International Journal on Data Science and Technology
2019; 5(2): 40-44
http://www.sciencepublishinggroup.com/j/ijdst
doi: 10.11648/j.ijdst.20190502.12
ISSN: 2472-2200 (Print); ISSN: 2472-2235 (Online)

Social Media Data Extraction Method Benchmarking Comparison

Zhenhuan Sui

Department of Integrated Systems Engineering, The Ohio State University, Columbus, USA

Email address:
suizhenhuan@gmail.com

To cite this article:
Zhenhuan Sui. Social Media Data Extraction Method Benchmarking Comparison. *International Journal on Data Science and Technology.* Vol. 5, No. 2, 2019, pp. 40-44. doi: 10.11648/j.ijdst.20190502.12

Received: April 7, 2019; **Accepted**: August 13, 2019; **Published**: August 28, 2019

Abstract: Social media has become more and more widely used nowadays. As the most popular media, a lot of information spread through Twitter, especially given the fact that U.S. President Trump has used Twitter as his main official free news publication outlet. Therefore, social media platforms like Twitter have become the important sources to extract information and then the information could be further analyzed through text analytics models for decision-making problems. In this paper, we first investigate several text analytics methods and then multiple tweets retrieving methods/software will be investigated: Twitter Analytics, Application for Twitter, Python plus Tweepy, and Next Analytics. Seven criteria related to features are applied to compare the methods for ease of use, extraction timing and capability to accommodate big data. Given that our results may be approximate because we might not be able to observe all the capability and features of the software, our results show that Python plus Tweepy method is the most ideal one when applying to big data projects (millions of tweets or above) and real time text data extraction. Next Analytics is the software that could retrieve historical text message in a more convenient way through Excel and is able to trace back further in time period, which could give much better capabilities in social media analysis.

Keywords: Natural Language Processing, Text Analytics, Twitter Analysis, Social Media, Software Analysis, Big Data Analysis

1. Introduction

Twitter is an online social network, a micro blogging service where "tweets" are issued by participants to followers. Twitter appeals to a broad range of users covering every age group, and is internationally one of the most popular social media platforms. tweets are short, less than or equal to 140-characters. Status updates appear on users timelines. Similar to a Facebook, there is a home screen but only with short status updates. The contents of a tweet could include: mentions of other users, hash tags, URL's, and pictures or other attached media.

Previous literature relating to extracting data from Twitter and social networks have tended to focus on specific software languages and advanced usage (e.g. python) and not commercial packages [2, 3]. Our focus here is to review selected commercial offerings relevant to non-experts in python or other programming languages.

Studies have provided a workshop on using Twitter for multiple objectives including prediction [4]. A user may favorite (similar to "like") or retweet, which forwards an original user's tweet via the new user's account. Upon logging in to Twitter, a user sees their home timeline, tweets from users they follow listed chronologically. Clicking on a specific user will lead to a user timeline, the collection of tweets written by them. A user must be "following" another user so that their tweets to show up on their home timeline. Twitters following model is asymmetrical; following someone does not mean they need to follow you back. Celebrities have millions of followers. Twitter is a useful resource for data mining since tweets are public information. There is no privacy concern. Also, the enormous user base provides a mass amount of information for marketing, political targeting, sentiment analysis, monitoring, prediction, e.g., of the stock market, or other possible objectives.

The primary purpose of this article is to help the readers and new users select an approach for extracting tweets for analysis. Also, criteria are established that can help users and developers evaluate extraction software. In Section 2, the four extraction methods that are considered are described. Admittedly, there is arbitrariness in selecting the four but we focused on those which currently have the highest google page ranks. Then, seven criteria are described and used to evaluate the four methods in Section 3. Section 4 summarizes the findings and implications as well as suggests opportunities for future research.

2. Data and Methods

2.1. Data Analytics Methods

Text data is usually semi-structured or unstructured, but the computer cannot directly calculate the text data, and it needs to be converted into the computable model [5]. The most commonly used text representation model is the Vector Space Model (VSM), and in VSM, after the word segmentation, the text first removes the stop words, and then counts the word frequency, and finally expresses it in the vector form. In addition, the text is simplified to the so-called BOW (bag of words), ignoring the word order, grammar and syntax of a text, which is only regarded as a set or a combination of words, in which each word is independent. At present, BOW has become a common mode of the text categorization [6, 7].

2.1.1. Bayesian Classification

The Bayesian classification algorithm is based on the Bayesian theory. It is a text classification algorithm using the prior probability and the conditional probability. It has the characteristics of the simple implementation, the high accuracy and the fast speed. The Bayesian algorithm is based on the assumption of independence. That is, the effect of an attribute on a given class is independent of the value of other attributes. The independence hypothesis is too restrictive, and it is often not valid in the practical application, so in many cases its classification accuracy cannot be guaranteed [8].

2.1.2. k-Nearest Neighbor

The k-nearest neighbor algorithm is a case-based negative learning algorithm. The idea of the algorithm is to count k most similar sample categories of a sample in the feature space, and then determine the category of the sample to be classified by the weighted voting. KNN classifier only stores instances and traverses training samples for each unknown input, so its algorithm efficiency is very low when dealing with a large number of the data to be classified [9].

2.1.3. Support Vector Machine

The support vector machines (SVM) is a machine learning technology developed by V. Vapnik and his Bell Laboratory team. SVM is a linear classifier, which adopts the principle of the structural risk minimization [10]. Its characteristic is that it can simultaneously minimize the empirical error and maximize the geometric edge area. Finally, the classification problem is transformed into solving the optimal decision-making hyperplane problems. This method belongs to the category of the statistical learning theory, which studies the machine learning law under the small samples. It has the good adaptability to small samples, overcomes the phenomenon of "over-learning", and has the relatively good performance indicators. The two most important factors affecting the classification performance of SVM are the error penalty parameters and the kernel functions [11].

2.1.4. Neural Network

The neural network is a simulation of the nervous system. In the text categorization, a neural network consists of a group of neurons, whose input units usually represent terms, and the output units represent the categories or the class interest, and the connection weights of the neurons represent the conditional dependencies. The simplest neural network for the text categorization is the perceptron. The perceptron is actually a linear classifier, which converts the classification problem into a correction problem for the error classification. By iterating and updating all the training instances, the number of the error classifications is lower than a certain threshold, and the weight of each input component connected to the perceptron is obtained [12, 13].

2.2. Twiter Extraction Methods

To retrieve Twitter text data for feeding the topic models, we have investigated the following four methods.

2.2.1. Twitter Analytics

The Twitter analytics dashboard [14] is an add-on for users with advertiser status. These users can find detailed information on how their outgoing tweets are performing based on a few different criteria. The Twitter analytics tool provides public data only. If a user account has privacy settings and they do not follow the advertisers, their data will not be provided. This makes retrospective analyses for the previously unfollowed (apparently) impossible.

The dashboard is an intuitive tool for social media marketers. On the dashboard, the user can see a maximum date range of 91 days of their past tweets performance. The dashboard shows the user the top ten accounts their followers follow, ranked by percentage. This information can be used to better understand what kind of information your followers are interested in on Twitter.

The data provided by Twitter includes a tweet impressions, link clicks, retweets, detail expands, favorites, embedded media clicks, user profile clicks, and replies. An impression is the number of times the tweet is read. Link clicks refers to the number of times the URL was clicked. Detail expands is the number of times the tweet was clicked on the view details. A graph displays the past month of data, if a user wants to compare different months the data can be downloaded to a CSV file. This information would help the user determine which of their tweets was most effective in reaching their audience, or whether a certain time was most effective. There is also a feature that tracks follower increases or decreases and information on follower's location and gender. Tracking the audience's interests is a key feature of this method.

2.2.2. Application for Twitter

An analytics application called "Follow the Hashtag" [15] has many useful features. The main dashboard section includes: total tweets, total impressions, total potential tweet impressions in followers timelines, and results from multiplying each contributor's keyword repetitions and it´s number of followers and adding all contributors potential impressions. Other outputs include the total audience, total potential audience, result of adding each contributor followers number, impressions / audience, and impressions per user of a tweet with searched keyword.

Algorithms are utilized to analyze Twitter contributors' gender and percentage of males and females. Follow the hashtag allows users to export data to a PDF or CSV. Both will produce detailed sheets including: summary, top tweets, top users, gender, reach, charting data, geolocation, and stream data (tweet content, country etc.). The reports also include the best hour of the day, best day of the week for a hashtag's performance. An influence section is also included showing user by keyword repeats, top users by influence score, and top users by keyword. An influence score shows the largest contributors in a searched keyword. There is a historical data feature where the user can recover tweets *only up to 60 days old*. A Twitter picture analysis is available which shows all the pictures related to a Twitter search, useful for picture based Twitter, or to get a general overview. An aggregated key repeats chart shows aggregated repetition values over time of the most repeated words related to your search. This chart shows the evolution of each related keyword discovering how a searched keyword is related to others over time. Overall, the usability and built-in analytics shine but the extraction is somewhat limited.

2.2.3. Python Plus Tweepy

Python is a widely used programming language that is easy to use, and effective for text analytics. In particular, it is used in multiple software programs for exporting data from Twitter. Tweepy [16] is a python library that uses Twitter's application programming interface (API) to access public data. An API is a way for other programs to enter a given program. To access Twitter's streaming API, the user must create an app on Twitter. Once the Twitter API access is granted the user needs the API key. This API must be secret with an access token and access token secret. A script file is utilized to access live tweets, the file can search for specific keywords or usernames. The file can be run for any length of time depending upon the amount of data the user wishes to collect. This program can only pull current, live tweets, no historical data can be provided. The data can be exported to a CSV using a line of code. In our experiments we found that the interface was difficult to understand and the derived CSV was somewhat miss-parsed, i.e., the fields were not cleanly usable in all cases.

2.2.4. Next Analytics

Next Analytics [17] is a paid software program used for video and social media analytics, including Google, Facebook, Twitter, Instagram, and YouTube. Here, we focus on Twitter analytics. Next Analytics for Twitter is primarily associated with Microsoft Excel as an add-in function. Next Analytics could extract all the tweets going back to the start of Twitter itself. Users can select to extract tweets from their own account, followers' accounts, or any specific account. Yet, the extraction is on an account basis unlike, e.g. Twitter analytics. The output from Next Analytics is also in the format of excel including the information account name, tweet text, post time, account favorite number, account friend number, account follower number, and retweet number. The formatting of the output is well done so that not much effort is needed to clean up after the extraction before analysis.

3. Benchmarking

3.1. Criteria

The first criteria relates to which sources of tweets can be tracked. Message sources show where the data is from. For both Python plus Tweepy and Follow the Hashtag application methods, when the searching keywords are inputted, the two methods will search for the tweets with the keyword across the world. For the Twitter Analytics and Next Analytics methods, the data and analyses will only be from the activities of the user account. Moreover, Next Analytics also has the function for users to choose the data sources: account users themselves, followers, friends, or even specific accounts the users want to analyze. Therefore, if the analysis is targeted to specific users, the Twitter Analytics and Next Analytics methods should be applied.

Second, the analysis duration gives out the time range of data. For the Python plus Tweepy method, the data is real time. The program will search for and output the tweets with the searching keywords from the time point of the start of the program until the stop command is inputted. As long as there is a tweet with the keyword posted online, the program will output it instantaneously. For analysis duration, Twitter Analytics and Follow the Hashtag application methods analyze the data in history for the time range specified by users. The time range can look back for up to two years. For Next Analytics, the software could extract data back to whenever Twitter keeps for the account holders.

Third, some extraction methods show the individual tweets and others do not. All four extraction methods permit the outputs of individual messages but Follow the Hashtag emphasizes the statistics and meta data making viewing individual tweets less direct.

All of the software permit significant customization and flexibility. The fourth criteria relates to the level of detail of derived outputs as subjectively assessed in our testing. We find that Follow the Hashtag and Python plus Tweepy offer relatively sophisticated output, far beyond extracting the tweets themselves. The fifth relates to the ease for which summary statistics about the tweets can be obtained. These are the numerical portions of the outputs.

Again, some of the software such as Follow the Hashtag permit the easy derivation of detailed statistics. The sixth criteria is our subjective assessment of user friendliness. Here, we find that the programming environment is significantly less friendly than the others which have fairly standard graphical user interfaces. Finally, we also subjectively assessed how easy it is to derive outputs of various formats. Again, all of the software permits significant customization. Yet, we focus on emphasis and ease and find much greater potential for Python plus Tweepy than the others.

3.2. Comparison

In this section, the four different extraction methods are compared using seven criteria. The results are shown in Table 1. Of all of the criteria, the ability to extract tweets historically (criteria 2) is the most important in our applications. Therefore, Next Analytics shines for our needs. Also, we ourselves are capable of producing summary statistics so the strength of Follow the Hashtag is less relevant. The extreme potential for customization of Python plus Tweepy also makes that software relevant for consideration.

Table 1. *Comparison Matrix.*

	Twitter Analytics	Follow the Hashtag	Python plus Tweepy	Next Analytics
Message Sources	Users/followers	Whole network	Whole Network	Users/followers/friends /specific accounts
Analysis Duration	History (relatively limited apparently)	History (relatively limited apparently)	Real time	History (can go back for multiple years)
Displaying Message	Yes	Only partially	Yes	Yes
Output information detail level (message, like, forward)	Relatively limited	Extensive	Extensive	Relatively limited
Summary Statistics	Yes	Yes	No	No
User Friendly	Yes	Yes	No	Yes
Output format	Excel	Excel	Many format (txt, Exce)	Excel

3.3. Applications and Industry Usage

In this section, we propose suggestions on how the various software might support activities in different industries. Each of the software programs has a strength, even Twitter Analytics might offer minimal installation. As Twitter is more and more popular as the means of communication and information publication, these alternative extraction methods could be used by different users to extract data and information for business needs.

Python plus Tweepy is a method extracting real-time original data and has a great ability handling large-scale data. Hence, this method could be applied to the cyber security industry. Hypothetically, high end users such as the Department of Defense (DoD) could apply this method to extract keywords posted on Twitter in real time. As long as the terrorists publish a tweet with dangerous messages, the DoD could monitor the account activity and take appropriate actions. Moreover, this method satisfies the business need for high-end news-based trading in high-frequency trading (HFT) on Wall Street. HFT traders could use Python plus Tweepy to track company names, key words, and trading news on Twitter at any given time. For example, the Wall Street Journal posts a message that profit of Google this year goes up. The trader will use the Python plus Tweepy method to track the Wall Street Journal Twitter account and “Google’s profit goes up” information. Then, the Tweepy software can extract information in real time and transfer it to another program which will identify the keywords and semantics and further process the information to output the command on buying Google stocks. All these processes are carried out within microseconds or even nanoseconds automatically on computers. Therefore, for the business objective of real-time information and fast processing on the raw data, this method is the best.

For the Twitter Analytics method, its message source is primarily from the activities of the user account and the data statistics that is available directly to the users. Therefore, the result could be directly used by marketing professionals. Like the advertisements on Gmail and Facebook accounts, marketing departments of retailing companies could use this method to get the posting message of specific customers and doing further analysis with keywords. Then, the companies would know what products the customer may be interested in and target users to send the corresponding advertisements.

The Follow the Hashtag application method could also be applied in a similar way with more built-in statistical information but also more installation burden. As this application also provides the information about the best hour of the day, best day of the week for hashtags' performance, retailers could use this function to know when their product related keywords are most active. Then, they could target their advertising and sales forces on those active time points for better sales efficiency. Moreover, because its message source is from the whole Twitter network, media or fashion industries could benefit from it. The media or fashion industries could set the statistical analysis duration to be only within recent months. Then, the statistical analysis will give the most popular keywords in this period of time. The media or fashion industries could use the results to analyze news or fashion trends and in turn customize their business to accommodate the public’s needs and trends.

For Next Analytics, because it has many functions analyzing Google, Facebook, Twitter, Instagram, and YouTube, it could be applied to a range in analytics for different media. Moreover, because it performs the best when analyzing historical data with user-friendly interface, it is a good method in extracting and analyzing historical tweet text. Moreover, the output also includes the total retweet number [4], and the number of retweets can be used as an important indicator in the prediction model for social events and changes. Therefore, if users want to use Twitter to analyze historical data for cyber security industry, media or fashion industries, or even use the results to analyze whether a Hollywood movie will be a hit, Next Analytics method should be selected.

4. Conclusions and Future Work

This paper analyzed four popular methods of extracting tweets and data mining from Twitter based on seven criteria. We also provide recommendations for different types of user related to different hypothetical industry usages. For example, Twitter Analytics method is good for business needs on specific customers targeting and customer behavior predicting. The Follow the Hashtag application method is good for public trend and topic analysis, which could aid for users desiring significant statistical capabilities. Python plus Tweepy has a significant advantage on large scale raw data and information extracting. Therefore, it could satisfy the business needs of fast and large raw tweet data extracting and processing by relatively sophisticated users. Finally, Next Analytics offers relatively good capability for extracting historical data. Hence, it could be used for building forecasting models or reconstructing historical events.

There are numbers of opportunities for future research. First, commercial software is always changing so the analysis here could be repeated with criteria that expand as all capabilities grow. Yet, users will have specific needs and we found very different abilities to meet those needs. For example, we found it difficult to study events over one year prior with software other than Next Analytics. Such differences, if they persist, will motivate updates to the results. Second, specifications for extraction software can be developed as standards by international organizations. The ability to generate helpful summary information from tweets will continue to grow and the need for analysis tools and related standards will grow also.

References

[1] Allen, T. T., Sui, Z., & Parker, N. L. (2017). Timely decision analysis enabled by efficient social media modeling. Decision Analysis, 14 (4), 250-260. https://doi.org/10.1287/deca.2017.0360.

[2] Russell, M. A. & Russell, M. (2011). 21 Recipes for Mining Twitter. O'Reilly Media, Inc.

[3] Moujahid, A. (2015) An Introduction to Text Mining Using Twitter Streaming API and Python. Data Analytics and More. N. p., n. d. Web. 04 May.

[4] Zaman, T. R., Herbrich, R., Gael, J. V., & Stern, D. (2010) Predicting information spreading in Twitter. Workshop on computational social science and the wisdom of crowds, nips 104 (45), 17599-601.

[5] Allen, T. T., Sui, Z., & Akbari, K. (2018). Exploratory text data analysis for quality hypothesis generation. Quality Engineering, 30 (4), 701-712.

[6] Porter, M. F. (1980) An algorithm for suffix stripping. Program. 14 (3), 130-137.

[7] Sui, Z. (2019). Social Media Text Data Visualization Modeling: A Timely Topic Score Technique, American Journal of Management Science and Engineering. 4 (3), 49-55. doi: 10.11648/j. ajmse.20190403. 12.

[8] Wang, Y., & Liu, H. (2013) Advances in the Machine Learning Methods, Wireless Internet Technology, 7, 89-90.

[9] Zhan, P. (2014) Talking about the Machine Learning Method, Network Security Technology and Application, 1, 145-146.

[10] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20 (3), 273-297.

[11] Li, B., Cong, Y., Tian, Z., & Xue, Y. (2014) Prediction and virtual screening of the selective inhibitors of MMP-13 to MMP-1 based on the molecular descriptors and the machine learning methods, Acta Physico-Chimica Sinica, 1, 136-137.

[12] Zha, Y., Sun, C., & Wang, K. (2015) Research on the Tax Loss of the Real Estate Industry Based on the Micro-data -- Empirical Analysis Based on the Machine Learning Method, China's Prices, 9, 109-110.

[13] Sun, C., & Wang, C. (2015) Application of the Machine Learning in the Credit Risk Prediction and Recognition, China's Prices, 12, 101-102.

[14] Twitter Analytics 2015. “Twitter Analytics”. https://analytics.twitter.com/about, N. p., n. d. Web. 05 May.

[15] Followthehashtag 2015. “Followthehashtag // Twitter Keyword Search Analytics, Influence, Geo Content Analysis Tool, and Much More.” https://www.followthehashtag.com/, N. p., n. d. Web. 04 May.

[16] Tweepy 2015. “Tweepy”. http://www.tweepy.org/, N. p., n. d. Web. 05 May.

[17] Next Analytics 2015. “Next Analytics”. https://www.nextanalytics.com/, N. p., n. d. Web. 05 May.