-Poisson stoplists*.

References

Blanchard, A. (2007). Understanding and customizing stopword lists for enhanced patent mapping, *World Patent Information, 29*(4), 308-316.

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: an introduction to co-word analysis. *Social Science Information, 22*, 191-235.

Callon, M., Law, J., & Rip, A. (Eds.). (1986). *Mapping the Dynamics of Science and Technology*. London: Macmillan.

Church, K. and Gale, W. (1995). Poisson mixtures. *Journal of Natural Language Engineering*, *2*, 163-190.

Fox, C. (1990). A stop list for general text. *ACM-SIGIR Forum*, *24*, 19-35.

Garfield, E. (1967). Primordial concepts, citation indexing and eIistorio-bibliography. *The Journal of Library History, 2*(3), 235-249. http://www.garfield.library.upenn.edu/essays/v6p518y1983.pdf

Jungiewicz, M. & Lopuszyński, M. (2014). Unsupervised keyword extraction from Polish legal texts. *Advances in Natural Language Processing*, 65–70. Springer LNCS.

Leydesdorff, L. (1989). Words and co-words as indicators of intellectual organization. *Research Policy*, 18(4), 209-223.

Luhn, P. (1958). The automatic creation of literature abstracts, *IBM Journal of Research and Development, 2*(2) 159–165

Makrehchi, M. & Kamel, M. (2008) *Automatic Extraction of Domain-specific Stopwords from Labeled Documents*. Berlin / Heidelberg: Springer.

Noyons E. C. M. & Raan A. F. J. (1998). Monitoring scientific developments from a dynamic perspective: Self-organized structuring to map neural network research. *Journal of the American Society for Information Science*, *49*(1), 68–81.

Rose, S., Engel, D., Cramer, N. & W. Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory*. John Wiley & Sons, Ltd.

Salton, G. & Yang, S. (1973). On the specification of term values in automatic indexing, *Journal of Documentation,* 29(4), 351–372.

Sinka, M.P. & Corne, D.W. (2003). Towards modernised and web-specific stoplists for web document analysis, *IEEE/WIC International Conference on Web Intelligence,* 396-402.

Zipf, K. (1932). S*elective Studies and the Principle of Relative Frequency in Language*. Cambridge: MIT Press.