Looking Through the Twitter Glass: Bridging the Data – Researcher Gap

Completed Research Paper

Shohil Kishore
University of Auckland
s.kishore@auckland.ac.nz

Gabrielle Peko
University of Auckland
g.peko@auckland.ac.nz

David Sundaram
University of Auckland
d.sundaram@auckland.ac.nz

Abstract

Social media data provides researchers with insight into how social media platforms are used, and how they shape the lives of their respective users. While exploring and analyzing data has vast research potential, the process for doing so is tedious, time-consuming, and difficult, especially for non-technical researchers. In most cases, researchers require funds and technical personnel to even begin exploring data-related ideas. For those that do have the resources, low quality and high latency data is typically the end result. To decrease the overall gap between the data and the researcher, we propose a method and prototypical web-based system that utilizes Twitter's latest API offering to enable researchers to collect high quality data themselves at minimal cost. The system has been evaluated through multiple implementations at multiple universities. Our initial findings are positive, indicating that our design and prototype alleviate the major issues faced by non-technical researchers collecting social media data.

Keywords

Social Media Research, Twitter Data, Data Collection, Data – Researcher Gap, Data Latency, Social Media Data Mining.

Introduction

Social media research (SMR) has been the focus of social researchers for over a decade (Zeng et al. 2010). With the mass adoption of social media platforms such as Facebook, Twitter, and Instagram, new avenues for research have become prevalent, particularly from the perspective of data. Social media data opens gateways that provide researchers with insight into how these platforms are used, and how they impact the lives of their respective users (Mahrt and Scharkow 2013). While methods of collecting and interpreting social media data have existed for years, the process for doing so is still tedious, time-consuming, expensive and generally difficult especially for non-technical social researchers (Lomborg and Bechmann 2014).

Take the perspective of the social researcher who intends to explore an idea. Collecting social media data poses one of the first challenges. In many cases, people collect data manually or use web scraping software, both of which are costly tasks. In the latter case, the researcher must secure funds to hire a developer, clearly detail their requirements and needs, and wait for the software to be constructed before data collection can even begin to occur. Even if all of these steps are successful, scraping usually results in low quality data which cannot be validated by others using the same methodology (Ruths and Pfeffer 2014). The feedback loop for exploring a supposedly simple idea becomes cumbersome, inefficient and long drawn out.

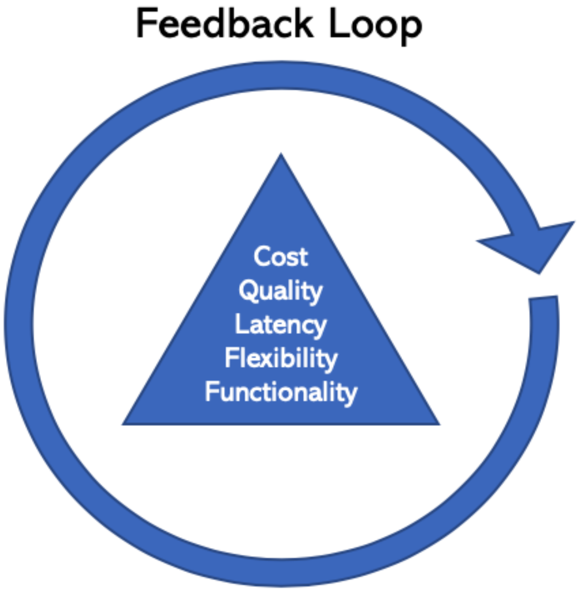

Figure 1: How Cost, Quality, Latency, Flexibility and Functionality Impact Feedback

This article presents an architecture that attempts to minimize the feedback loop by focusing on five core constraints: (1) overall cost, (2) data quality, (3) data latency, (4) flexibility, and (5) functionality to enable the initial exploration of ideas and reduce the gap between high quality data and the researcher (Figure 1). To fulfill these requirements, we propose and build a prototypical web-based system using Twitter's latest Application Programming Interface (API). While there are a number of social media platforms that allow API access, Twitter's developer community, reasonable pricing structure, data accessibility, high data quality and minimal data latency makes it the ideal platform for most researchers, especially those looking to explore the validity of an idea. The interactive system therefore enables non-technical researchers to collect replicable data with minimal effort and cost, which can then be analyzed to obtain valuable insight in a shorter period of time.

Challenges of Social Media Research

Methods of Data Collection

In the past, the most popular method of automated data collection involved web crawlers (Heydon and Najork 1999). Web crawlers are software programs designed to traverse public webpages, "scrape" potentially relevant information and download the data in an analyzable format (Glez-Peña et al. 2013). Crawlers are still widely popular today, and both paid as well as open-source versions exist online. For example, Scrapy is an extremely popular open-source web crawler written in Python (Scrapy 2019) and ScrapeStorm is a user-friendly web scraping tool with monthly paid plans (ScrapeStorm 2019).

However, there is no assurance of quality or validity when extracting data using a crawler (Myllymaki 2002). As they only have access to some public webpages, crawlers are unable to capture all data, especially on social media platforms (Bošnjak and Oliveira 2012). Furthermore, crawlers have frequently been portrayed as unethical as they do not need to seek user permission before collecting data (Thelwall and Stuart 2006).

Platform APIs address some of these concerns. As individual platforms, such as Twitter, Reddit and Facebook, control access to their respective APIs, they can provide direct access to their data (Lomborg and Bechmann 2014). Researchers, for example, can simply pay platforms such as Twitter for access to high-quality, full-fidelity and comparatively ethical data. This method of data collection has steadily gained popularity over time and has been widely adopted by researchers collecting social media data for both large-scale quantitative (Boyd and Crawford 2012; Bruns and Stieglitz 2012) and smaller scale qualitative (Lomborg 2012) research studies.

Twitter Data Collection

Until 2017, Twitter only provided two types of access to their data: the Standard Search API and the Enterprise Search API (Tornes 2017). Standard Search is free but only retrieves a sample of Tweets published in the last seven days. On the other hand, Enterprise Search has access to Twitter's entire archive and retrieves complete data. However, it is expensive as it primarily targets commercial entities (Twitter 2019a). Not only did this restrict businesses from leveraging insights from Twitter, but it also restricted researchers from collecting and analyzing data (Boyd and Crawford 2012; Ghosh et al. 2013; Morstatter et al. 2013). Twitter's latest offering, known as the Premium Search API, fills the gap between the two by providing users with access to a set number of Tweets for a set price. For example, their most recent pricing structure allows users to collect up to 100,000 Tweets for $99 USD (Twitter 2019b). They also provide assurances that their Premium Search API is low-latency and full-fidelity (Twitter 2019c).

A number of software solutions already utilize this API to enable interactions with Twitter. One of the most popular solutions is known as search-tweets-python which was created by the Twitter Development team and allows users to collect Twitter data using the command line (Twitter 2017). Similar solutions utilizing the Twitter Premium Search API also exist (Koepp 2013).

While these solutions prove to be useful to those with a technical background, non-technical researchers looking to explore an idea are restricted. Both of these solutions require some level of technical understanding, environment setup, interaction with configuration files, and occasional interaction with code, especially when there are errors. This article proposes an architecture and a prototypical web-based solution that removes these barriers and enables the collection of high quality data in a short period of time. The proposed artefact is entirely web-based therefore interactions with code or configuration files are unnecessary and setup is minimal.

The Data – Researcher Gap

Knowledge discovery in databases

The Knowledge Discovery in Databases (KDD) model is an extremely popular iterative method of deriving useful knowledge from raw data (Fayyad et al. 1996). It provides a structure to support the identification of data mining goals, selection of relevant data, cleaning and preprocessing of data, feature selection and transformation, utilization of appropriate data mining methods, exploratory analysis, pattern identification and interpretation, and finally, knowledge discovery. As depicted in Figure 2, each of these steps occurs iteratively; different data or data mining methods may be selected which may change patterns and interpretation, potentially yielding additional information. This indicates that gathering insights from data is a challenging and time-consuming process.

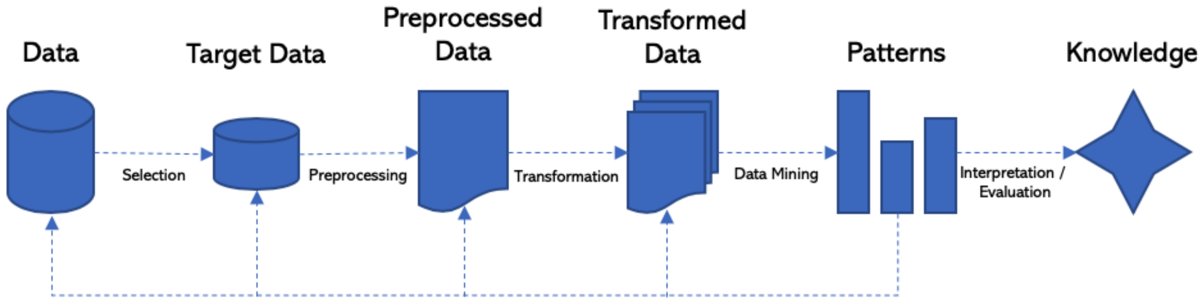

Figure 2: Knowledge Discovery in Databases Model (adapted from Fayyad et al. 2002)

Understanding the data – researcher gap

While the KDD model is still relevant today, it is common to find or generate datasets that exist outside of traditional systems such as databases. Specifically, these datasets reside in huge unstructured and semi-structured volumes, are generated at a rapid pace, and are uniquely formatted (Katal et al. 2013). Twitter's Premium Search API, for example, returns Tweet objects that includes time data, user data, location data and retweet data all in a format specific to Twitter. The skills required to collect data from the API requires technical skills outside the realm of most social researchers. This means that a developer would have to be involved, and software would most likely have to be constructed to collect data. As depicted in Figure 3, these obstructions create a significant gap between the initial idea generated by the researcher and the actual collection of data. Before social media data can be collected, funds need to be secured, a developer needs to be hired and informed, and software has to be created or modified.

Moreover, as data exploration and mining occurs, the researcher may realize that the idea is weak which means that time and funds were wasted, or that more data needs to be collected, requiring additional developer support. Simply put, data mining methodologies such as the KDD model do not take into account that collecting data for some projects, such as SMR, can be a challenging, iterative process that must occur before data selection, preprocessing, transformation, mining and interpretation can occur.

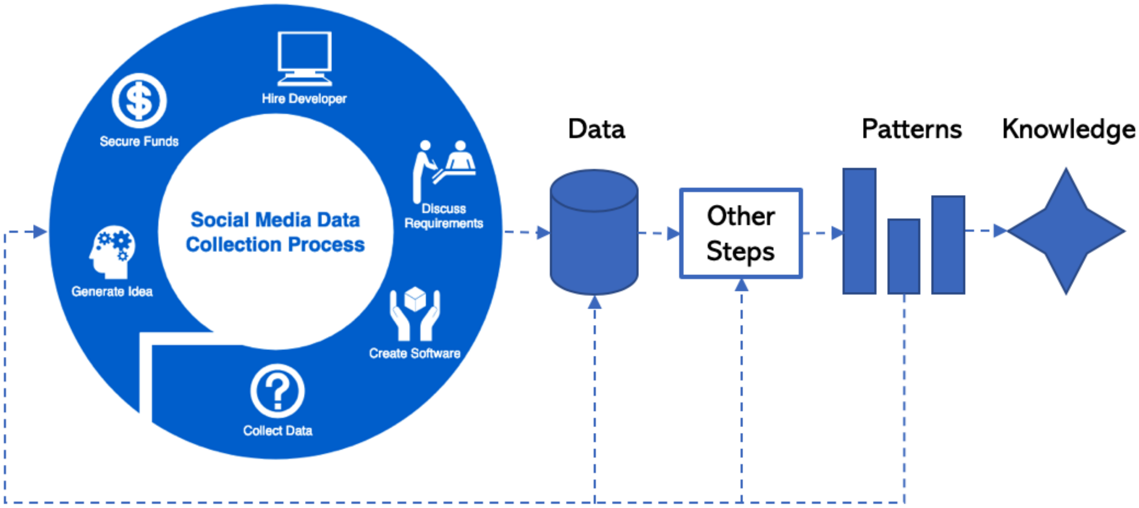

Figure 3: Current Social Media Data Collection Process

Bridging the Gap: Proposing an Adapted Model

To enable non-technical researchers, we must focus on decreasing the data – researcher gap. Ideally, researchers starting a new project should be able to explore an idea themselves, without the requirement of technical skill or the support of developers. This allows the creative process to occur more naturally and organically allowing less time to be wasted. It also removes the extreme reliance on a technician (such as a software developer) when testing ideas, and the consequent delays associated with the technician being a bottleneck resource.

We realized that a simplified web-based system, which will now be referred to as the Data Toolkit, could make the 'Securing Funding', 'Hiring a Developer,' 'Discussing Requirements' and 'Creating Software' steps redundant. The total amount of steps decreases from six in Figure 3 to three in Figure 4 through the usage and implementation of the Data Toolkit. This system allows researchers to focus on generating ideas and collecting data. Of course, the system must be easy to setup and run, and generate high quality data for minimal cost, but in doing so, researchers would be more inclined to study SMR related topics and less concerned about the technicalities of a project.

Clearly, this type of system would be useful to all researchers but in particular to:

1. Researchers with limited resources and skill
2. Researchers investigating time-critical events
3. Researchers evaluating the validity of an idea
4. Projects that may require iterative data collection
5. Projects that require high quality datasets

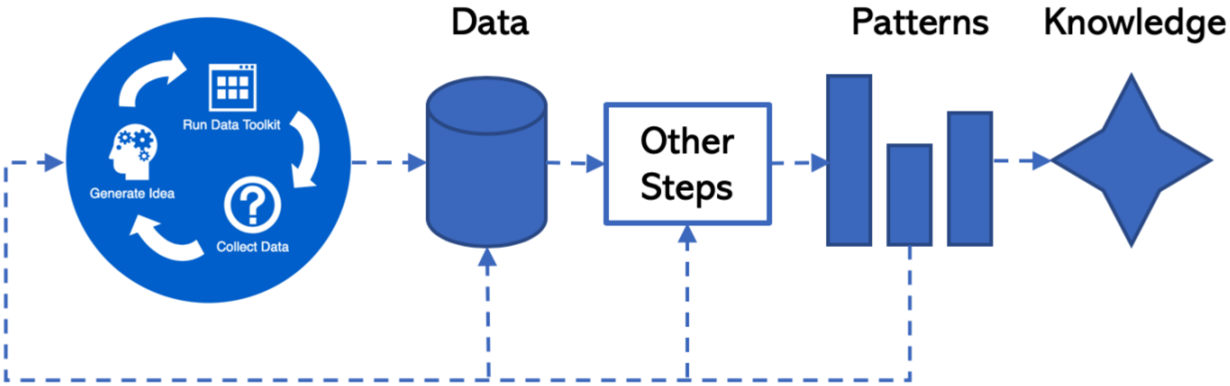

Figure 4: Proposed Social Media Data Collection Process

By allowing researchers to collect data themselves, they can quickly test if an idea is relevant or not and move on to a larger scale study or a different idea. The Data Toolkit would significantly decrease the feedback loop and the research cycle, allowing many more iterations as well as wide ranging explorations.

Most projects also tend to be time critical, but when there are projects that are particularly close to the wire, the Data Toolkit becomes invaluable in exploring alternative hypotheses, carrying out due diligence in a short period of time and retaining the maximum amount of information value. For example, Figure 5 depicts that there is high information value in an event when it occurs. However, value rapidly decreases over time, particularly due to delays in data collection and data analysis (Hackathorn 2002).

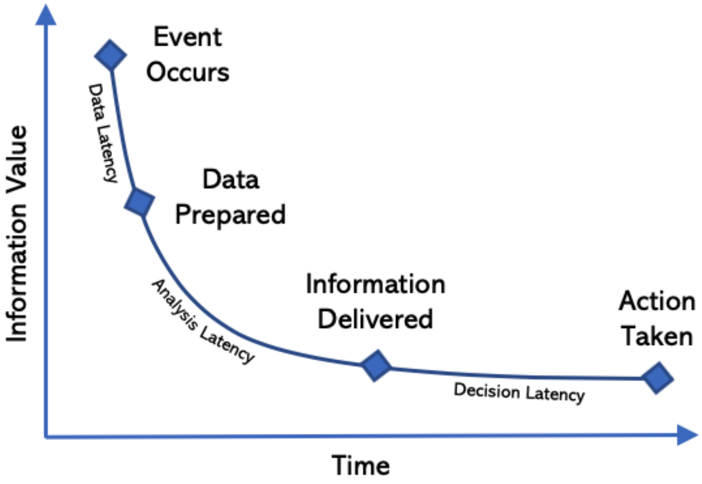

Figure 5: Information Value and Latency (adapted from Hackathorn 2002)

Closely linked to the benefits associated with time and latency is quality. Usually time and quality are inversely related. However, with the usage of such a system, high quality is also assured by utilizing Twitter's Premium Search API which allows access to all data. Beyond reducing time and increasing quality the Data Toolkit also leads to less wastage of time during exploration. That is, since data extraction is not expensive or overly time-consuming, researchers are free to explore many possible hypotheses ultimately leading to more robust research.

Application of Method and System

Software implementation

Based on the model in Figure 4 we decided to implement the Data Toolkit, an open-source web-based system that attempts to lower the barriers to entry when collecting social media data. As discussed, we chose to utilize Twitter's Premium Search API as the backbone of our system as it is relatively inexpensive and allows for the collection of full-fidelity data. Specifically, these were the system requirements:

1. Social researchers with minimal technical skill must be able to setup and run the system themselves.
2. The cost of the overall project must be comparatively low, or at least less than hiring a software developer.
3. The system must decrease the overall feedback loop, allowing data to be collected in an analyzable format in a short time-frame.
4. The data must be replicable and of a high-quality.

With these requirements in mind and utilizing design science methodology (March and Smith 1995), a working prototype was created. Through multiple iterations of development and evaluation, three Python scripts were written to support basic data exploration (twitter-count.py), collection (twitter-search.py) and aggregation (twitter-merger.py). After running each of these scripts, data is collected and combined into a single file which can be exported to analysis environments such as Power BI or SPSS. A web-based platform has been built on top of these scripts to enable ease of use. Non-technical researchers can simply download the software and use the web interface to explore, collect and aggregate data without having to interact with code or configuration files. This system is entirely open-source and available for free. The software can be found here: https://github.com/shohil-kishore/twitter-data-toolkit.

Implementation evaluation

While the artefact is a prototype, it has been used iteratively for small projects, and as part of two larger studies at the University of Auckland and one study at the University of Oxford. Data collection at the University of Auckland was carried out by a technician where 100,000 and 80,000 Tweets were collected for each project, respectively. At the University of Oxford, data was collected by a student with a non-technical background. The process for collecting and aggregating data proved to be reasonably straightforward, as the student at the University of Oxford was able to collect 20,000 Tweets for less than $100 USD on a tight deadline. Thus, we were able to fulfill our core requirements of requisite functionality, low cost, low latency, high quality and flexibility to cater to a wide range of requirements and different levels of expertise, ultimately retrieving data that satisfied non-technical researcher requirements.

According to March and Smith (1995), design science evaluation involves the process of rigorously examining if the evaluand achieves its stated purpose. In our case, the system was designed with five core requirements in mind, all of which have been met and tested against. As the system is currently a prototype, we intend to build on it through multiple phases of development and evaluation, and continue expanding system functionality while utilizing design science principles.

Conclusion

When data is required for a SMR project, difficulties such as cost, project flexibility, technical personnel requirements, data quality and data latency arise. These challenges cause delays and introduce barriers to entry for researchers looking to simply explore the validity of an idea, particularly those examining a time-critical occurrence. We examine the impact of these impediments in the context of a full-scale study that involves iterative data collection, cleaning and analysis.

To enable idea exploration and formation, we designed a system utilizing Twitter's latest API to enable non-technical researchers to collect full-fidelity data themselves at a low cost. While Twitter does not represent all social media interactions online, it is widely utilized and therefore valuable. We intend to add additional functionality to the Data Toolkit, primarily to generalize it to other social media platforms, such as Reddit, improve system flexibility without increasing complexity and add basic data visualization capabilities to enable exploration.

We also implemented the Data Toolkit, a web-based prototype based on this design that minimizes these issues when collecting and analyzing social media data, enabling researchers to get feedback on a project idea in a shorter time-frame. Finally, we evaluated the artefact through multiple implementations at multiple universities. Our findings were positive, indicating that our design and prototype alleviated the major issues faced by non-technical researchers when evaluating an idea and carrying out SMR.

REFERENCES

Bošnjak, M., and Oliveira, E. 2012. "TwitterEcho - A Distributed Focused Crawler to Support Open Research with Twitter Data," in *Proceedings of the 21st International Conference on World Wide Web*, Lyon, France: ACM, pp. 1233–1240.

Boyd, D., and Crawford, K. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon," *Information Communication and Society* (15:5), pp. 662–679. (https://doi.org/10.1080/1369118X.2012.678878).

Bruns, A., and Stieglitz, S. 2012. "Quantitative Approaches to Comparing Communication Patterns on Twitter," *Journal of Technology in Human Services* (30:3–4), pp. 160–185. (https://doi.org/10.1080/15228835.2012.744249).

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996. "From Data Mining to Knowledge Discovery in Databases," *Al Magazine* (17:3), pp. 37–54. (https://doi.org/10.1145/240455.240463).

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 2002. "The KDD Process for Extracting Useful Knowledge from Volumes of Data," *Communications of the ACM* (39:11), pp. 27–34. (https://doi.org/10.1145/240455.240464).

Ghosh, S., Zafar, M. B., and Bhattacharya, P. 2013. "On Sampling the Wisdom of Crowds: Random vs. Expert Sampling of the Twitter Stream," in *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management*, pp. 1739–1744. (https://doi.org/10.1145/2505515.2505615).

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., and Fdez-Riverola, F. 2013. "Web Scraping Technologies in an API World," *Briefings in Bioinformatics* (15:5), pp. 788–797. (https://doi.org/10.1093/bib/bbt026).

Hackathorn, R. 2002. "Current Practices in Active Data Warehousing."

Heydon, A., and Najork, M. 1999. "Mercator: A Scalable, Extensible Web Crawler," *World Wide Web 2.4* (2:4), pp. 219–229.

Katal, A., Wazid, M., and Goudar, R. H. 2013. "Big Data: Issues, Challenges, Tools and Good Practices," in *6th International Conference on Contemporary Computing, IC3 2013*, Noida, India: IEEE, pp. 404–409. (https://doi.org/10.1109/IC3.2013.6612229).

Koepp, C. 2013. "TwitterSearch." (https://github.com/ckoepp/TwitterSearch, accessed March 22, 2019).

Lomborg, S. 2012. "Researching Communicative Practice: Web Archiving in Qualitative Social Media Research," *Journal of Technology in Human Services* (30:3–4), pp. 219–231. (https://doi.org/10.1080/15228835.2012.744719).

Lomborg, S., and Bechmann, A. 2014. "Using APIs for Data Collection on Social Media," *Information Society* (30:4), pp. 256–265. (https://doi.org/10.1080/01972243.2014.915276).

Mahrt, M., and Scharkow, M. 2013. "The Value of Big Data in Digital Media Research," *Journal of Broadcasting and Electronic Media* (57:1), pp. 20–33. (https://doi.org/10.1080/08838151.2012.761700).

March, S. T., and Smith, G. F. 1995. "Design and Natural Science Research on Information Technology," *Decision Support Systems* (15:4), pp. 251–266. (https://doi.org/10.1016/0167-9236(94)00041-2).

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. M. 2013. "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose," in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, Cambridge, MA: AAAI, pp. 400–408.

Myllymaki, J. 2002. "Effective Web Data Extraction with Standard XML Technologies," *Computer Networks* (39:5), pp. 635–644. (https://doi.org/10.1016/S1389-1286(02)00214-1).

Ruths, D., and Pfeffer, J. 2014. "Social Media for Large Studies of Behavior," *Science* (346:6213), pp. 1063–1064. (https://doi.org/10.1126/science.346.6213.1063).

ScrapeStorm. 2019. "ScrapeStorm: AI-Powered Visual Web Scraping Tool." (https://www.scrapestorm.com, accessed February 21, 2019).

Scrapy. 2019. "Scrapy, a Fast High-Level Web Crawling & Scraping Framework for Python." (https://github.com/scrapy/scrapy, accessed February 21, 2019).

Thelwall, M., and Stuart, D. 2006. "Web Crawling Ethics Revisited: Cost, Privacy, and Denial of Service," *Journal of the American Society for Information Science and Technology* (57:13), pp. 1771–1779. (https://doi.org/10.1002/asi.20388).

Tornes, A. 2017. "Introducing Twitter Premium APIs." (https://blog.twitter.com/developer/en_us/topics/tools/2017/introducing-twitter-premium-apis.html, accessed March 23, 2019).

Twitter. 2017. "Python Twitter Search API." (https://github.com/twitterdev/search-tweets-python, accessed March 23, 2019).

Twitter. 2019a. "Unleash the Power of Twitter Data." (https://developer.twitter.com/en/enterprise, accessed March 23, 2019).

Twitter. 2019b. "Configure Search Tweets: Full Archive." (https://developer.twitter.com/en/pricing/search-fullarchive, accessed February 19, 2019).

Twitter. 2019c. "Search Tweets - Premium Search APIs." (https://developer.twitter.com/en/docs/tweets/search/overview/premium.html, accessed February 18, 2019).

Zeng, D., Chen, H., Lusch, R., and Li, S. H. 2010. "Social Media Analytics and Intelligence," *IEEE Intelligent Systems*, pp. 13–16. (https://doi.org/10.1109/MIS.2010.151).