A Smart System for Twitter Corpus Collection, Management and Visualization

Gaspar Brogueira, Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal
Fernando Batista, Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal
Joao P Carvalho, University of Lisbon, Lisbon, Portugal

ABSTRACT

Social networks have become popular and are now becoming an alternate mean of communication, used to share information on various topics, ranging from politics or sports to simple aspects of everyday life. Twitter messages (tweets) are shared in real time and are essentially public, making them a useful source of information for areas such as tourism, marketing, health, and safety. This paper describes an information system that involves the creation and storage of a corpus of tweets, written in European Portuguese and published within the Portuguese territory. The system also involves a REST API that allows access to the stored information, and a web-based dashboard that makes it possible to analyze and visualize indicators concerning the stored data.

KEYWORDS

Geolocation, Information System, MongoDB, Social Networks, Twitter

INTRODUCTION

A large part of today’s society lives hyper-connected. Despite their distance, people are linked to their friends and to other people whose lives they want to follow, either through photos posted on Instagram, content shared on Twitter or Facebook, or even videos uploaded on YouTube. The existing frameworks make it now possible to widely expose interests, tastes and other characteristics through social networks, and facilitate the way someone can follow a person’s life, either by its geographical location widespread via Foursquare, by the vacation photographs on Flickr, or by the communities in which one participates. The social network, as defined by Boyd et al. (2007), is a web-based service that allows a person to build a public or semi-public profile within a defined system and manage a list of users with whom it maintains a connection. Such person accesses both to its list of connections and connections of other users by viewing their status updates. Social networks tend not only to become an environment to follow the life of friends, but also an interaction platform with companies, brands, applications or services. The importance of social networks in the contemporary society is undeniable. Common practices in social networks include advertising an event and comment or disclose an idea, making them a privileged environment to the expression of individual opinions and consequently one of the main trainers of social perception as well as of the world around us. Today, important events are often commented on social networks, even before they become “public notice” and even news agencies have had to adapt and begin using social networks as a source of information.

DOI: 10.4018/IJTHI.2017070102

Copyright © 2017, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
Despite their importance, many questions about the effect of social networks in society are yet to be properly analyzed, mostly due to the lack of extensive and flexible data collections, storing information shared on social networks during certain periods of time. The existence of extensive data collections makes it possible to properly process and analyze the information in a wider range of different perspectives. Such information constitutes a valuable resource that can be used for analyzing questions, such as: what issues, hot topics or events were shared on over the time on social networks; who are the main actors within the themes; the origin and spread of topics (timeline); what makes certain event become important in social networks; how long it takes for an event to have an impact on social networks and society; and the role of the “main players” in the spread of events. The creation of such data collections is far from a trivial problem, involving many technical obstacles, ranging from the restrictions imposed by each one of the social networks on the data access to the required dedicated and intelligent hardware and software platforms.

This paper describes an architecture for collecting, storing, analyzing, and processing tweets produced in Portugal and written in European Portuguese. The resulting Information System combines four modules to collect, store, view, and access the data. The proposed architecture overcomes most of the limitations imposed by Twitter, and makes it possible to retrieve most of the required data over the time. Based on the proposed system, we have created a massive dataset of tweets, written in Portuguese by Portuguese users. All the available information concerning a tweet is stored in our dataset, thus keeping the information suitable for a large spectrum of studies. The existing database can be easily accessed, not only by complex queries, written by a developer, but also by researchers that can use the REST API in order to focus their studies in more specific subjects. Based on a dataset achieved by the proposed architecture, an analysis over 28 Million Portuguese geolocated tweets, produced in Portugal in a 1-year period, was recently concluded. The extracted information about the behavior of the geolocated Portuguese Twitter community revealed that with this information it is possible to extract overall population indicators. The existing pipeline is a crucial component in the MISNIS platform, recently described by Rosa et al., (2016).

This paper is structured as follows: Section 3 overviews the proposed information system; Sections 4 and 5 describe the tools for collecting, storing and processing information extracted from Twitter; Sections 6 and 7 present the tools developed for access and visualization of data; Section 8 discusses the resources created within this framework, and finally Section 9 presents the main conclusions and the future work.

Related Work

The analysis of the content and information shared on social networks has been proved useful in various fields, including Politics, Marketing, Tourism, Public Health, and Safety. Twitter is amongst the most widely used social networks, making available about 500 million tweets every day (https://about.twitter.com/company), on average. Twitter provides free access to part of the information produced by its users through public APIs (Application Programming Interface), and the popularity of Twitter as a source of information has led to the development of numerous applications and to new research methods in various fields, such as: decision making (Gruber et al., 2015), sentiment analysis (Saif et al. 2012; Spencer and Uchyigit, 2012; Rill et al., 2014), trend detection (Lau et al., 2012; Kaleel and Abhari, 2015); or model training in speech recognition (Kong et al., 2014). The rest of this section overviews literature that reports Twitter as a source of information, also focusing on studies involving Portuguese tweets and data collections, and describes some of the existing platforms for data collection from Twitter.

Twitter has been extensively used as source of information for a vast number of Natural Language Processing tasks, including topic detection (Saravanan et al., 2013) and sentiment analysis (Batista and Ribeiro, 2012; Souza and Vieira, 2012; Duarte, 2013). The resultant knowledge extracted, especially from sentiment analysis, may be directly applied a vast range of domains, including the detection consumer sentiment variations about products and services (Ghiassi et al., 2013), marketing
Related Content

Prescription to Remedy the IT-Business Relationship
www.igi-global.com/chapter/prescription-remedy-business-relationship/29327?camid=4v1a

Between Individuals and Teams: Human Resource Management in the Software Sector
www.igi-global.com/chapter/between-individuals-teams/22436?camid=4v1a

E-Governance in the Developing World: An Overview on Bangladesh
www.igi-global.com/article/governance-developing-world/74061?camid=4v1a
Digital Storytelling
www.igi-global.com/chapter/digital-storytelling/136678?camid=4v1a