xTAS and ThemeStreams

Extendable Text Analysis Service and its Usage in a Topic Monitoring Tool

Ork de Rooij
University of Amsterdam
Science Park 904
1098 XH Amsterdam, The Netherlands
orooij@uva.nl

Tom Kenter
University of Amsterdam
Science Park 904
1098 XH Amsterdam, The Netherlands
tom.kenter@uva.nl

Maarten de Rijke
University of Amsterdam
Science Park 904
1098 XH Amsterdam, The Netherlands
derijke@uva.nl

ABSTRACT

xTAS is an extendable multi-user text analysis service for large scale multi-lingual document analysis developed at the University of Amsterdam. It can process large amounts of documents in a timely manner through a web interface that can be used by multiple users at once. In this demonstration paper we present recent additions which include semanticization, on the fly TF-IDF model generation and on the fly co-occurrence metrics. Furthermore, we demonstrate ThemeStreams, a novel topic monitoring tool built on top of xTAS.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text Analysis

General Terms

Algorithms, Performance, Experimentation

Keywords

text analysis, web service, distributed processing, microblog visualization

1. INTRODUCTION

xTAS[1] is an integrated set of text analysis services for processing documents in a timely manner. It is available through a web API that can be used by multiple users at once. xTAS includes tools for stemming, tokenization, named entity recognition, part–of–speech tagging, sentiment analysis and various types of aggregation on top of this. The purpose of xTAS is to run text processing tasks as fast as possible, without concerning users about databases, storage or result caching.

The software can run multiple tasks in parallel, possibly on different machines (nodes). xTAS is built solely with open source software. It uses Celery [2] to distribute tasks between nodes. By default MongoDB [4] is used to store documents and results though other options are available as well.

The software is extendable. Additional functionality can easily be added through a plugin architecture.

In what follows we describe recent additions to xTAS and we present ThemeStreams, a novel topic monitoring tool built on top of xTAS.

2. XTAS

Recent additions and improvements to xTAS include:

- Semanticization[2]

  xTAS can semantically enrich texts by linking entities mentioned in it to their Wikipedia article.

- On the fly TF-IDF model generation and application

  TF-IDF models based on a user selected series of documents can be trained on the fly. The models can be used to provide TF-IDF statistics for words in new documents.

- Co-occurrence metric calculation

  A variety of co-occurrence metric calculation methods were added to xTAS, including maximum likelihood estimate, point wise mutual information, log likelihood ratio and χ^2. This enables users to calculate the co-occurrence of entities in a set of documents.

- Automatic language identification

  If the language of a document is not supplied xTAS can automatically determine it. Currently this is implemented by using TextCat [6].

- Support for multiple document stores

  Besides mongoDB [4], xTAS can communicate directly with Apache Solr [1] or ElasticSearch [3]. These stores can be used as a document repository as well as a result cache.

- Response time improvements

  Analysis of xTAS usage over time shows that named entity recognition is a frequently requested and time consuming analysis. In order to keep response times to

[1] See http://xtas.net

[2] Semanticization, the process of linking mentions of concepts in a text to the articles in an external knowledge base they denote, is also referred to as entity linking or Wikification.

DIR 2013, April 26, 2013, Delft, The Netherlands.
.

near-real time speeds xTAS keeps several NER models (for all supported languages) in memory on each xTAS node.

3. THEMESTREAMS

ThemeStreams[3] is a visual interface that helps answer the question *"Who is talking about what?"*. It does so for topics in the Dutch political landscape by showing the ebb and flow of conversations about particular themes trough time. While there are many topic monitoring tools available, the novelty of ThemeStreams lies in its ability to present the user with a quick overview of the relative frequency of posts a particular group of users issued on a certain subject. ThemeStreams is based on tweets posted on Twitter by four groups of people:

- politicians (ministers, members of parliament, but also the local ranks of politicians in municipalities and provinces)
- political journalists (news paper journalists as well as talk show hosts of political television shows)
- lobbyists (people pushing the people who are active in politics)
- other influencers (these include (satirical) columnists, politically engaged celebrities and stand-up comedians)

The harvesting of these tweets started late 2011. At the time of writing, we follow about 1400 individual users, who, together with all people participating in conversations with these *inner circle* users yield a set of just over 3.9M tweets.

The interactive visual interface is aimed at giving insight into the ownership and dynamics of themes being discussed. It enables users to answer questions such as *Who put this issue on the map?*, *Who picked up on this topic?*, *Is this topic gaining momentum?* ThemeStreams allows users to explore streams of tweets either from a fixed set of predefined themes or through a search box. It uses stream graphs [5] to indicate how the four influence groups discuss a specified theme, thereby depicting the volume, the "aliveness" and ownership of a topic.

The interface indicates the time a tweet was posted, the influence group the poster belongs to and the number of people which reacted to a statement (which can be used to estimate the "size" and "lifetime" of statement). Initially a combined word cloud is shown with words colorized by the group they originate from. Users can zoom in to parts of the stream for more detail. Doing so results in individual word clouds being displayed per influence groups during the selected period.

Initial usability studies were carried out with university staff members and media analysts working for a communication agent. We found that ThemeStreams was intuitive to understand and it was easy to inspect parts of a tweet stream in detail. The combined clouds proved to be insightful for a fast overview of data. The individual clouds proved to be useful for inspecting relative word usage between groups. We also found a need for depicting the most represented speakers within a group.

4. FUTURE WORK

xTAS is actively being used in a number of research and production environments. As such, work on xTAS is ongoing and features are being deployed in close collaboration with end users. Currently, we focus on adding support for temporal tagging and for easier deployment on large clusters.

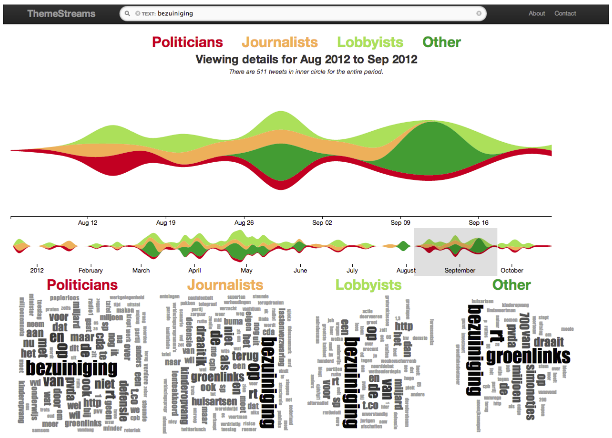

Figure 1: ThemeStreams - A visual interface that answers the question *Who is talking about what?*. Tweets are shown in a stream graph, categorized by their authors and weighted by their conversational influence. Parts of the stream can be selected and detailed word clouds per group pop up to show what was being said by whom during that period in time.

A more detailed user study of ThemeStreams is currently in progress. Also we are looking into additional application scenarios for ThemeStreams, like discourse analysis over time in other domains such as news paper archives.

5. ACKNOWLEDGEMENTS

This research was partially supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreements nr 258191 (PROMISE Network of Excellence) and 288024 (LiMoSINe project), the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.061.814, 612.061.815, 640.004.802, 727.011.005, 612.001.116, HOR-11-10, the Center for Creation, Content and Technology (CCCT), the BILAND project funded by the CLARIN-nl program, the Dutch national program COMMIT, the ESF Research Network Program ELIAS, the Elite Network Shifts project funded by the Royal Dutch Academy of Sciences (KNAW), and the Netherlands eScience Center under project number 027.012.105.

6. REFERENCES

[1] Apache solr. http://lucene.apache.org/solr/.

[2] Celery: Distributed Task Queue. http://celeryproject.org/.

[3] elasticsearch. http://www.elasticsearch.org/.

[4] MongoDB. http://www.mongodb.org/.

[5] L. Byron and M. Wattenberg. Stacked graphs–geometry & aesthetics. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1245–1252, 2008.

[6] W. B. Cavnar, J. M. Trenkle, et al. N-gram-based text categorization. *Ann Arbor MI*, 48113(2):161–175, 1994.

[3] See an online demo at http://themestreams.xtas.net/