CATS: Collection and Analysis of Tweets Made Simple

Ciprian-Octavian Truica
Computer Science and Engineering Department
Faculty of Automatic Control and Computers
University Politehnica of Bucharest
Splaiul Independentei, Nr. 313, Sector 6
Bucharest, Romania
ciprian.truica@cs.pub.ro

Adrien Guille
Data Warehousing, Engineering and Mining Team (ERIC)
Université Lumière Lyon 2
5 avenue Pierre Mendes France
Bron, France
adrien.guille@univ-lyon2.fr (corresponding author)

Michael Gauthier
Center for Research in Terminology and Translation (CRTT)
Université Lumière Lyon 2
86 rue Pasteur
Lyon, France
michael.gauthier@univ-lyon2.fr

CSCW '16 Companion, February 27 - March 02, 2016, San Francisco, CA, USA
ACM 978-1-4503-3950-6/16/02. http://dx.doi.org/10.1145/2818052.2874320

Abstract

Twitter presents an unparalleled opportunity for researchers from various fields to gather valuable and genuine textual data from millions of people. However, the collection process, as well as the analysis of these data require different kinds of skills (*e.g.* programing, data mining) which can be an obstacle for people who do not have this background.
In this paper we present CATS, an open source, scalable, Web application designed to support researchers who want to carry out studies based on tweets. The purpose of CATS is twofold: (i) allow people to collect tweets (ii) enable them to analyze these tweets thanks to efficient tools (*e.g.* event detection, named-entity recognition, topic modeling, word-clouds). What is more, CATS relies on a distributed implementation which can deal with massive data streams.

Author Keywords

Twitter; Data Collection; Data Analysis; Web application

ACM Classification Keywords

Information systems [Data mining, Crowdsourcing]

Introduction

Twitter is a popular micro-blogging service available in most of the world, which allows users to publish short 140-character messages, *i.e.* tweets. Through these tweets, people share, discuss and forward information about vari-

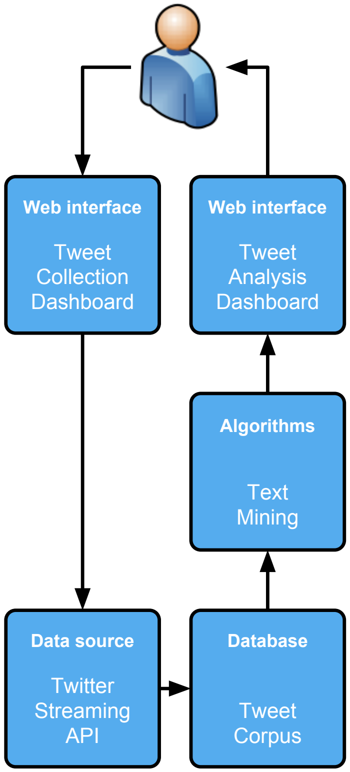


Figure 1: Researchers interact with the tweet collection dashboard to create corpora which suit their needs; CATS then collects data via the Twitter streaming API and stores them into distinct databases (*i.e.* one per corpus). Eventually, through the tweet analysis dashboard, researchers can explore these corpora in details with the help of text mining and visualization tools.

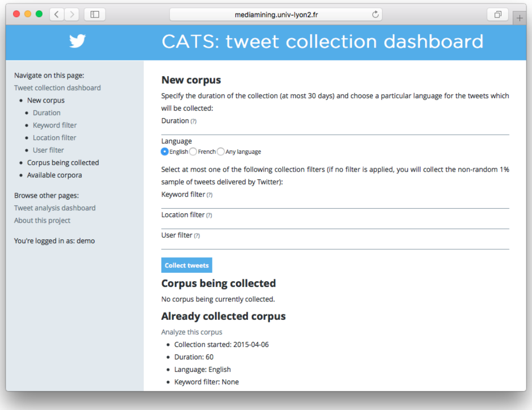
ous kinds of events in real-time. On the other hand, Twitter provides an unprecedented opportunity for researchers from numerous fields (*e.g.* health, linguistics, politics) to collect valuable textual data on a continuous basis. Collecting a large amount of tweets, storing them and analyzing them efficiently requires both programing and data mining skills. This is a problem for a lot of researchers who do not have those skills, nor the resources to work with computer scientists, and this is especially true in social sciences. Thus, there is a need for a tool that would allow these researchers to easily carry out studies based on tweets. This tool should be (i) free, (ii) easy to use, and (iii) able to handle large amounts of data. Even though there are already several tools to tackle this issue, none of them combines the three aforementioned requirements.

We address this issue with regards to these three points by proposing CATS (Collection and Analysis of Tweets made Simple), which is an (i) open source[1][2], (ii) Web based and (iii) distributed application for easily collecting and analyzing large scale corpora of tweets. The purpose of CATS is to enable any researcher to create a corpus of tweets with specific demands, even without any programing background. Not only does CATS allow researchers to collect tweets, it also enables them to analyze these tweets efficiently thanks to advanced text mining techniques and adapted visualizations.

Proposed Tool

CATS is meant to be an evolutive application. It will be continuously updated with the latest corpus analysis tools developed by the labs involved in the project. In this section, we present the current features, the implementation and the performance of CATS. Figure 1 gives an overview of the way CATS works.

[1] Sources: https://github.com/CATS-Project/CATS
[2] Wiki: https://github.com/CATS-Project/CATS/wiki



CATS: tweet collection dashboard

New corpus

Specify the duration of the collection (at most 30 days) and choose a particular language for the tweets which will be collected:

Duration (?)

Language
● English ○ French ○ Any language

Select at most one of the following collection filters (if no filter is applied, you will collect the non-random 1% sample of tweets delivered by Twitter):

Keyword filter (?)

Location filter (?)

User filter (?)

Corpus being collected

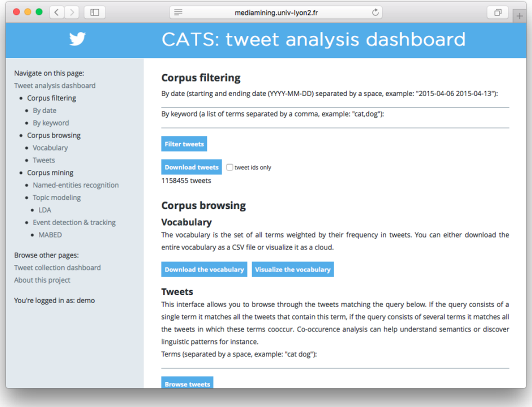
No corpus being currently collected.

Already collected corpus

- Collection started: 2015-04-06
- Duration: 60
- Language: English
- Keyword filter: None

You're logged in as: demo

Figure 2: The collection dashboard allows the creation of a new corpus. It also lists corpora which are being collected or have already been collected.



CATS: tweet analysis dashboard

Corpus filtering

By date (starting and ending date (YYYY-MM-DD) separated by a space, example: "2015-04-06 2015-04-13"):

By keyword (a list of terms separated by a comma, example: "cat,dog"):

□ tweet ids only

1158455 tweets

Corpus browsing

Vocabulary

The vocabulary is the set of all terms weighted by their frequency in tweets. You can either download the entire vocabulary as a CSV file or visualize it as a cloud.

Tweets

This interface allows you to browse through the tweets matching the query below. If the query consists of a single term it matches all the tweets that contain this term, if the query consists of several terms it matches all the tweets in which these terms cooccur. Co-occurence analysis can help understand semantics or discover linguistic patterns for instance.

Terms (separated by a space, example: "cat dog"):

You're logged in as: demo

Figure 3: The analysis dashboard allows filtering, browsing and mining a corpus of tweets.

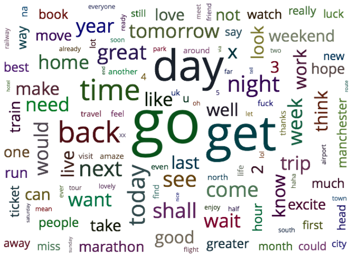
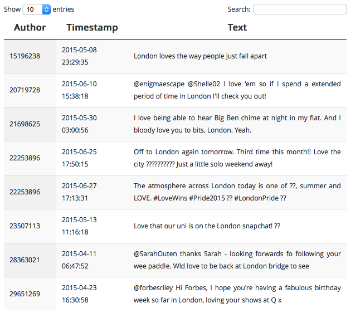


Figure 4: Vocabulary for tweets containing “London”.


Show 10 entries Search:

Author	Timestamp	Text
15196238	2015-05-08 23:29:35	London loves the way people just fall apart
20719728	2015-06-10 15:38:18	@enigmaescape @Shelle02 I love 'em so if I spend a extended period of time in London I'll check you out!
21698625	2015-05-30 03:00:56	I love being able to hear Big Ben chime at night in my flat. And I bloody love you to bits, London. Yeah.
22253896	2015-06-25 17:50:15	Off to London again tomorrow. Third time this month!! Love the city ?????????? Just a little solo weekend away!
22253896	2015-06-27 17:13:31	The atmosphere across London today is one of ??, summer and LOVE. #LoveWins #Pride2015 ?? #LondonPride ??
23507113	2015-05-13 11:16:18	Love that our uni is on the London snapchat! ??
28363021	2015-04-11 06:47:52	@SarahOuten thanks Sarah - looking forwards fo following your wee paddle. Wld love to be back at London bridge to see
29651269	2015-04-23 16:30:58	@forbesriley Hi Forbes, I hope you're having a fabulous birthday week so far in London, loving your shows at Q x

Figure 5: Browsing the tweets in which “‘London” and “love” co-occur.

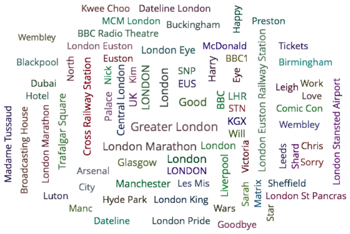


Figure 6: Named-entities mentioned in tweets containing “London”.

Features

Tweet collection. Researchers registered on CATS have their own private workspace. They can collect a corpus of tweets (see Fig. 2) during the period of their choice (*e.g.* a week, a month), using one of four collection methods: tweets matching one or more keywords; tweets published by a set of Twitter accounts; tweets geo-tagged within bounding boxes (*i.e.* pairs of coordinates); the non-random 1% sample of tweets delivered by Twitter. In all cases, researchers can specify a particular language for the tweets which will be collected.

Tweet analysis. Researchers can analyze a corpus at any time thanks to the analysis dashboard (see Fig. 3) which is updated daily with the latest tweets collected. It features three modules: (i) corpus filtering, (ii) browsing and (iii) mining.

The filtering module enables researchers to define sub-corpora based on the content or publication date of tweets. Thus, CATS users can analyze and compare results obtained through the browsing and mining modules on specific parts of the corpus.

The browsing module is composed of two functions. First, the vocabulary browser designed to provide an overall view of the set of tweets the researcher wants to analyze. The vocabulary is the list of all the words present in the corpus ordered according to their frequency after lemmatization. It can either be downloaded or visualized in CATS as an interactive word-cloud (see Fig. 4) connected to the tweet browser.

Secondly, the tweet browser provides a detailed view of specific words in context by displaying all the tweets in which they appear (see Fig. 5). It can either be accessed by manually specifying words or by clicking on a word in the aforementioned word-cloud.

The mining module is composed of three sections: named-entity recognition, topic modeling and event detection.

Named-entity recognition consists in automatically locating names of people, organizations and locations in tweets [1]. It is possible to interact with the named-entities in the same two ways as with the vocabulary (see Fig. 6).

Topic modeling is a way of automatically discovering hidden themes pervading a collection of tweets. Topic models can help organizing, understanding, and summarizing large amounts of tweets. As of now, one topic model is available in CATS, namely Latent Dirichlet Allocation (LDA) [2].

Event detection consists in automatically identifying and describing (temporally and textually speaking) events reported in tweets. Currently, CATS relies on Mention-anomaly-based Event Detection (MABED) [6]. The results of both the topic modeling and event detection algorithms are displayed in the same way: *i.e.* an ordered list of topics/events, each topic/event being described by a weighted list of the most meaningful words.

It is possible to download any corpus or sub-corpus of tweets, as well as any kind of data generated with the aforementioned functions (*e.g.* vocabulary, named-entities, topics, events) at any moment in a CSV file, so that CATS users can use other tools to analyze their corpora.

Implementation

CATS is a Web based, parallelized and distributed application written in Python. It collects tweets through the Twitter streaming API and stores them via a NoSQL, document-oriented database management system, namely MongoDB [3], so that researchers have their own database. CATS scales horizontally by adding more nodes (*i.e.* MongoDB instances), using either a shared everything architecture on a single server or a shared nothing architecture on a cluster of servers.

In addition to the meta-data provided by Twitter about each

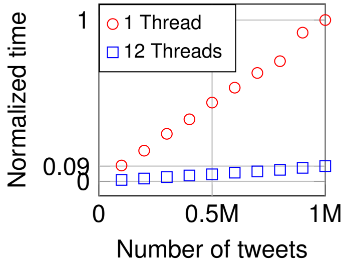


Figure 7: Duration of the meta-data extraction versus number of tweets.

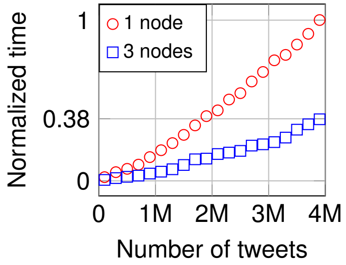


Figure 8: Vocabulary extraction runtime comparison versus number of tweets.

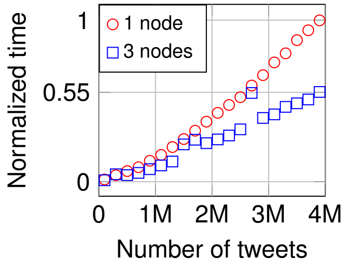


Figure 9: Named-entity recognition runtime comparison versus number of tweets.

tweet (*e.g.* language, geo-location), CATS extracts and stores advanced meta-data using natural language processing functions available through the NLTK Python package [1]. The meta-data are either directly accessible by CATS users (*e.g.* named-entities mentioned in tweets) or serve as an input for mining algorithms (*e.g.* lemmatized tweet content) in order to provide optimal results.

Performance

In order to maximize efficiency, most of the operations performed by CATS are parallelized and distributed using MapReduce. Figure 7 highlights the scalability of CATS, by showing that (i) meta-data extraction time grows linearly in the number of tweets, and that (ii) it drops by a factor of 10 when comparing the single-thread instance of CATS with a 12-thread one. Figure 8 shows that the processing time drops by an average factor of 6 for the extraction of the vocabulary, while it drops by an average factor of 4 for named-entity recognition by going from one to three MongoDB instances on a single server.

Demonstration

During the demo, we will showcase the different functions CATS has to offer, with the aim of demonstrating their usefulness for researchers from various fields.

CSCW attendees will be able to analyze two corpora; a large scale corpus gathered in a previous study [5] (several million tweets), and one which will be collected in real-time through the hashtag of the conference (#CSCW2016) with the aim of enhancing the conference experience for attendees.

We will provide credentials to every interested attendee, so that they can access CATS anytime during the conference. This will allow them to have a hands-on experience with CATS, enabling them to analyze the topics pervading the discussions happening on the hashtag of the conference, or to identify trending papers and presentations for example. We will also offer longer-term credentials for attendees who are interested in using CATS for their own research via our dedicated server[3].

Conclusion

CATS is the result of the cooperation between researchers from social and computer sciences, and aims at bridging the gap there is between these two disciplines. The end goal of CATS is to represent a common platform for researchers from any background, so that anyone can take advantage of the ever-growing resources that social media have to offer. In this paper, we have described the current state of CATS. Several improvements are being developed, among which are an algorithm for opinion mining [4] and more sophisticated visualizations.

References

[1] S. Bird, E. Klein, and E. Loper. 2009. *Natural Language Processing with Python.* O'Reilly.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

[3] Kristina Chodorow and Michael Dirolf. 2010. *MongoDB: The Definitive Guide* (1st ed.). O'Reilly.

[4] M Dermouche, L Khouas, J Velcin, and S Loudcher. 2015. A Joint Model for Topic-Sentiment Modeling from Text. In *SAC '15*.

[5] M. Gauthier, A. Guille, A. Deseille, and F. Rico. 2015. Text Mining and Twitter to Analyze British Swearing Habits. *Handbook of Twitter for Research* (2015).

[6] A. Guille and C. Favre. 2015. Event detection, tracking and visualization in Twitter: a mention-anomaly-based approach. *Social Network Analysis and Mining* 5 (2015), 18:1–18:18.

[3] URL: http://mediamining.univ-lyon2.fr/cats