Creating Stopword Lists for Historical Languages

Patrick J. Burns

April 26, 2017

Proposal

Stopwords are words that are filtered from documents prior to text analysis tasks, usually words that are either high frequency or semantically non-selective.[1] Through the removal of such words, text analysis tasks, including supervised machine learning, clustering, information retrieval, and text summarization,[2] benefit in areas like noise reduction, feature reduction, or speed optimization. There are stopwords lists available online for Latin and Greek, for example via the *Perseus Project* and *stopwords-json*[3] but these lists offer little documentation about their sources or creation. Moreover, since the Perseus list is not published as a self-contained dataset, it does not provide systematic version control or persistent identifiers for proper citation.

I propose to use the time at the Global Philology *Big Corpora of Historical Text* seminar at Universität Leipzig on July 10-11 for developing best practices for stopword list creation for Latin, Greek, and other historical languages. I will review recent publications on stopword list creation in other languages,[4] replicate corpus-based experiments based on this literature, and work with seminar participants to arrive at the appropriate methods for developing similar lists in our target languages.

The outcome of the seminar would be an article presenting results of stopword list creation methods for Latin as well as a white paper suggesting the best path forward for other languages represented in the Global Philology project. These results, i.e. the stopword list istelf, would also be "published" in 1. the Classical Language Toolkit,[5] 2. a branch of the natural language processing tool spaCy that I am developing for the Latin language,[6] and 3. other open-source venues such as *stopwords-json* and *stop-words*. Furthermore, the stop word dataset would be available for use under a CC0 for use in Open Greek and Latin and related projects.[7] By ensuring that the Latin stopword list is properly documented, version controlled, and citable, I am for the RAD paradigm with this dataset, that is it will be replicable, aggregable, and data-driven,[8] and as such will be better suited to be included in other text analysis projects.

[1] Manning et al. 2008, 26.

[2] See, for example, http://text-analytics101.rxnlp.com/2014/10/all-about-stop-words-for-text-mining.html.

[3] See http://www.perseus.tufts.edu/hopper/stopwords; https://github.com/6/stopwords-json/blob/master/dist/la.json. The *Tesserae Project* (http://tesserae.caset.buffalo.edu/) uses a default stop list based on corpus frequency. Note too here that elsewhere there is a desideratum for Latin and Greek stop words, e.g. https://pypi.python.org/pypi/stop-words.

[4] E.g. Lo, He, and Ounis 2005; Zou et al. 2006; Alajmi, Saad, and Darwish 2012; Daowadung and Chen 2012; Raulji and Saini 2016.

[5] http://cltk.org/.

[6] https://spacy.io/. I am currently developing a Latin parameter set for spaCy, which is available on Github here: https://github.com/diyclassics/spaCy/tree/latin/spacy/la/.

[7] https://wiki.creativecommons.org/wiki/CC0_use_for_data.

[8] Haswell 2005.

References

Alajmi, A., E. M. Saad, and R. R. Darwish. 2012. "Toward an Arabic Stop-Words List Generation." *International Journal of Computer Applications* 46 (8): 8–13. https://pdfs.semanticscholar.org/eff7/4e0e013679251909324679f441af4ff7bedf.pdf.

Daowadung, P., and Y. H. Chen. 2012. "Stop Word in Readability Assessment of Thai Text." In *2012 IEEE 12th International Conference on Advanced Learning Technologies*, 497–99. doi:10.1109/ICALT.2012.9.

Haswell, R. H. 2005. "NCTE/CCCC's Recent War on Scholarship." *Written Communication* 22 (2): 198–223. doi:10.1177/0741088305275367. http://journals.sagepub.com/doi/abs/10.1177/0741088305275367.

Lo, R. T., B. He, and I. Ounis. 2005. "Automatically Building a Stopword List for an Information Retrieval System." In *5th Dutch-Belgium Information Retrieval Workshop*. Utrecht, The Netherlands. http://terrierteam.dcs.gla.ac.uk/publications/rtlo_DIRpaper.pdf.

Manning, C. D., P. Raghavan, H. Schütze, and others. 2008. *Introduction to Information Retrieval*. Vol. 1. 1. Cambridge: Cambridge University Press.

Raulji, J. K., and J. R. Saini. 2016. "Stop-Word Removal Algorithm and Its Implementation for Sanskrit Language." *International Journal of Computer Applications* 150 (2). http://www.ijcaonline.org/archives/volume150/number2/raulji-2016-ijca-911462.pdf.

Zou, F., F. L. Wang, X. Deng, S. Han, and L. S. Wang. 2006. "Automatic Construction of Chinese Stop Word List." In *Proceedings of the 5th WSEAS International Conference on Applied Computer Science*, 1010–5. https://pdfs.semanticscholar.org/c543/8e216071f6180c228cc557fb1d3c77edb3a3.pdf.