xtas 3, the eXtensible Text Analysis Suite

Lars Buitinck
Netherlands eScience Center
Maarten de Rijke
Informatics Institute, University of Amsterdam

xtas, the eXtensible Text Analysis Suite, version 3,[1] is a collection of natural language processing and text mining tools, brought together in a single software package with built-in distributed computing and support for the Elasticsearch document store. Version 3 of xtas is a redesign based on the lessons learned from earlier versions (De Rooij, Vishneuski and De Rijke 2012), especially with regard to ease of use and ease of installation. xtas is open source software, distributed under the terms of the Apache License.

Compared to competing text analysis toolkits such as GATE (Cunningham et al. 2013) and NLTK (Bird 2006), xtas promotes simplicity and scaling. A basic installation works like a Python module that can be used interactively, even by inexperienced programmers such as students. Built-in package management and a simple interface take away the hassle of installing, configuring and using the multitude of existing libraries and applications. Inputs to xtas modules are free text, while outputs are designed to be human-readable and easy to process with simple tools, rather than relying on custom (XML or other) formats.

Distributed computing is supported by means of the RabbitMQ/Celery job queuing middleware, so that large sets of documents can be processed on clusters of machines. Scaling up to distributed computing is a gradual process, requiring minimal changes to scripts. A REST (web service) interface permits users of other programming languages to access xtas functionality, submit jobs and retrieve text analysis results.

By connecting to Elasticsearch, currently the go-to solution for storage of, search in, and analytics on large document collections, xtas enables semantic search, e.g., querying for named entities or temporal expressions in addition to full-text search. xtas functionality consists partly of wrappers for existing packages, with automatic installation of software and data; and partly of custom-built modules coming out of research. Currently offered are various parsers for Dutch and English (Alpino, CoreNLP, Frog, Semafor), named entity recognizers (Frog, Stanford and custom-built ones), a temporal expression tagger (Heideltime) and a sentiment tagger based on SentiWords.

Finally, xtas's open architecture makes it possible to include custom code, run this in a distributed fashion and have it communicate with Elasticsearch to provide document storage and retrieval.

References

Bird, S. (2006) ‘NLTK: the natural language toolkit’, *Proc. COLING/ACL Interactive presentation sessions*, pp. 69–72.

Cunningham, H. et al. (2013) ‘Getting more out of biomedical documents with GATE's full lifecycle open source text analytics’, *PLoS Comput. Biol.* 9.2.

De Rooij, O., Vishneuski, A. and De Rijke, M. (2012) ‘xTAS: Text Analysis in a Timely Manner’, *DIR*.

[1] http://xtas.net