IJCSN - International Journal of Computer Science and Network, Volume 6, Issue 6, December 2017
ISSN (Online) : 2277-5420
www.IJCSN.org
Impact Factor: 1.5

Research Aids for Social Media Analytics

[1] Anita Kumari Singh; [2] Mogalla Shashi

[1] Research Scholar,
Department of Computer Science & Systems Engineering,
Andhra University, Visakhapatnam, India

[2] Professor,
Department of Computer Science & Systems Engineering,
Andhra University, Visakhapatnam, India

Abstract - Social media analytics helps us get simplified insights on data generated on various social media platforms by millions of people through variety of activities. The core of analytics is identifying the patterns and activities on these platforms so that social media professionals can make better social media strategies. This paper introduces the general approach to social media analytics, and then elaborates on the various phases of analytics. Essentially, identifying credible sources of mineable social media content is very important. The applications for collecting data are authenticated using OAuth and proceed for collecting the required contents suitable for the purpose of analytics using APIs and built-in language libraries. Data is pre-processed and data modeling techniques are applied to gain valuable insights. Presenting the results in simple understandable form using appropriate visualization techniques is again important to interpret the results.

Keywords - *Social Media Analytics, open source tools.*

1. Introduction

Social networking platforms are the most favorite sites for individuals to express their views and share their thought openly to the world. Social Media Analytics is gathering data produced through Social Media platforms like Facebook, Twitter, LinkedIn, WhatsApp, Wikia, YouTube, Pinterest, Instagram, Tumblr, Reddit, Snapchat, Gab, Google+, Baidu Tieba, VK, WeChat, Weibo and many others in the form of comments, tweets, posts, likes, shares and links for extracting valuable information from the enormously large data using social media analytics tools for taking right business decisions. Analytics for Social Media also involves development and evaluation of tools and frameworks that collect, monitor, analyze, summarize, and visualize social media content. Social Media Analytics is commonly used to learn customer sentiment and is helpful in gaining insights on product reviews and skim that data for creating better marketing strategies and improved customer service. Researches on Social Media broadly concentrated toward two main categories Social Media Text Mining and Social Media Graph Mining. Social media text mining has attracted numerous researchers and most of the work in this paper concentrates on analytics for social media textual content.

1.1 General Approach for Social Media Analytics

The general approach for social media analytics starts by identifying the credible and mineable source of information form the diverse social media platforms used for Microblogging, Blogging, Community-based Question Answering (CQA), Chats, Forums, Media Sharing, and Hybrid Applications which generate huge volumes of noisy, distributed, unstructured and dynamic data. The next step is getting authentication from the websites for scraping data, as we are given access to extract the data while protecting the account credentials at the same time. The next big step is to extract the right content form these platforms which suits the purpose of analytics using APIs and built in language libraries. Preprocessing and cleaning the data is an important phase which should be performed meticulously because high quality data is essential for better analytics. Data Modeling and Analytics is applied to the cleaned up data to build models for social media analytics. Finally graphically visualizing the results to interpret the analytics performed is important to dispatch the obtained results in a concise and simplified form. A variety of prevailing open source tools and techniques are available to graphically visualize the results in different formats.

The current work focuses on all these phases and lists out various open source tools which aid in all the phases. The rest of the paper is organized as follows; section 2 briefs the importance of identifying good sources of data. Section 3 is about authentication using OAuth, section 4 deals with extracting data from social media platforms. Section 5 presents pre-processing techniques and tools while section 6 is on analytics using social media data. In section 7 data visualization techniques with emphasizes on open source tools for visualization is presented.

2. Sources of Data

Abundance of freely available online data has attracted many researchers to perform various levels of analytics. Data from social media sites have been used in varied subject areas like anthropology, psychology, linguistics, and medicine [1]. Social media analytics is commonly used for taking better business decisions, reputation management, marketing, customer service, sales and many others. Social media content include blogs, posts, comments, tweets, images, videos, links, network etc. Most of the academic researchers have used unstructured textual data found in tweets and facebook posts, apart from the other extractable contents. Twitter being a micro blogging service attracts billions of users and its 140 character limit on tweets allows spread of data in a real time. On an average of around 6,000 tweets are produced on Twitter per second, which corresponds to over 350,000 tweets sent per minute and 500 million tweets per day and sums up to around 200 billion tweets per year. The live Twitter Usage Statistics are available on the webpage *internetlivestats.com.* All this add up to the wealth of online data and gives us good opportunities to build models which convert this data into information.

3. Authenticating with Social Media Websites

Most of the social media platforms allow access to their data using APIs but we need to authenticate ourself to access the data using OAuth. OAuth is an open standard for access delegation, commonly used to get permission for websites or applications to access information on other websites without revealing the credentials for privacy reason [2]. As per *outh.net,OAuth* is defined as an open protocol to allow secure authorization in a simple and standard method from web, mobile and desktop applications [3]. Using OAuth is a very simple way to interact with protected data on the websites while it being a safer and more secure way for people to access data. The specific steps to authenticate, the secret keys and access tokens may vary from social media platform to platform but the general procedure is the same [4].

4. Extracting Data

After we go through the authentication process we can now link up our application with the social media platforms and start collecting data. Most of the prominent programming languages have built in libraries and tools, to support data extraction from these platforms which are discussed in the next section.

4.1 Collecting tweets

Twitter allows you to interact with its data that is tweets and several other attributes for tweets. The Variants of Twitter API are

- Search API which queries twitter for tweets based on a specific keyword and returns a collection of relevant tweets matching to the query.
- Streaming API a real time stream of tweets filtered by userID, keywords, geographic location or random sampling.
- REST API is used to retrieve a portion of the most recent tweets of a particular twitter user.

These APIs generally produce output from twitter in a JSON format. There are a plenty of freely available third party tools which perform social media analytics in various capacities ranging from just crawling data, capturing social interactions, preprocessing, analyzing the sentiment, storing and visualizing [5] whose details are given in the next section.

Built-in libraries for Twitter

- tweepy :Twitter for Python
- twython: It’s a pure Python wrapper for the Twitter API to collect both normal and streaming data in the form of tweets.
- twitteR: Is a R based twitter client which provides an interface to the Twitter API.

- Twitter4J: Is an unofficial Java library for the Twitter API to integrate your Java application with the Twitter service.

- Apache Flume: Though not a library it can be used to extract data from Twitter and store it in Hadoop Distributed File System (HDFS) for handling large volume of data through big data technologies.

- Similarly we can access twitter using scripting language like Java, PHP, Ruby, Hadoop Scripting Language (HIVE) and many others.

4.2 Collecting Facebook data

The best way to extract data externally is by using Facebook's Graph API apart from other built-in tools to scrape Facebook data.

- The Graph API is the primary way to extract data in Facebook platform. Facebook Graph API allows your application program to read and write Facebook Social Graph. It provides data in the form of text as well as graphs.
  Built-in libraries for Facebook

- We can make use of Graph API and scrape data using python.

- Rfacebook: The package Rfacebook lets you to access Facebook data using R programming.

- PHP library: The facebook SDK is used for text and graphs from facebook using PHP for web page development.

- Apache Flume: We can configure Apache Flume to automatically gather Facebook data using Graph API.

5. Preprocessing and Data Cleaning

The most important phase after collecting the required quantity of data is to preprocess and clean it to make it ready for all the analytics. Various preprocessing methods can be applied on the textual data generated from the social media platforms. Some of the most common pre-processing techniques on textual data include

- Tokenizing: Tokenization is the task of chopping up a sentence or document into pieces, called tokens and throwing away punctuation at the same time.

- Stemming: Stemming is reducing inflected (derived) words to their word stem its base or root form (e.g., cats to cat).

- POS tagging: Part-of-Speech tagging (POS tagging) is the process of marking words in a corpus to a particular part of speech, based on its definition and context.

- Named Entity Recognition (NER): NER is used to locate and classify named entities in text to pre-defined categories like names of persons, place, organizations, quantities, monetary values, percentages, etc.

- Normalizing: Normalization is canonicalizing tokens such that matches occur despite differences in character sequences of the tokens. Standard way to normalize is to implicitly create equivalence classes.

- Stop Words Removal: The most common words including articles and prepositions (e.g., and, an, the, he) which appear to be of little value in analytics is excluded from the vocabulary.

Additional support for preprocessing textual data generated from microblogging platforms particularly Twitter is listed below

- Filtering English only tweets: While extracting data we can restrict out search space to only English language tweets.

- Retweets: A Retweet is a re-posting of a Tweet. One can Retweet their own Tweets or Tweets from someone else. Retweets or all tweets starting with RT should be removed before proceeding to the next stage.

- URL removal: A lot of users include URLs in their tweets which complicate the sentiment analysis process. To remove url's from the tweets we can either replace them with a string "url" or delete them from the tweets.

- Remove of special characters by replacing with strings like “tag” for @, “excl” for !, “ques” for ?, and others as “symb” for .,:,;,+,-,=,/,etc. [6].

- Emoticons: Emoticons are an integral part of tweets; we can extract the emoticons from the tweets for sentiment analysis or even ignore them by replacing those using suitable strings.

- Slang words translation: Tweets often contain a lot of slang words (e.g., lol, omg). These words are usually important for sentiment analysis, but may not be included in sentiment lexicons [7]. Internet Slang Word Dictionary [8] can be used to convert these slang words into their original forms and add them to the tweet.

In the next part lists out some of the automated tools for preprocessing Tweets and further perform analytics over the data.

- TrendMiner: It is an open-source framework for text processing of streaming social media data [9].
- TwitIE: TwitIE is also an open-source information extraction pipeline for social media data particularly microblogs. It performs most of the preprocessing tasks like language identification, tokenizing, part-of-speech tagging and normalizing [10].
- TwitterZombie: The software is used to gather data for a series of search phrases simultaneously using Amazon’s cloud computing platform [11] and supports basic preprocessing on collected data.
- TwitterEcho: TwitterEcho is an open source twitter crawler used to collect and preprocess data generate by user communities [12].

6. Analytics for Social Media Data

Data modeling on data gathered from social media platforms particularly depends on the domain and targeted case for performing the analysis. Several standard data modeling algorithms to perform opinion mining (sentiment mining), clustering, recommendations, anomaly/spam detection, correlations and segmentations could be used to extract only the gist of data from the enormously large collection of social media content. Techniques and tools vary depending on the type of analytics being performed as per the requirements of the organization. The most trending topic for data modeling on Social media is presented below.

6.1 Opinion Mining or Sentiment Analysis

Sentiment analysis refers to the use of natural language processing (NLP), text analytics, computational linguistics, and biometrics in identifying, extracting, quantifying and studying the affective states and subjective information [13] to capture the emotion of the speaker or writer. The simplest form of sentiment analysis is classifying the polarity of the text or sentence to be positive, negative, or neutral. Sentiment classification also makes uses of the emotional states such as "angry", "sad", and "happy". There is big range of open source as well as proprietary tools for sentiment analysis made available by different vendors. In depth analysis of the users feedback on products is supported by automated tools by all major IT giants like Google, Microsoft etc.

However this paper provides a list of open source tools that support sentiment analysis

- NLTK (Natural Language Toolkit): It is text classification process which gives the analysis of text enter by you as positive sentiment, negative sentiment, or if it's neutral in real time [14].

- R: Text Mining(TM) package has a framework for text mining applications within R. [15].

- WEKA: WEKA is an open source collection of machine learning algorithms for data mining tasks written in java [16].

- RapidMiner: Some part of RapidMiner software is open source used for simple and fast real time data science [17].

- GATE: Has generic sentiment analysis tool for sentiment analysis and Voice of the Customer [18].

- Social Mention: Social Mention is a web based link for real-time social media search and analysis [19].

- TweetStats: Is a web based link useful for analyzing your twitter statistics in the form of graphs [20].

- KNIME: Knime analytics platform is a leading open solution for data analytics, reporting and integration [21].

- OpenRefine: OpenRefine is a tool for working with messy data which help us explore large data sets [22].

7. Visualization

Visualization is a process of creating interpretable images, diagrams or animations to communicate both abstract and concrete ideas. Visualization helps us better understand data on a graphical platform. A picture is worth a thousand words. When we take the case of visualizing social media there are numerous options available for simple to complex visualization.

- Heat map: A heat map is a graphical representation of data which visualize data through variations in color. It is a combination of colored rectangles, each representing an attribute element.

- Word cloud: A word cloud (tag cloud) is a visual representation of text data which is used to perceive the most prominent terms the words which appear most frequently in the source text [23]. There are a plenty of online free tools apart from built in libraries in the programming languages to build beautiful word clouds of different shape, colors, themes and font, just to name a few *Wordle, WordClouds.com, WordItOut, TagCrowd* and many more.

- Timeline: It's a linear order of events that have occurred. Timeline visualization for social media generally uses bar chart to show the occurrences of online posts or updates at different time intervals in reverse chronological order.

- Map: The area of representing data using map visualization is developing rapidly and it has many practical applications. We can plot data on geographical maps and visualize it for various purposes. Example tweets could be drawn on the world map at the location where they were tweeted mostly.
- Affinity: Close resemblance or connections in social media content can be plotted and visualized. Example frequent posts, hashtags, people, URLs etc. are drawn as a graph to show important actors and any relationship they have among them.

- Decision Trees: A decision tree is a tree where each node divides the observations based on some feature variable such that every element in one group belongs to the same category.

Leading open source automated tools with main emphasis on visualization aspects of twitter contents are presented below

- TweetTracker: TweetTracker is a powerful tool used to track, analyze, and understand activity on Twitter build at Arizona State University [24].

- tweetXplorer: TweetXplorer is a system with effective visualization techniques[25].

- TwitInfo: TwitInfo is a powerful system for visualizing and summarizing events on Twitter after browsing a large collection of tweets [26].

- Trendsmap: It has the trending hashtags and topics from twitter mapped with it in real-time. We can visualize top trending topics globally, nationally, and in our own your city [27].

- Twitalyzer: Twitalyzer is free web-based analytics tools that allow automating, exporting, and managing twitter data to measure effectiveness, account impact, engagement, clout, and velocity of individual twitter accounts [28].

- Sentiment viz: It is web based twitter sentiment visualization project to visualize the sentiment, trending topics, heatmaps, tag cloud, timeline, map and affinity for tweets developed at North Carolina State University, United States [29].

The above mentioned tools are used to present the result of social media analysis focusing on the contents of the social media. However certain applications demand for visualization tools that present the gist of the social networks structure. Some of the tools for social network structure analysis are presented below

- NetworkX: NetworkX is suitable for operation on large real-world graphs and is reasonably efficient, scalable and highly portable framework for social network analysis [30].

- MuxViz: It is a framework for interactive visualization and exploration of multilayer networks like social media networks [31].

- iGraph: Is a library that provides options for descriptive network analysis and visualization with the power of R, Python, and C/C++.

- Gephi: Is an open source network analysis and visualization package written in Java [32].

8. Conclusion

Social media data has a lot of knowledge hidden in its huge volumes of noisy, unstructured and dynamic collection of data. Thus social media analytics has become extremely important issue for organization and modern researchers to extract promising outcomes using the emerging fields like Bigdata Analytics, Deep Learning and others. This paper is an overview of various phases involved in social media analytics, with focus on most of the open source tools useful for researchers while there a numerous proprietary tools for more complicated and real time analytics.

References

[1] Social Media Data Research and Use https://www.lib.ncsu.edu/social-media-archives-toolkit/research-and-use/research

[2] Whitson Gordon. "Understanding OAuth: What Happens When You Log Into a Site with Google, Twitter, or Facebook". Retrieved 2016-05-15

[3] OAuth https://oauth.net/

[4] Ravindran, Sharan Kumar, and Vikram Garg. Mastering social media mining with R. Packt Publishing Ltd, 2015.

[5] Goonetilleke, Oshini, et al. "Twitter analytics: a big data management perspective." ACM SIGKDD Explorations Newsletter 16.1 (2014): 11-20.

[6] Ji, Xiang, et al. "Twitter sentiment classification for measuring public health concerns." Social Network Analysis and Mining 5.1 (2015): 13.

[7] Singh, Tajinder, and Madhu Kumari. "Role of Text Pre-processing in Twitter Sentiment Analysis." Procedia Computer Science 89 (2016): 549-554

[8] Internet & Text Slang Dictionary & Translator www.noslang.com

[9] Preotiuc-Pietro, Daniel, et al. "Trendminer: An architecture for real time analysis of social media text." (2012) www.trendminer-project.eu

[10] Bontcheva, Kalina, et al. "TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text." RANLP. 2013.

[11] Black, Alan, et al. "Twitter zombie: Architecture for capturing, socially transforming and analyzing the Twittersphere." Proceedings of the 17th ACM international conference on Supporting group work. ACM, 2012.

[12] Bošnjak, Matko, et al. "Twitterecho: a distributed focused crawler to support open research with twitter data." Proceedings of the 21st International Conference on World Wide Web. ACM, 2012.

[13] Wikipedia contributors. "Sentiment analysis." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 24 Nov. 2017. Web. 9 Dec. 2017.

[14] Natural Language Toolkit http://www.nltk.org/, http://text-processing.com/demo/

[15] https://cran.r-project.org/web/packages/textmining/textmining.pdf

[16] WEKA https://www.cs.waikato.ac.nz/ml/weka/

[17] https://rapidminer.com/the-core-of-rapidminer-is-open-source/

[18] General Architecture for text engineering https://gate.ac.uk/sentiment/

[19] SocialMention http://socialmention.com/

[20] TweetStats http://www.tweetstats.com/

[21] KNIME(Konstanz Information Miner) https://www.knime.com/

[22] OpenRefine http://openrefine.org/

[23] Martin Halvey and Mark T. Keane, An Assessment of Tag Presentation Techniques, poster presentation at WWW 2007, 2007

[24] Kumar, Shamanth, et al. "TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief." ICWSM. 2011.

[25] Morstatter, Fred, et al. "Understanding twitter data with tweetxplorer." Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013.

[26] Marcus, Adam, et al. "Twitinfo: aggregating and visualizing microblogs for event exploration." Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 2011.

[27] Trendsmap: Real-time Twitter Trending Hashtags and Topics. https://www.trendsmap.com/

[28] Twitalyzer: Serious analytics for social business. http: //twitalyzer.com

[29] https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

[30] Aric A. Hagberg, Daniel A. Schult, Pieter J. Swart, Exploring Network Structure, Dynamics, and Function using NetworkX, Proceedings of the 7th Python in Science conference (SciPy 2008), G. Varoquaux, T. Vaught, J. Millman (Eds.), pp. 11–15.

[31] Muxviz http://muxviz.net/

[32] Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating

networks. International AAAI Conference on Weblogs and Social Media.