On Building a Reusable Twitter Corpus

Richard McCreadie[1], Ian Soboroff[2], Jimmy Lin[3],
Craig Macdonald[1], Iadh Ounis[1], Dean McCullough[2]

[1] University of Glasgow, Glasgow G12 8QQ, UK
[2] National Institute of Science and Technology, MD 20899, USA
[3] University of Maryland, MD 20742, USA

{richard.mccreadie, craig.macdonald, iadh.ounis}@glasgow.ac.uk[1],
{ian.soboroff, dean.mccullough}@nist.gov[2], jimmylin@umd.edu[3]

ABSTRACT

The Twitter real-time information network is the subject of research for information retrieval tasks such as real-time search. However, so far, reproducible experimentation on Twitter data has been impeded by restrictions imposed by the Twitter terms of service. In this paper, we detail a new methodology for building and distributing Twitter corpora, developed through collaboration between the Text REtrieval Conference (TREC) and Twitter. In particular, we detail how the first publicly available Twitter corpus – referred to as Tweets2011 – was distributed via lists of tweet identifiers and dedicated tweet crawling software. Furthermore, we analyse whether this distribution approach remains robust over time, as tweets in the corpus are removed by users. Tweets2011 was successfully used by 58 participating groups for the TREC 2011 Microblog track, and our results attest to the robustness of the crawling methodology over time.

Categories and Subject Descriptors: H.3.3 [Information Storage & Retrieval]: Information Search & Retrieval

General Terms: Experimentation, Performance

Keywords: Twitter, Corpus Creation, Reproducibility

1. INTRODUCTION

Twitter is a communications platform on which users can send short, 140-character messages, called “tweets”, to their “followers” (other users who subscribe to those messages). Conversely, users can receive tweets from people they follow via a number of mechanisms, including web clients, mobile clients, and SMS. As of Spring 2011, Twitter has over 140 million active users worldwide, who collectively post over 340 million tweets per day. Twitter is an active research area in the information retrieval (IR) field [1]. However, previously it has not been possible to build and distribute reusable tweet corpora due to restrictions placed upon researchers by Twitter's terms of service.[1] Indeed, this has resulted in two prior unsuccessful attempts to share Twitter data by Stanford and Edinburgh universities.

The Text REtrieval Conference (TREC) is a workshop series that aims to improve the state of the art in information access task effectiveness through building sharable *test collections*. Beginning in 2011, TREC ran the Microblog track that investigated tweet search and ranking [2]. In line with the TREC aims and in collaboration with Twitter, a new methodology was developed to build and distribute a publicly available Twitter dataset, known as *Tweets2011*. In this paper, we describe this new methodology, whereby a corpus like Tweets2011, is distributed as a set of tweet identifiers and a tweet crawling tool for downloading the identified tweets. However, since researchers separately download the tweets themselves at different times, and the set of tweets available is not static (as tweets can be deleted), the exact composition of the corpus will vary depending upon when it is downloaded. We also examine how robust the distribution methodology is over time, by comparing crawls of Tweets2011 made at different points in time. Our results show that over the time period tested, the changes in the corpus over time had no noticeable effect on the systems that participated in the TREC 2011 Microblog track.

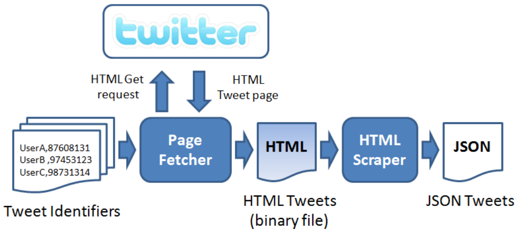

Figure 1: Illustration of tweet crawling.

2. DISTRIBUTION METHODOLOGY

The Twitter terms of service forbids third parties from data redistribution, which means that researchers that have gathered tweets cannot share them. To overcome these constraints, a compromise had to be reached. In particular, a collection would not consist of the tweets themselves, but rather (username, tweet id) pairs and associated software for reconstructing the tweets.

The Twitter REST API provides flexible access to any available tweet; nearly all common Twitter capabilities can be programmatically accessed, e.g., posting new tweets, retweeting, following a user, searching, etc. The API is generally available to the public, although by default it is rate limited; the most common unauthenticated connection places a limit of 150 requests per hour. This restriction makes it impractical to gather large number of tweets for offline processing. Historically, Twitter has lifted the API request limit for some clients based on a particular IP address or an authentication token, but this capability is no longer offered.

[1] `twitter.com/tos`

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

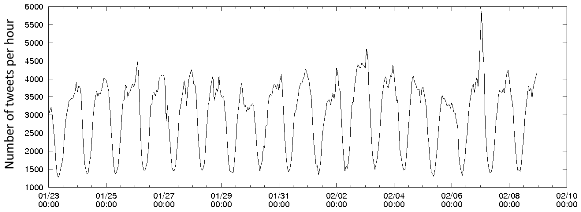

Figure 2: Tweets2011 tweet distribution over time.

The software for reconstructing tweets is an asynchronous HTTP fetcher that downloads each tweet individually. For researchers that had access to the REST API without rate limit restrictions, that can be used to download each tweet in JavaScript Object Notation (JSON) format. However, for researchers without this access, the fetcher instead crawls raw HTML pages from the `twitter.com` site and reconstructs the tweets in JSON format, as shown in Figure 1.

3. TWEETS2011 AND TREC 2011

We used the above methodology to distribute the Tweets2011 collection to the participants of the TREC 2011 Microblog track, which is now available to everyone.[2] To create Tweets2011, we identified a common set of tweets (username, tweet id pairs), for distribution. In particular, we created a sample from tweets posted during the period from January 23rd to February 8th, 2011 (inclusive). It was important to sample, as even for users who were not rate limited, downloading billions or more tweets in a sequential manner would not be practical. Instead, some spam removal was performed and then approximately 1% of the remaining tweets were sampled for the corpus, resulting in a set of approximately 16 million tweets. The distribution of these tweets over the two week period is shown in Figure 2. The chosen time period includes the Egyptian revolution as well as the US Superbowl, and a spike of tweeting activity on February 6th is easily observed. Moreover, to ensure the corpus was representative of the multi-lingual tweet retrieval environment, no language filtering was performed.

For the evaluation of TREC 2011 participating systems, 49 topics were created. From the pool of 50,324 tweets formed from the participants runs for these topics, 2,965 were judged relevant.

4. COLLECTION DEGRADATION

Tweets2011 is unique in the history of information retrieval test collections in that it will degrade over time. Twitter users can delete their tweets, or mark their accounts as private, and after that point, these tweets will not be part of the collection. This is an experimental challenge because system effectiveness may be affected by missing tweets even if those tweets are not relevant, for example due to altered collection statistics. If two systems are compared on different versions of the collection, we worry that those comparisons may not be valid.

Figure 3 illustrates the effect as we have been able to observe it in the short lifetime of the collection so far. When someone downloads the collection in HTML format, which was the case for nearly all participants, each tweet has an HTTP status code associated with it. HTTP 404 indicates a deleted tweet, 403 indicates a protected tweet, and 302 indicates a retweet. Status 301 represents a new report code from the Twitter API and we are not entirely certain what it indicates; preliminary investigations suggest that it occurs for users that have changed their screen name since the tweets were originally sampled, but the frequency seems much too high for that to be the only explanation. For the 4th of November 2011 crawl, 3,424,155 (21.2%) of the original 16,141,809 tweets were unavailable. Note that for some status codes, the number of missing tweets is not monotonically increasing. This behavior can result from crawler (download) errors, as well as when users mark their accounts as private and subsequently open them again.

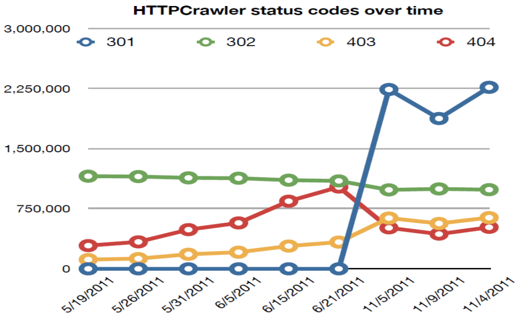

Figure 3: HTTP statuses of participant crawls of the Tweets2011 collection over time.

We measured the effect of this collection decay by re-crawling the collection after all participants had done so, and removing 200 tweets that were pooled and judged for the 49 topics, but which were subsequently deleted or protected. We then computed the precision at rank 30 of all participating runs, and compared the ordering of runs by P@30 to the official results using Kendall's τ. The correlation was 0.99, indicating that missing tweets seemed to affect all participating systems equally if at all.

Further decay in the collection may render it unusable at some point in the future. We do not recommend comparing evaluation scores done after TREC to the official TREC results, because of decay concerns. Rather, experimenters should compare multiple systems (or versions of the same system) that all use the same crawl, or closely contemporaneous crawls. This is not a concern for the official TREC results since participants all crawled the collection during a short window of time.

5. CONCLUSIONS

We detailed a new methodology for building and distributing tweet corpora, developed through collaboration between the Text REtrieval Conference (TREC) and Twitter. In particular, we described how tweet corpora can distributed via lists of tweet identifiers in tandem with tweet crawling software. This distribution approach will work for any collection of tweets. We detailed the first corpus to be distributed in this manner, Tweets2011. We recognized that the corpus may change over time as tweets are deleted, and through an analysis of different versions of the Tweets2011 corpus downloaded by TREC participants, we showed that the distribution method, at least to date, is robust to changes in the underlying corpus.

6. REFERENCES

[1] M. Busch, K. Gade, B. Larson, P. Lok, S. Luckenbill and J. Lin. Earlybird: Real-Time Search at Twitter. In *Proc. of ICDE'12.*

[2] I. Ounis, C. Macdonald, J. Lin and I. Soboroff. Overview of the TREC-2011 Microblog Track In *Proc. of TREC'11.*

[2] `trec.nist.gov/data/tweets/`