Strategies for Collecting, Processing, and Analyzing Tweets from Large Newsworthy Events

Nick Ruest
York University
4700 Keele St.
Toronto, Ontario M3J 1P3
ruestn@yorku.ca

KEYWORDS

Web archiving, Twitter

#WomensMarch, #Aleppo, #paris, #bataclan, #parisattacks, #porteouverte, #jesuischarlie, #jesuisahmed, #jesuisjuif, #charliehebdo, #panamanpapers, and #exln42 are all different hashtags, but they share several things in common. They are all large newsworthy events. They are datasets that each contain over a million tweets. Most importantly these collections raise some interesting insights in collecting, processing, and analyzing large newsworthy events[2].

Collecting tweets from these events can be challenging because of timing. Tweets can be collected from the Filter API[1] and Search API[2]. Both having their own caveats. The Filter API only captures the current Twitter stream, and is limited to collecting up to 1% of the overall Twitter stream. The Search API allows you to collect more than 1% of the overall Twitter stream[1], but one can only collect up to 18,000 every 15 minutes, and is limited to a 7 day window. Generally, using a strategy of using the Filter and Search API to capture a given event is the best.

DocNow's twarc[3] includes a number of utilities to process a dataset after collection. These tools allow a researcher, librarian, or archivist to filter their dataset(s) down to what is needed for appraisal, and then accession. Noteworthy tools include; deduplication, source, retweets, date/times, users, and hashtags.

DocNow's utilities[4] can be further used to curate related collections. One can extract all the urls of a dataset, unshorten them, and extract the unique urls to use as a seed list for a web crawler to capture websites related to a given event. One can also extract all of the image urls, and download all images associated with a dataset, which then can be used for image analysis[3], presentation, and/or preservation.

In conclusion, this presentation provides an overview of collection strategy, insights from processing and analysis, ensuing web crawls, and image presentation from each collection.

---

[1] https://dev.twitter.com/streaming/public
[2] https://dev.twitter.com/rest/public/search
[3] http://github.com/docnow/twarc
[4] https://github.com/DocNow/twarc/tree/master/utils

REFERENCES

[1] Kevin Driscoll and Shawn Walker. 2014. Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data. *International Journal of Communication* 8 (2014), 1745–1764.
[2] Ian Milligan, Nick Ruest, and Jimmy Lin. 2016. Content Selection and Curation for Web Archiving: The Gatekeepers vs. The Masses. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (JCDL '16)*. ACM, New York, NY, USA, 107–110. https://doi.org/10.1145/2910896.2910913
[3] Nick Ruest. 2016. 1,203,867 #exln42 images. (March 2016). http://ruebot.net/post/1203867-elxn42-images