Finally, a Downloadable Test Collection of Tweets

Royal Sequiera and Jimmy Lin
David R. Cheriton School of Computer Science
University of Waterloo, Ontario, Canada
{rdsequie,jimmylin}@uwaterloo.ca

ABSTRACT

Due to Twitter's terms of service that forbid redistribution of content, creating publicly downloadable collections of tweets for research purposes has been a perpetual problem for the research community. Some collections are distributed by making available the ids of the tweets that comprise the collection and providing tools to fetch the actual content; this approach has scalability limitations. In other cases, evaluation organizers have set up APIs that provide access to collections for specific tasks, without exposing the underlying content. This is a workable solution, but difficult to sustain over the long term since someone has to maintain the APIs. We have noticed that the non-profit Internet Archive has been making available for public download captures of the so-called Twitter "spritzer" stream, which is the same source as the Tweets2013 collection used in the TREC 2013 and 2014 Microblog Tracks. We analyzed both datasets in terms of content overlap and retrieval baselines to show that the Internet Archive data can serve as a drop-in replacement for the Tweets2013 collection, thereby providing the research community with, *finally*, a downloadable collection of tweets. Beyond this finding, we also study the impact of tweet deletions over time and how they affect the test collections.

1 INTRODUCTION

Test collections—comprised of a corpus of documents, a set of information needs, and associated relevance judgments—lie at the heart of the Cranfield Paradigm [2] for information retrieval research. In most cases, researchers can acquire the document collection under study: in the 1990s, these were on physical CD-ROMs or DVDs delivered via postal mail; today, hard drives are shipped instead. What if it were not possible to distribute document collections for research use? One example is a collection of tweets: Twitter's terms of service forbid redistribution of such data. This is not a Twitter-specific problem, as similar challenges exist with electronic medical records, emails, and a host of other sensitive collections researchers may wish to study.

Over the past several years, the community has experimented with and developed alternative evaluation approaches for cases where the distribution of documents is challenging, collectively known as "Evaluation as a Service" (EaaS) [1, 3]. Specifically for tweets, TREC organizers have built a search API for researchers to perform evaluation tasks without bulk access to the raw collection [4]; this approach was deployed in both the TREC 2013 and TREC 2014 Microblog Track evaluations.

The Internet Archive[1]—a nonprofit digital library with the mission of providing "universal access to all knowledge"—has been making available captures of the so-called Twitter "spritzer" stream (an approximately 1% sample of public posts) for download. Putatively, this is the same source that was used to construct the Tweets2013 collection used in the TREC 2013 and 2014 Microblog Tracks. A natural question, therefore, is how this dataset compares to the official Tweets2013 collection and if it can be used as a drop-in replacement for evaluation purposes. The main contribution of this paper is in answering these questions: we find that, *yes*, the publicly downloadable Internet Archive data is substantially similar to the official Tweets2013 collection. We observe around 95% overlap in terms of tweet content, and retrieval baselines on the Internet Archive data yield effectiveness that is statistically indistinguishable from the official API. Thus, the information retrieval community *finally* has access to a downloadable collection of tweets for research, obviating the need for the service API.

Beyond the contribution of validating a downloadable Twitter test collection, this paper also takes a closer look at deleted tweets. The fact that users can delete their tweets means that any collection is constantly changing, and the size of the collection monotonically decreases over time (since there is no "undelete" option). We present an analysis of deleted tweets in the Tweets2013 collection over the past several years to quantitatively characterize the delete process and to examine effects on retrieval effectiveness. We find that although the collection indeed degrades over time, and almost a fifth of tweets from the raw Tweets2013 collection have been deleted as of December 31, 2016, these deletes appear to have minimal impact on the integrity of the test collections built on the tweets.

2 BACKGROUND AND RELATED WORK

Restrictions on the redistribution of tweets have long been a hurdle to building test collections for information seeking on social media streams. The TREC Microblog Tracks, which ran from 2011 to 2015, have wrestled with this issue and experimented with two different solutions. The track organizers built the Tweets2011 collection that was used in TREC 2011 and 2012 [6]. To circumvent the no-redistribution limitation, the organizers devised a process whereby NIST distributed the *ids* of the tweets (rather than the tweets themselves). Given these ids and a downloading program developed by the organizers (essentially, a crawler), a participant could "recreate" the collection [5]. Since the downloading program

[1] https://archive.org/

SIGIR'17, August 7–11, 2017, Shinjuku, Tokyo, Japan

978-1-4503-5022-8/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3077136.3080667

accessed the twitter.com site directly, the tweets were delivered in accordance with Twitter's terms of service.

The "download it yourself" approach successfully addressed the no-redistribution issue for the purposes of a shared evaluation, as evidenced by 59 participating groups in the TREC 2011 Microblog Track (one of the largest ever in the history of TREC). Beyond TREC, this approach has been adopted by other communities for sharing collections of tweets. However, distribution via re-downloading exhibits scalability limitations. In particular, the speed of the downloading program, which has built-in rate limiting (imposed voluntarily for robotic "politeness"), sets a practical upper bound on collection size. The Tweets2011 collection originally contained 16 million tweets, which is small by modern standards, especially considering that tweets are short.

The Tweets2011 collection also identified another issue with tweet collections in general, explored by McCreadie et al. [5]: they degrade over time due to deletes. Based on recrawls of the collection several months after its original release, the authors concluded that the deletes did not impact the relative effectiveness of runs submitted to TREC 2011. However, to our knowledge, the effects of deletes over much longer periods of time have not be studied.

In order to tackle the scalability challenges associated with the "download it yourself" approach, for the TREC 2013 Microblog Track the organizers implemented an evaluation-as-a-service solution. They gathered a collection of tweets centrally, but instead of distributing the tweets, the organizers provided a service API through which participants could access the tweets to complete the evaluation task. To build the official collection, organizers developed an open-source crawler using the twitter4j Java library[2] to gather tweets from Twitter's public sample stream,[3] colloquially known as the "spritzer" stream. This level of access is available to anyone with a Twitter account and does not require special authorization. The organizers crawled all tweets between February 1 and March 31, 2013, UTC (inclusive). According to the TREC 2013 Microblog Track overview: The collection was gathered from two separate virtual machine instances on Amazon's EC2 service, one on the east coast of the US, and the other on the west coast of the US. The redundant setup guarded against network outages and other operational issues during the collection period. Fortunately, no downtime was experienced, so one of the copies was simply designated as the official collection. In total, the organizers reported gathering 259 million tweets, although at the time of the evaluation, the collection behind the API was reduced to 243 million tweets after the removal of deleted tweets.

The API for accessing the Tweets2013 collection provided basic search capabilities using the open-source Lucene search engine. In addition to returning the text of the retrieved tweets, the API returned associated metadata about the time of the post, the user making the post, and other properties such as the number of retweets, whether the tweet was a reply, etc. Although the setup essentially limited participants to reranking tweets, this is not unlike multi-stage ranking architectures that are common today [9]. Additional meta-evaluations have shown that using the API does not appear to affect the diversity of the submitted runs [8] and a retrievability analysis does not reveal any substantive issues that arise from not having access to the entire collection [7].

Although the evaluation-as-a-service approach is a workable solution for some tasks, the biggest challenge of the approach is sustainability over the long term, since someone must ultimately devote resources to the service, manage access, troubleshoot issues, etc. This is an open-ended commitment for the life of the collection: as a point of comparison, TREC test collections from the 1990s are still being used today. It is difficult to imagine anyone supporting the API for two decades. For one, the software behind the service will have long become obsolete.

[2]http://twitter4j.org/en/index.html
[3]https://dev.twitter.com/docs/streaming-apis

Source	Count
$\lvert\mathcal{T}\rvert$	259,035,603
$\lvert\mathcal{A}\rvert$	246,615,368
$\lvert\mathcal{T} \cup \mathcal{A}\rvert$	260,382,756
$\lvert\mathcal{T} \cap \mathcal{A}\rvert$	245,268,215
$\lvert\mathcal{T} - \mathcal{A}\rvert$	13,767,388
$\lvert\mathcal{A} - \mathcal{T}\rvert$	1,347,153

Table 1: Collection statistics, where $\mathcal{T}$ represents the raw Tweets2013 collection and $\mathcal{A}$ represents the Tweets2013-IA collection from the Internet Archive.

3 COLLECTION STATISTICS

In this paper, we examine two tweet datasets available from the Internet Archive for public download:

- ArchiveTeam JSON Download of Twitter Stream 2013-02: https://archive.org/details/archiveteam-twitter-stream-2013-02
- ArchiveTeam JSON Download of Twitter Stream 2013-03: https://archive.org/details/archiveteam-twitter-stream-2013-03

According to the Internet Archive, the above datasets are:

> A simple collection of JSON grabbed from the general Twitter stream, for the purposes of research, history, testing and memory. This is the "Spritzer" version, the most light and shallow of Twitter grabs.

Putatively, this is the same source that the Tweets2013 collection was created from. We downloaded these tweets and compared them against the official Tweets2013 collection (collected by the organizers). In Table 1 we present some basic collection statistics for the raw Tweets2013 collection, which we denote as $\mathcal{T}$, and the above datasets downloaded from the Internet Archive, which we refer to as Tweets2013-IA and denote as $\mathcal{A}$ for short. Note that $\mathcal{T}$ is *not* the collection exposed via the official API, since deletes were applied to it before the TREC evaluations.

Twitter's streaming API is formatted in JSON and comprises messages of two types: actual tweet content and delete messages. These statistics consider tweet JSON messages only. Due to transient network issues, some messages are delivered more than once, and therefore all reported statistics in this paper are on *unique* counts. For all experiments in this paper, data manipulation is performed using Spark on our Hadoop cluster since the datasets are large; for reference, the raw Tweets2013 collection (including all tweets and deletes) is 107 GB compressed.

Overlap statistics between $\mathcal{T}$ and $\mathcal{A}$ are shown in Table 2. Most importantly, we see that there is approximately 95% overlap in tweet content between the publicly downloadable datasets from

Collection	Overlap
$1 - \|(\mathcal{T} - \mathcal{A})\|/\|\mathcal{T}\|$	94.69%
$1 - \|(\mathcal{A} - \mathcal{T})\|/\|\mathcal{A}\|$	99.45%

Table 2: Overlap analysis between the Tweets2013 ($\mathcal{T}$) and Tweets2013-IA ($\mathcal{A}$) collections.

the Internet Archive and the raw Tweets2013 collection. It seems that the latter is nearly a superset of the former, as there are very few tweets in $\mathcal{A}$ that are not in $\mathcal{T}$.

4 DELETION ANALYSIS

Per the Twitter Developer Agreement, one must "delete content that Twitter reports as deleted or expired".[4] Alongside the tweet content, Twitter's streaming API also delivers delete messages. This means that, in order to precisely follow the agreement, gathering any Twitter content from the API also incurs an open-ended liability to monitor the stream indefinitely for delete messages.

From the perspective of IR evaluation, this means that any collection of tweets is unstable and will degrade over time—the collection size will monotonically decrease, since there is no "undelete" feature. Although McCreadie et al. [5] have previously examined this issue, their analysis was over a much smaller collection and a much shorter time span. Here, we characterize deletes on the raw Tweets2013 collection over a much longer period of time and examine its impact on associated test collections.

The deletion data in our analysis come from two long-term crawls of the Twitter spritzer stream from April 2013 through December 2016 (inclusive). Due to occasional crawler failures, we take the union of delete messages observed across both crawls as the "ground truth". The notation D(YY/MM-YY/MM) refers to deletes observed between the specified years and months, inclusive. D(13/02-13/03) is observed directly in the raw Tweets2013 collection, while all other deletes come from the sources described above.

Deletion statistics are shown in Table 3, where we provide numbers for a few noteworthy periods: We show the count of deletes that are directly observed as part of the collection (in February and March of 2013). The period from February to June 2013 (inclusive) captures the deletes that were applied for the service API made available for TREC 2013 and TREC 2014. Also of interest are the delete aggregates at yearly intervals, i.e., the counts of deletes through the end of 2013, 2014, 2015, and 2016.

In Table 3 we also show the effects of removing the deleted tweets from the raw Tweets2013 collection $\mathcal{T}$ and also the Tweets2013-IA collection $\mathcal{A}$. From the table, we see that by the end of 2016, deletes have reduced the collection to 211m for $\mathcal{T}$ and 199m for $\mathcal{A}$, down from the original sizes of 259m and 247m, respectively. Figure 1 plots the number of deletes by month on both the raw Tweets2013 collection and the Internet Archive data. Although we do see that the number of deletes drops off after the initial few months, there is still a substantial number of deletes even years after the tweets were originally posted. The total size after applying all deletes is shown in Figure 2; as expected, we see a steady degradation of the collection over time.

The next obvious question is how these deletes affect test collections from the TREC 2013 and 2014 Microblog Tracks that have been built on the Tweets2013 data. The answer is shown in Table 4, which lists for various conditions the number of relevant documents and qrels (all judgments in the pool, regardless of relevance) that would have disappeared as a result of the deletes. We see that by the end of 2016, a little over 5% of the relevant documents would have been deleted. This is a smaller value than the fraction of the entire collection that is deleted, which means that deletes are more likely to affect non-relevant documents.

[4] https://dev.twitter.com/overview/terms/agreement-and-policy

Source	Count
$\|\mathcal{T}\|$	259,035,603
$\|\mathcal{A}\|$	246,615,368
$\|D(13/02\text{-}13/03)\|$	10,631,099
$\|D(13/04\text{-}13/06)\|$	5,091,183
$\|D(13/07\text{-}13/12)\|$	7,197,460
$\|D(14/01\text{-}14/12)\|$	96,98,613
$\|D(15/01\text{-}15/12)\|$	7,928,857
$\|D(16/01\text{-}16/12)\|$	7,496,871
$\|\mathcal{T} - D(13/02\text{-}13/03)\|$	248,404,504
$\|\mathcal{A} - D(13/02\text{-}13/03)\|$	234,337,730
$\|\mathcal{T} - D(13/02\text{-}13/06)\|$	243,313,321
$\|\mathcal{A} - D(13/02\text{-}13/06)\|$	230,893,086
$\|\mathcal{T} - D(13/02\text{-}13/12)\|$	236,115,861
$\|\mathcal{A} - D(13/02\text{-}13/12)\|$	223,695,626
$\|\mathcal{T} - D(13/02\text{-}14/12)\|$	226,417,248
$\|\mathcal{A} - D(13/02\text{-}14/12)\|$	213,997,013
$\|\mathcal{T} - D(13/02\text{-}15/12)\|$	218,488,391
$\|\mathcal{A} - D(13/02\text{-}15/12)\|$	206,068,156
$\|\mathcal{T} - D(13/02\text{-}16/12)\|$	210,991,520
$\|\mathcal{A} - D(13/02\text{-}16/12)\|$	198,571,285

Table 3: Deletion statistics, applying deletes observed in the Twitter "spritzer" stream over time.

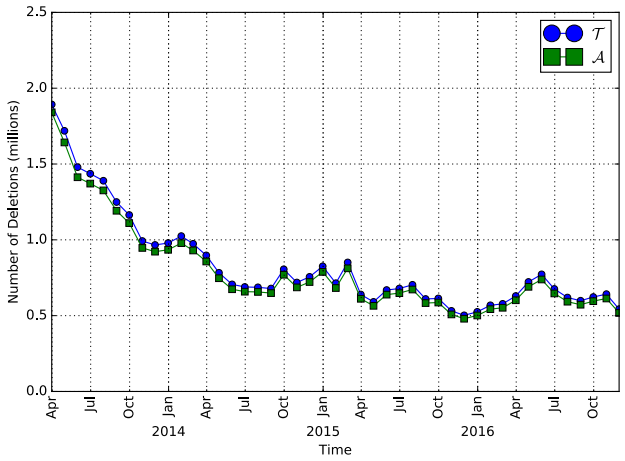

Figure 1: Number of tweets deleted from Tweets2013 and Tweets2013-IA over time.

5 RETRIEVAL EXPERIMENTS

In our final set of experiments, we examined the effectiveness of baseline retrieval techniques on some of the variant collections

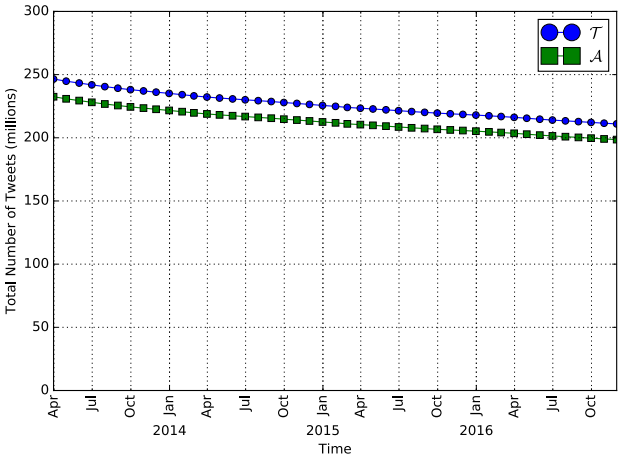

Figure 2: Size of the Tweets2013 and Tweets2013-IA collection over time after applying observed deletes.

Source	missing reldocs	missing qrels
$\lvert\mathcal{T} - D(13/02\text{-}13/12)\rvert$	220 (1.12%)	1,820 (1.41%)
$\lvert\mathcal{A} - D(13/02\text{-}13/12)\rvert$	209 (1.06%)	1,707 (1.32%)
$\lvert\mathcal{T} - D(13/02\text{-}14/12)\rvert$	539 (2.74%)	4,456 (3.45%)
$\lvert\mathcal{A} - D(13/02\text{-}14/12)\rvert$	513 (2.61%)	4,190 (3.24%)
$\lvert\mathcal{T} - D(13/02\text{-}15/12)\rvert$	816 (4.15%)	6,576 (5.09%)
$\lvert\mathcal{A} - D(13/02\text{-}15/12)\rvert$	776 (3.95%)	6,193 (4.79%)
$\lvert\mathcal{T} - D(13/02\text{-}16/12)\rvert$	1,095 (5.57%)	8,500 (6.58%)
$\lvert\mathcal{A} - D(13/02\text{-}16/12)\rvert$	1,042 (5.30%)	7,997 (6.19%)

Table 4: Deletion statistics over relevance judgments; percentage of total is shown in parentheses.

explored above. The reference point for comparison is the official Thrift API, which served a collection of 243 million tweets (taking into account deletions up until the time the API was deployed for the evaluation). For our experiments, we used exactly the same code base[5] as the API, which was built on top of the open-source Lucene search engine (although in our case, we had direct access to the Lucene indexes).

For evaluation, we used 60 topics from TREC 2013 and 55 topics from TREC 2014. Ranking was performed using Lucene's implementation of query-likelihood, just as with the API. Following standard practice, we retrieved up to 1000 hits per topic and measured effectiveness in terms of average precision (AP) and precision at 30 (P30), the two official metrics used in the evaluations. We report results with the official *original* NIST qrels. However, it would certainly be reasonable to remove deleted tweets from the judgments. We do so and report results under the *modified* qrels condition.

Experimental results are shown in Table 5 for both the original and modified qrels. The condition denoted $\mathcal{T} - D(13/02\text{-}13/06)$ attempts to replicate the data conditions of the official API; our results are very close but not exactly the same because we only consider deletes at monthly increments. The condition $\mathcal{A} - D(13/02\text{-}13/06)$

[5]http://twittertools.cc/

	Original		Modified	
Track	**AP**	**P30**	**AP**	**P30**
Official Thrift API	0.3198	0.5278	-	-
$\mathcal{T} - D(13/02\text{-}13/06)$	0.3120	0.5278	0.3120	0.5278
$\mathcal{A} - D(13/02\text{-}13/06)$	0.2951	0.5130	0.2951	0.5130
$\mathcal{T} - D(13/02\text{-}16/12)$	0.2996	0.5220	0.3158	0.5220
$\mathcal{A} - D(13/02\text{-}16/12)$	0.2864	0.5130	0.3013	0.5130

Table 5: Effectiveness measures on TREC 2013 and 2014 Microblog Track topics over different data conditions.

represents the best that a researcher can obtain in replicating the official API using publicly available resources. Based on paired t-tests, we do not find any significant differences (at $p < 0.01$) between the official Thrift API and these two data conditions, for both metrics (AP and P30), with either the original qrels or the modified qrels.

The last two rows in Table 5 show the state of the collection as of December 31, 2016 if all deletes were applied. We also do not find any significant differences between these two data conditions and the official Thrift API, for both metrics and both qrel conditions.

6 CONCLUSIONS

The Tweets2013 collection, used in the TREC 2013 and TREC 2014 Microblog Tracks, serves as the basis of the most comprehensive evaluation resource for *ad hoc* retrieval on social media to date. Hampering its availability, however, is the API-based access mechanism. However, courtesy of the Internet Archive, researchers now can directly download tweets from the same period and same source as the official Tweets2013 collection. Our analyses confirm that, indeed, the Internet Archive data can serve as a drop-in replacement for evaluation purposes. We share with the community all code and data associated with analyses in this paper as well as instructions for replicating reported data conditions to serve as the basis of future work.[6] *Finally*, researchers now have a downloadable test collection of tweets!

REFERENCES

[1] Allan Hanbury, Henning Müller, Krisztian Balog, Torben Brodt, Gordon V. Cormack, Ivan Eggel, Tim Gollub, Frank Hopfgartner, Jayashree Kalpathy-Cramer, Noriko Kando, Anastasia Krithara, Jimmy Lin, Simon Mercer, and Martin Potthast. 2015. Evaluation-as-a-Service: Overview and Outlook. *arXiv:1512.07454*.
[2] Donna Harman. 2011. *Information Retrieval Evaluation*. Morgan & Claypool Publishers.
[3] Jimmy Lin and Miles Efron. 2013. Evaluation as a Service for Information Retrieval. *SIGIR Forum* 47, 2 (2013), 8–14.
[4] Jimmy Lin and Miles Efron. 2013. Overview of the TREC-2013 Microblog Track. In *TREC*.
[5] Richard McCreadie, Ian Soboroff, Jimmy Lin, Craig Macdonald, Iadh Ounis, and Dean McCullough. 2012. On Building a Reusable Twitter Corpus. In *SIGIR*. 1113–1114.
[6] Iadh Ounis, Craig Macdonald, Jimmy Lin, and Ian Soboroff. 2011. Overview of the TREC-2011 Microblog Track. In *TREC*.
[7] Jiaul H. Paik and Jimmy Lin. 2016. Retrievability in API-Based "Evaluation as a Service". In *ICTIR*. 91–94.
[8] Ellen M. Voorhees, Jimmy Lin, and Miles Efron. 2014. On Run Diversity in "Evaluation as a Service". In *SIGIR*. 959–962.
[9] Lidan Wang, Jimmy Lin, and Donald Metzler. 2011. A Cascade Ranking Model for Efficient Ranked Retrieval. In *SIGIR*. 105–114.

Acknowledgments. This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada, with additional contributions from the U.S. National Science Foundation under IIS-1218043 and CNS-1405688.

[6]https://github.com/castorini/Tweets2013-IA