Cities 86 (2019) 37–50

Contents lists available at ScienceDirect

Cities

journal homepage: www.elsevier.com/locate/cities

Analysis of Twitter messages using big data tools to evaluate and locate the activity in the city of Valencia (Spain)

Angel Martín[a,*], Ana Belén Anquela Julián[a], Fernando Cos-Gayón[b]

[a] Department of Cartographic Engineering, Geodesy and Photogrammetry, Universitat Politècnica de València, C\Camino de Vera s/n, 46022 Valencia, Spain
[b] Department of Architectural Constructions, Universitat Politècnica de València, C\Camino de Vera s/n, Valencia 46022, Spain

ARTICLE INFO

Keywords:
Twitter
Big data
Apache Spark
MongoDB
Urban infrastructure

ABSTRACT

This paper presents the big data architecture and work flow used to download georeferenced tweets, store them in a NoSQL database, analyse them using the Apache Spark framework, and visualize the results. The study covers a complete year (from December 10, 2016 to December 10, 2017) in the city of Valencia (Eastern Spain), which is considered to be the third most important in Spain, having a population of nearly 800,000 inhabitants and a size of 135 km^2. The concepts of a specific event map and a specific event map with positive or negative sentiment are developed to highlight the location of an event. This approach is undertaken by subtracting the heat map of a specific day from the mean daily heat map, which is obtained by taking into account the 365 days of the studied period.

This paper demonstrates how the proposed analysis from tweets can be used to depict city events and discover their spatiotemporal characteristics. Finally, the combination of all daily specific events maps in a single map, leads to the conclusion that the city of Valencia city has appropriate urban infrastructures to support these events.

1. Introduction

Events in a city, such as concerts, sports, special days (Halloween or Black Friday) and rallies, are a social and economic factor that city authorities must keep in mind (Clark, Kearns, & Cleland, 2016), and they should be conveniently planned in suitable places. If these places do not exist, they must be built or conditioned well in advance. In the case of spontaneous events or those not known a priori, they should be conveniently studied to take the appropriate measures if they recur.

For the study of the viability of the infrastructures where events are located, big data technology can be highly useful. In general terms, the recent proliferation of big data has contributed to smart city transformation, representing digital traces of human activities (Lim, Kim, & Maglio, 2018) and offering the possibility to re-imagine and regulate urban life (Kitchin, 2014). The new information and communication technologies (ICET) related to big data, are additionally used to compute sustainable urban forms or urban designs (Bibri & Krogstie, 2017).

The referring "big data" referred to in the previous paragraph is associated with digital devices, transactions and interactions across digital networks; sensed data generated by a variety of sensors and actuators embedded into objects or environments that regularly communicate their measurements; the scanning of machine-readable objects, such as travel passes, passports, or barcodes on parcels that register payments and movements through a system; and machine to machine interactions across the internet of things or social media (Kitchin, 2014), which are the primary data inputs into this research.

With the rise of social media, people write, obtain and share information almost instantly on a 24/7 basis.

In this context, the possibility of including spatial and temporal information in the social media messages is generating a wide range of applications: disaster management for various types of hazards, such as earthquakes (Addair, Dodge, Walter, & Ruppert, 2014; Crooks, Croitoru, Stefanidis, & Radzikowski, 2013; Sakaki, Okazaki, & Matsuo, 2010), hurricanes (Huang & Xiao, 2015), tsunamis (Mersham, 2010), agricultural droughts (Enenkel et al., 2015), or floods (Restrepo-Estrada et al., 2018, Wang, Mao, Wang, Rae, & Shaw, 2018); impact on voting behaviour from the propagation of political ideologies in social networks (Correa & Camargo, 2017); or intelligent transportation systems and smart cities (Khan, Anjum, Soomro, & Tahir, 2015; Figuereido, 2017; Gao, Wang, Padmanabhan, Yin, & Cao, 2018; Kousiouris et al., 2018; Wu, Zhang, Shen, Mo, & Peng, 2018).

Social media can also offer interesting opportunities in urban studies, spatiotemporal demographic analyses, human dynamics (understood as human activities and interactions in space and time, Yuan,

* Corresponding author.
E-mail addresses: aemartin@upvnet.upv.es (A. Martín), anquela@cgf.upv.es (A.B.A. Julián), fcosgay@csa.upv.es (F. Cos-Gayón).

https://doi.org/10.1016/j.cities.2018.12.014
Received 11 May 2018; Received in revised form 29 November 2018; Accepted 17 December 2018
Available online 04 January 2019

2018), or urban planning. These factors include city dynamics over the course of the day (García-Palomares, Henar Salas-Olmedo, Moya-Gómez, Condeço-Melhorado, & Gutiérrez, 2018; Zhi et al., 2016), population mobility patterns (Blanford, Huang, Savelyev, & MacEachren, 2015; Lenormand et al., 2014; Longley & Adnan, 2016; Wu, Zhi, Sui, & Liu, 2014), identification of successful public spaces (Martí, Serrano-Estrada, & Nolasco-Cirugeda, 2017), city marketing and promotion (Zhou & Wang, 2014), impact of mega-events in neighbourhoods (Clark et al., 2016), identification of demographic groups based on user names (Luo, Cao, Mulligan, & Li, 2016), urban vibrancy (Wu, Zhang, et al., 2018) or representing the residents' perceptions of the quality and activity of the points of interest (e.g., shopping, residential, medial or education), which has a significant influence on housing values (Wu et al., 2016).

Among all social networks, Twitter data is ideal for these studies because it contains spatial and temporal information and can be easily used by a moving user in real-time. Neuhaus (2013) used mapped tweets locations to create a 3D landscape of message densities over fifteen cities around the world. Martínez and Gonzílez (2013) studied a dataset of over one million tweets from Mexico City and 1805 locations around the city and found significant differences in the way that Mexico City's inhabitants feel about weekdays. In Lenormand et al. (2014), Twitter, mobile phone and census data shows good agreement in terms of the analysis of mobility in the cities of Madrid and Barcelona. Blanford et al. (2015) created movements tracks by connecting the tweet date and time for each unique user in Kenya to identify how places are connected, thereby demonstrating the utility of social media data for mapping connectivity. To investigate the spatiotemporal characteristics of human mobility with a particular focus on the impact of demography in Chicago, Luo et al. (2016) exploited the explicit location footprints together with demographic information. The results show that demographic information, particularly race/ethnicity group, significantly affects urban human mobility patters. Deng et al. (2018) show the association between geotagged tweets and hourly electricity consumption at the building level on the campus of the State University of New York at Binghamton, and the results suggest a high correlation between electricity consumption and Twitter activity. García-Palomares, Henar Salas-Olmedo, Moya-Gómez, Condeço-Melhorado, and Gutiérrez (2018) show how in Madrid, land uses (e.g., offices, education, health, residential, transport or park and sport) was linked to activities that were activated during major time slots of the day (morning, afternoon, evening and night).

Many of the previous works use a geographic information system (GIS) as a tool for the storage and analysis of data, e.g., Lenormand et al. (2014), Blanford et al. (2015), Wu et al. (2016), García-Palomares et al. (2018), Shaw and Sui (2018) and Wu, Ye, Ren, and Du (2018).

However an important aspect to consider related with social data is that such unstructured/semi-structured but semantically rich data have been argued to constitute 95% of all big data (Gandomi & Haider, 2015). Therefore, this type of constantly accumulating data has been named big social data or social big data (Olshannikova, Olsson, Huhtamäki, & Kärkkäinen, 2017). Thus, social media can be considered to be an important source of information for big data storage and analysis, and would be good to treat it with big data tools. This approach would transform these data to a distributed and redundant data set, which would facilitate quick querying and reduce data loss, as well as considering data scale.

In this paper, we used Twitter data to automatically delineate the spatiotemporal location of events of interest and to identify the urban adequacy of those events. The study covers a complete year (from December 10, 2016, to December 10, 2017) in the city of Valencia (in eastern Spain), which is considered the third most important city in Spain, having a population of nearly 800,000 inhabitants and an area of 135 km^2.

As mentioned above, Twitter data require, a large storage capability, easy data manipulation and quick data access, preferably using a big data tool. Therefore, we opted to use the MongoDB NoSQL database as data storage software, which provides flexibility, scalability and adaptability. MongoDB is an open-source cross-platform document-oriented database that stores its data in documents using a JSON structure. This method allows the data to be flexible and not require a schema. MongoDB can be easily accessed and consulted using its own shell or through the PyMongo driver inside the python software code, which makes MongoDB a notably versatile and easy-to-use database. This property was the main reason for the choice of MongoDB as a database. For example, in Zhou and Xu (2017), MongoDB is used as a Twitter database, and R software is used for the analysis.

Finally, handling big data requires high performance computing or distributed data processing. The state-of-the-art industrial standard is the MapReduce model (Dean & Ghemawat, 2008). The frameworks of Apache Hadoop (Murthy et al., 2011) and Apache Spark (Zaharia et al., 2012) are the most prominent ones for the MapReduce open source implementation. Apache Hadoop is an open source software framework for deploying data-intensive distributed applications. This framework not only implements the MapReduce computational paradigm but also the Hadoop Distributed File System (HDFS), which is derived from the Google File System (GFS) (Ghemawat, Gobioff, & Leung, 2003) as a highly fault-tolerant system to manage big data files. For example, the use of Apache Hadoop for the analysis of geospatial events can be found in Li et al. (2016).

Similar to Hadoop, Spark supports MapReduce style computations while developing the notion of a Resilient Distributed Dataset (RDD), a read-only dataset that is partitioned across multiple machines and can be rebuilt if some partitions are lost. RDDs can be cached in memory and can be reused in multiple MapReduce-like operations. This approach can result in significant performance improvements for certain classes of applications compared to the Hadoop style MapReduce, which writes all intermediate results to disk. Apache-Spark also allows the use of Python programming, which makes it the ideal framework for this research, focusing all of the programming (downloading tweets, storage in MongoDB and analysis) in that language. For example, the use of Apache-Spark for the analysis of geospatial events can be found in Huang, Chen, Wan, and Peng (2017) or Asencio-Cortés, Morales-Esteban, Shang, and Martínez-Álvarez (2017).

The primary novel contributions of this paper are the following:

1. The proposed big data architecture and work flow based on free software and Python code for harvesting, storing, processing, analysing and visualizing social media data.
2. The previous work flow is generated to achieve the primary objective of the research: delineate the spatiotemporal location of events of interest in order to distinguish the urban adequacy for those events. This work flow is considerably more suitable for the treatment of the data volume associated with Twitter than a GIS, which is traditionally the system used in human dynamics analysis. Thus, this work contributes to the generation of background in the use of big data technologies for human dynamics analysis.
3. From a theoretical point of view, the generation of the so-called *specific event maps* and *specific event maps of positive or negative sentiments* are a new concept than can automatically provide the location of events. This new concept should work in collaboration with a word count routine to determine what a particular event is about.

The results obtained from the analysis can also be used to generate background knowledge of the dynamics of Valencia during a complete year and can be used in the future for decision-making or to anticipate issues related to forthcoming events in the city. The presented work can be applied to any city in the world because it is based on location by the coordinates of the tweets and their texts.

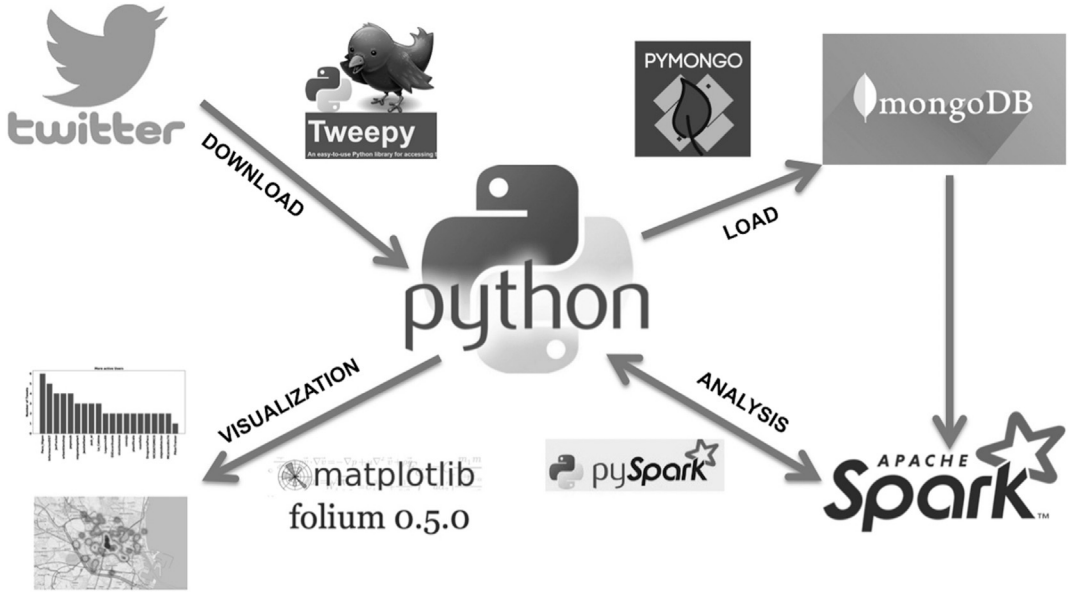

Fig. 1. Architecture implemented in the research.

2. Methodology

Twitter data has been collected and stored continuously since 2016, but for this manuscript, as a case study, only one year of data has been analysed.

The main core of the developed architecture is Python software (Fig. 1). In a first script, the Tweepy library was used to collect and parse Twitter messages as MongoDB documents. Only geolocated tweets in Valencia city were selected, which meant that around only 10% of the total tweets could be downloaded (Surprenant, 2012; Wang et al., 2018).

The downloaded tweets were saved directly to the MongoDB NoSQL database using the PyMongo library; therefore, every tweet is a document in the database containing the following fields: user identification, user screen name, date of tweet creation, latitude, longitude, tweet text, retweet count and favourite count.

A second script queried the database through the PyMongo library and generated a file with the result of the query. This file was then imported as an RDD (Resident Distributed Dataset) file into the Apache Spark framework for analysis (Zaharia et al., 2012). The code for the analysis was written in Python using the MapReduce concept (Dean & Ghemawat, 2008). The Pyspark library and the spark-submit command were important tools for this part of the architecture.

In the analysis, some special aspects were considered:

(1) Some users are related to fixed positions and do not correspond to real persons, for example, the user TrendsValencia or the user meteolivereta (a meteorological station that produces a tweet every hour with climatological information). These tweets were not taken into account in the analysis.

(2) The same user can post several tweets from the same location (e.g., home and work) in a short period of time. The number of such tweets can be extremely high in comparison with other users, leading to an overestimation of the presence of this type of user at these locations and times. It is therefore necessary to eliminate compulsive users who remain in the same place and write tweets in a short period of time.

(3) A sentiment analysis is performed with the text of the tweets. All tweets were written in Spanish or the Valencian language, and, given that they can include special characters, such as HTML tags, punctuation marks, mentions (Twitter user-names preceded by the "@"), and hashtags (thematic words preceded by the "#"), a first process of cleaning the words was performed. A second process was used to remove the so-called "non-words" by using a non-words list (with a total of 312 expressions). These words were removed because of their high frequency in natural language use and their minimal contribution in the linguistic representation of a text paragraph. The third process was to compare the words of the tweet with a list of positive words (4171 words as a corpus) and negative words (4259 corpus words). Finally, the difference between the positive and negative tweet words was used to determine if the tweet sentiment was positive or negative in general. An example of the sentiment analysis of Twitter messages can be found in Martínez and Gonzílez (2013), Corea (2016) or Wallgrün, Karimzadeh, MacEachren, and Pezanowski (2018).

(4) One of the major parts of data visualization is to map the results. Because some users wrote in the same locations (home or work, for example), a random number between 10 and −10 m was added to the latitude and longitude coordinates such that the tweet points with the same coordinates could be visually distinguished on a map and could be selected.

In the code of the second script, tweets could be selected from MongoDB by date, hour interval, user, sentiment or a combination of these attributes. After the selection, the script interacted with the Apache Spark framework and automatically gave the following results:

(1) A text file and figures (using the Matplotlib package) for the number of tweets per day, number of tweets per week day, number of tweets per month, number of tweets per user (indicating who is the most active user) and number of retweets per user (indicating who is the most influential user).

(2) A list with the total word count, positive and negative words count from the tweet text.

(3) Location and plot of the selected tweets using the Folium Python package and the OpenStreetMap cartographic base. Two different maps were generated automatically: a heat map and a map of the location of the tweets using circles for those with neutral sentiments, triangles for those with negative sentiments, and diamonds for positive sentiments. The software included a popup, ensuring that if any tweet was clicked on in the map, information about the date, user screen name and the tweet text was displayed.

3. Results

A complete year of geolocated tweets in the city of Valencia (from December 10, 2016, to December 10, 2017), were downloaded, saved in the MongoDB database and used for the analysis.

The first general analysis corresponds to the complete year, and the

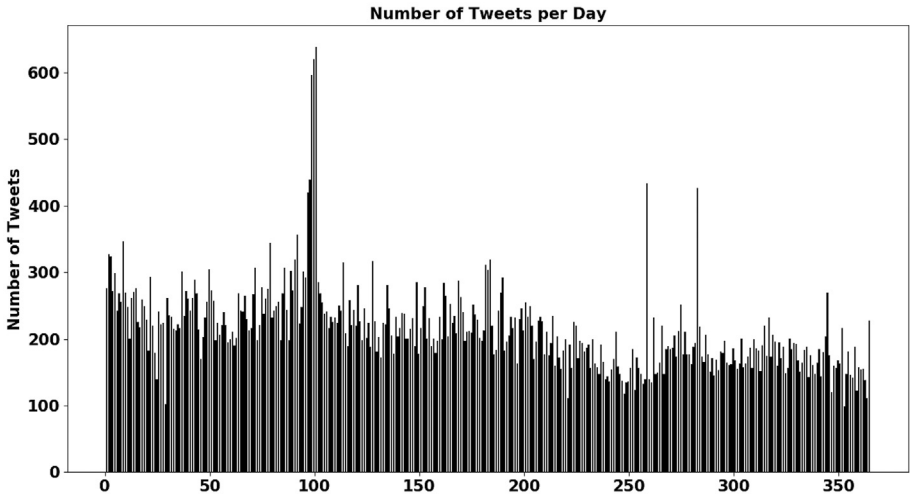

Fig. 2. Number of tweets per day.

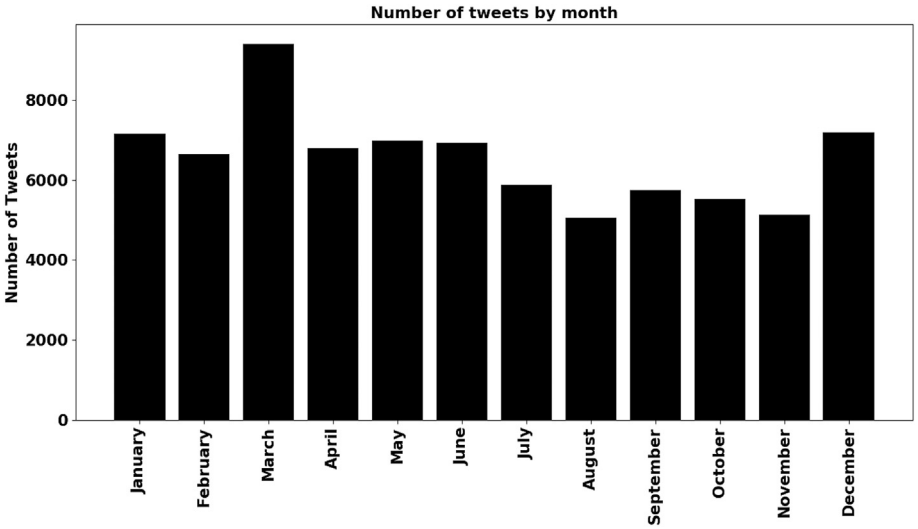

Fig. 3. Number of tweets per month.

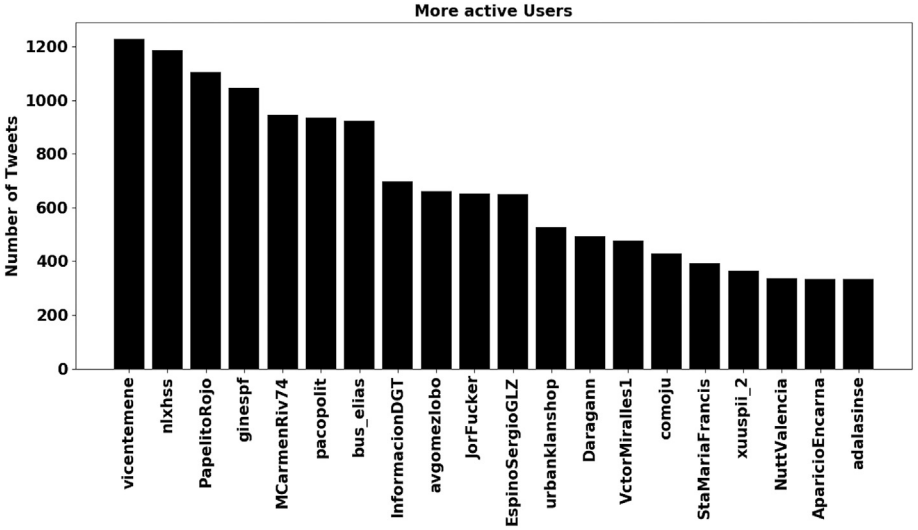

Fig. 4. Most active users.

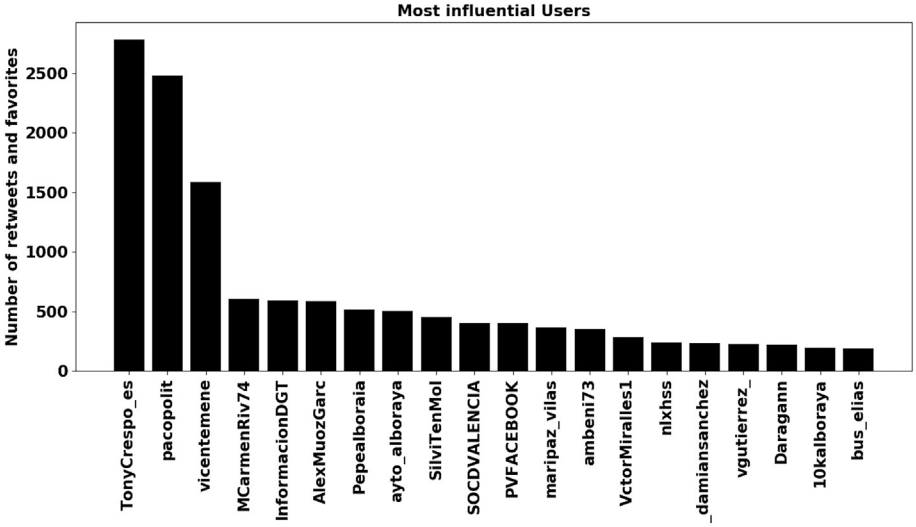

Fig. 5. Most influential users.

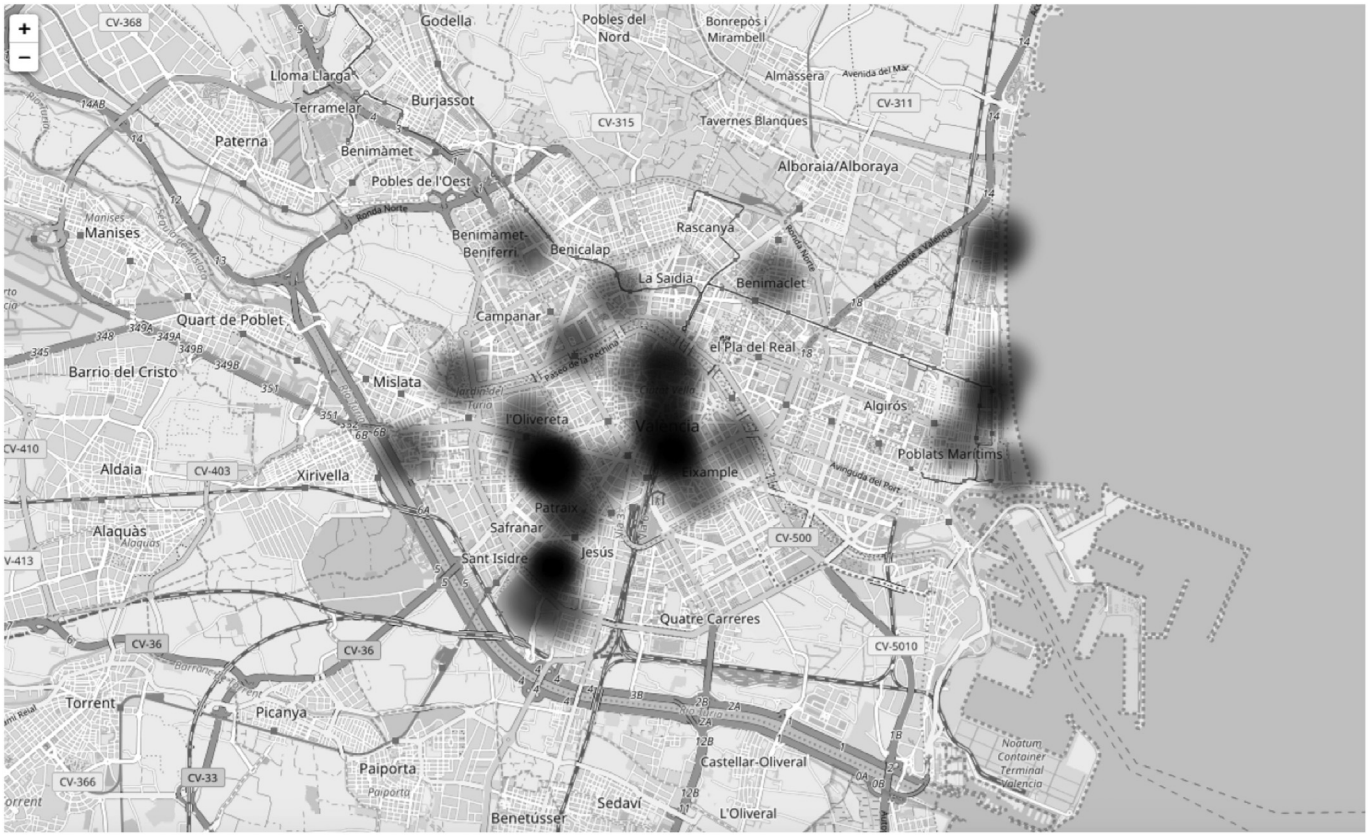

Fig. 6. Heat map for the most active user.

results were the following:

(1) Number of tweets per day (Fig. 2) with a mean value of 215 geolocated tweets per day, where three clear positive peaks can be seen. One of them corresponds to the traditional celebration of *the Fallas* (five consecutive days), and the other two correspond to two particular days and one user, who generate more than 425 tweets for each day with no relevancy (an example of the "compulsive users" defined above that should be eliminated in a deeper analysis).

(2) Number of tweets per month (Fig. 3) where March is the month with the largest number of tweets (corresponding with *the Fallas* month) together with December and January, corresponding with the Christmas period. The month with the least activity is August, coinciding with the holidays of a great part of the population of the city.

(3) Most active users (Fig. 4), referring to users who have written the most tweets can be seen.

(4) More influential users (Fig. 5), referring to users who have received the most retweets and favourites to their tweets. In this figure, three users seem to be the most influential such that for more information, a deep analysis focused on these users can be carried out.

After this first global analysis, a more local analysis can be performed: if we select, for example, only the tweets of the most active user we can include in the analysis the word count of the text of the tweets.

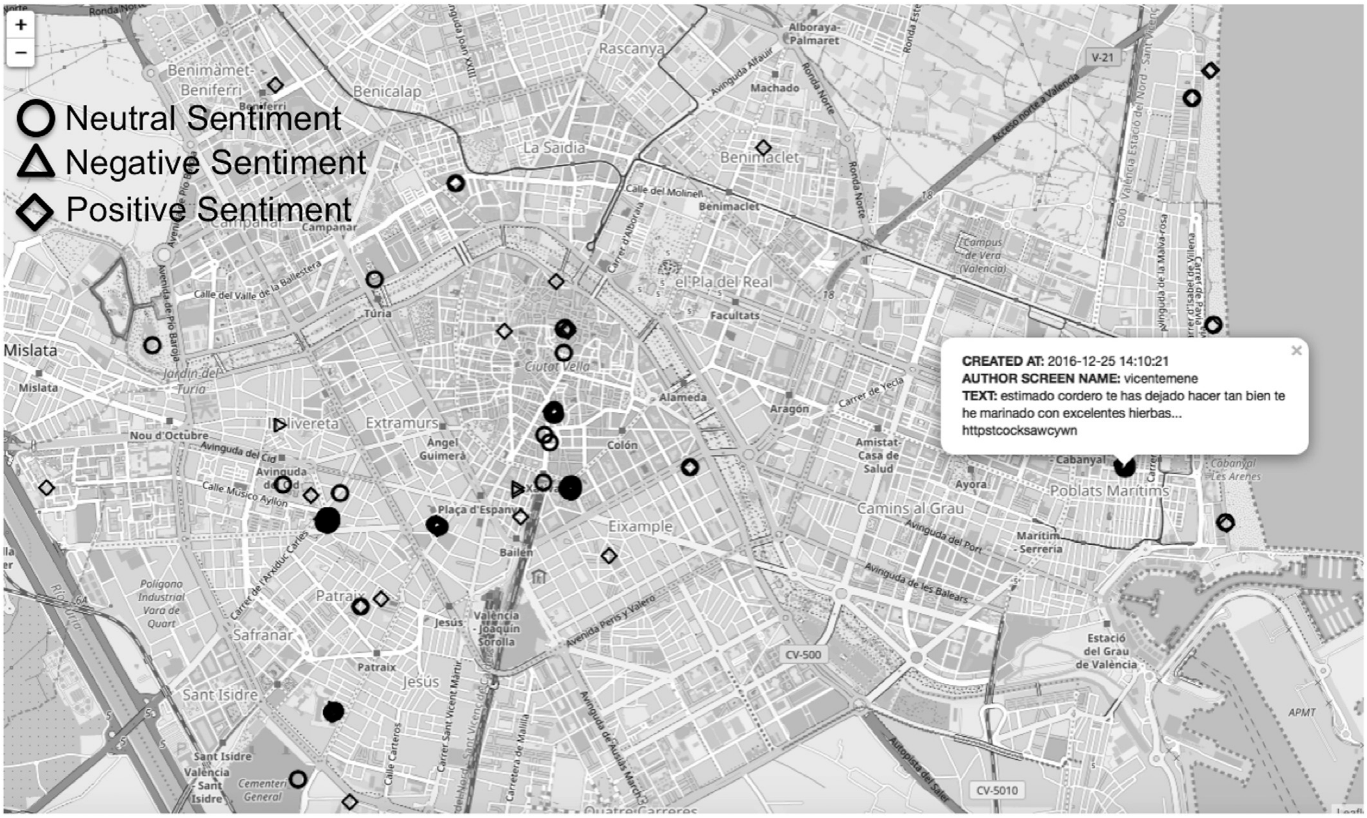

Fig. 7. Location of tweets for the most active user. Example of information associated with a tweet.

Table 1
Top five words during Christmas.

Word	Total	Positive word	Total	Negative word	Total
Christmas	372	Christmas	372	Past	41
Year	322	Happy	259	Bad	40
Happy	259	Best	206	Cold	33
Best	206	Thank you	186	Lack	30
Thank you	186	New	126	Pain	29

This analysis can identify common patterns in all of the analysis carried out, which indicates that there are more positive than negative sentiments among the tweets. For example, in this particular analysis, the two most positive words (translated to English) are love (written 55 times) and good (written 51 times) while the two most negative words are never (written 24 times) and pain (written 16 times).

However, most importantly, a local analysis can also include a clear map visualization of the results. Fig. 6 is the heat map of the most active user during the year, and Fig. 7 represents the locations of the tweets (triangles for negative sentiment tweets, diamonds for positive and circles for neutral). An example of the information that can be obtained from the popup if a specific tweet is clicked upon can also be seen in Fig. 7.

3.1. Deeper analysis during a specific period of days

A deeper analysis can be performed using a specific period of days or a concrete day (e.g., Christmas, the *Fallas*, Halloween, and the Black Friday.). For example, Table 1 provides the most used words, and top five positive and negative words (translated from Spanish) for the Christmas period (from December 10, 2016, to January 10, 2017). Figs. 8 and 9 are the heat maps of the activity during the day (from 8 AM to 10 PM) and night (from 10 PM to 8 AM), respectively. During the day, the activity occurs in the entire city, and during the night, the activity is focused in the city centre where the most popular restaurants and pubs are located.

Heat maps can be used to take a first look at the activity of a city. Fig. 10 is the heat map of Black Friday, and Fig. 11 is the heat map of a particularly important sporting event in the city, a football match between Valencia and Barcelona football clubs. Hot spots during Black Friday are located in the centre of the city (where the most important shops and commercial centres are located), while one of the most important hot spots is located in the football stadium on the match day. A clear idea about the activity can also be obtained using the top five words, Table 2.

The same maps and table can be obtained for the most active and/or influential user for a particular day or period of days, providing multiple possibilities for a deeper analysis.

3.2. Methodology to highlight a specific event

The main drawback of the proposed methodology is that heat and location tweet maps are too general to identify specific activities that occurred in the same places. For example, the heat map or location tweets maps for the Black Friday and Valencia Marathon days (Figs. 10 and 12 respectively) are similar; therefore, the only way to clearly understand what is occurring is to click on the location points of the tweets map and read the text of the tweets. However, a list with the top five words seems to show the activity well. For example, in Table 3, we can see the top five words for Halloween (October 31, 2017), the Valencia Marathon (November 10, 2017), International Epilepsy Day (February 13, 2017), Valentine's Day (February 14, 2017) or the reaction to the terrorist attack in Barcelona (August 25, 2017). With the word count routine, we have an idea about what the event is, but we cannot distinguish an exact place for it.

To obtain a heat map that only plots the exact place for a specific important event that occurred on a specific day, the following methodology was implemented:

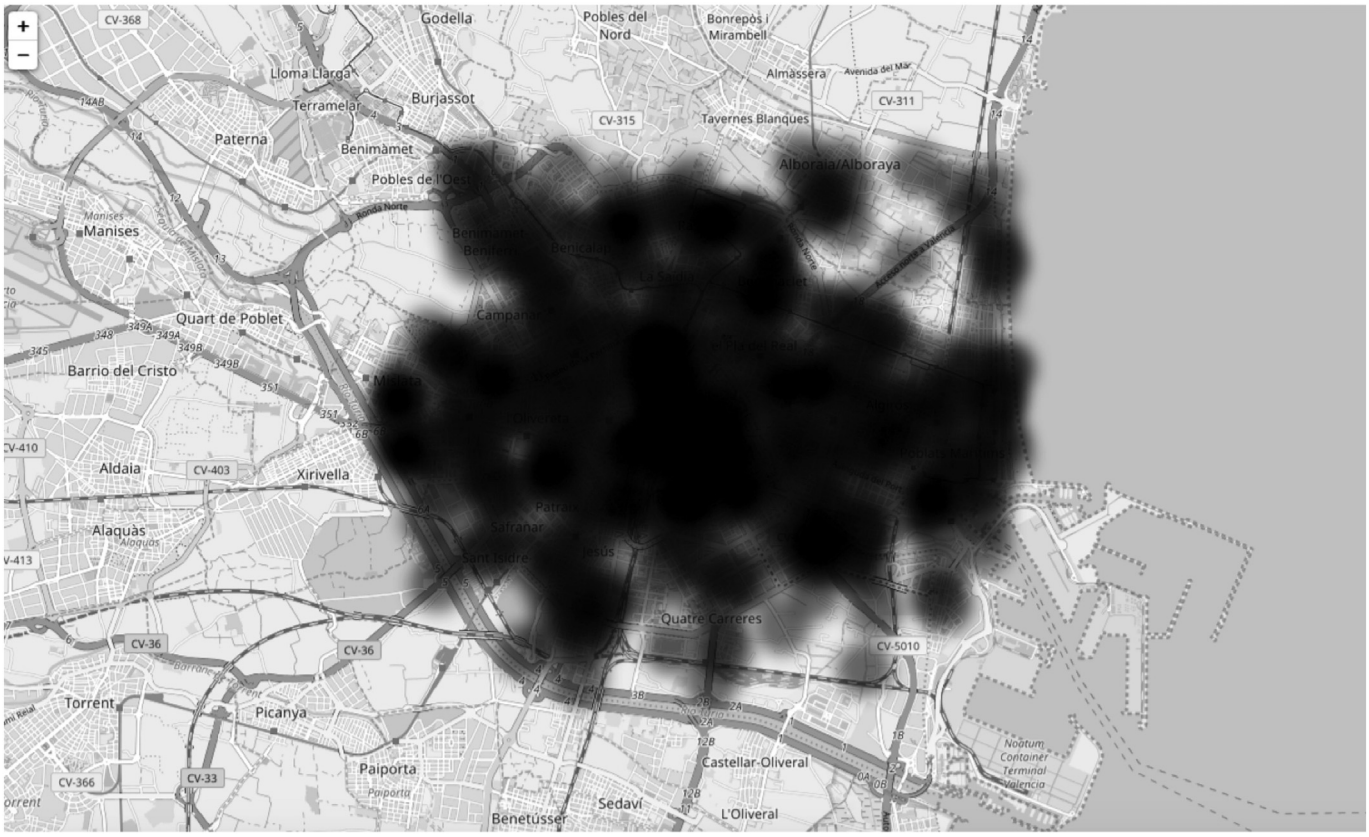

Fig. 8. Day heat map during Christmas.

Fig. 9. Night heat map during Christmas.

Fig. 10. Heat map for Black Friday.

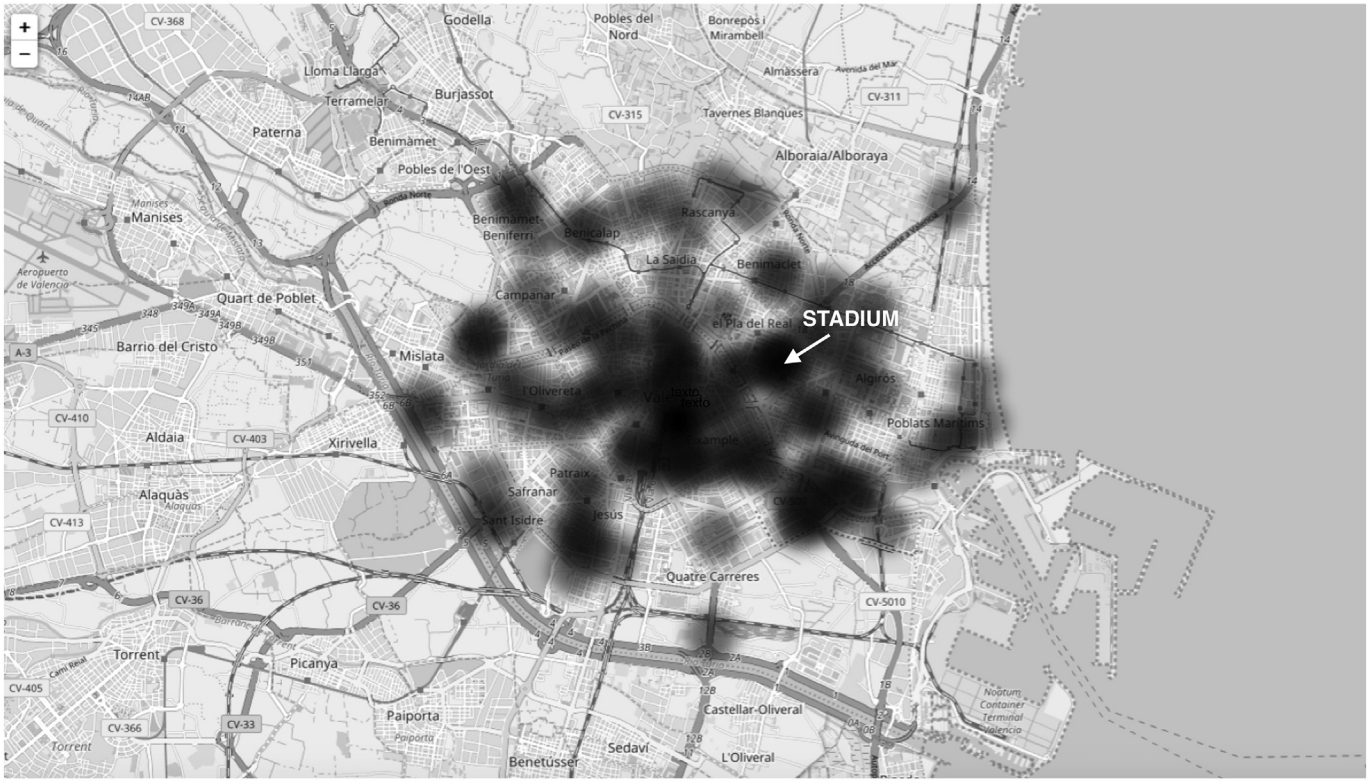

Fig. 11. Heat map for the Valencia-Barcelona football match.

1) A grid with 0.001 geographical degrees of grid space in latitude and longitude (100 m approximately) covering the city was constructed.

2) The summation of the inverse of the distance between a specific grid point and the coordinates of every tweet during the complete year is assigned as a value for that grid point. This value is computed for all the grid points; therefore, if we divide these values by 365, we obtain an idea of a mean heat map for a day.

Table 2
Top five words for Black Friday and Valencia vs Barcelona football match.

Top five words in Black Friday	Total	Top five words in football match	Total
BlackFriday or Black Friday	17	Mestalla (the name of the Stadium)	27
More	14	Stadium	21
City	6	Amunt (a Valencian word of encouragement)	11
Good	6	Spain	9
Home	6	Best	8

Fig. 12. Heat map for the Valencia Marathon day.

Table 3
Top five words for specific days.

10/31/17	Total	11/10/17	Total	2/13/17	Total	2/14/17	Total	8/25/17	Total
Halloween	15	MarathonValencia	31	Epilepsy	14	Valentine	39	Peace	67
November	11	Marathon	27	Thank You	10	Happy	31	Good	32
Night	7	Thank You	12	Letter	9	Love	16	Bad	25
Retention	6	Morning	9	Cake	9	Life	9	Love	18
Level	6	Best	9	Happy	9	Thank You	8	War	14

3) The subtraction of the values of the grid computed for a specific day and the values for the ideal mean day, computed in the previous step, can be used to highlight specific events in a specific place at a certain day (we call these maps, *specific events maps*).

To check this methodology, specific events have been tested: Fig. 13 is the specific heat map for the Valencia-Barcelona football match, where only the football stadium is highlighted. Fig. 14 corresponds to Ricky Martin's concert (2017-5-27), where the place of the concert is highlighted. Finally, if a *Fallas* day is selected, as in Fig. 15, it is expected that all of the city will present specific events. Conversely, those events or special days without a concrete specific location do not generate a specific event map, for example Black Friday, Halloween, The International Epilepsy Day or Valentine's Day.

Therefore, finally, if we check the list of the top words and the specific events maps a clear idea about an event and its location can be obtained.

If all of the specific events maps are combined in a unique map, a final important product can be obtained, where all of the specific events during one year can be viewed in conjunction. This final product can be very useful to check the locations where specific and exact important activities occur and check if the city has appropriate urban infrastructure to support them. This final map for the city of Valencia can be seen in Fig. 16, where the *Fallas* days have not been taken into account for the analysis.

Five important areas have been highlighted in this final map:

–Area 1: *Torres de Serrano* monument. Specific events related to the forthcoming *Fallas* announcement and the declaration (2016-12-1) of *Fallas* as an intangible heritage of mankind.
–Area 2: Specific location in *Alameda* garden. The start of a half

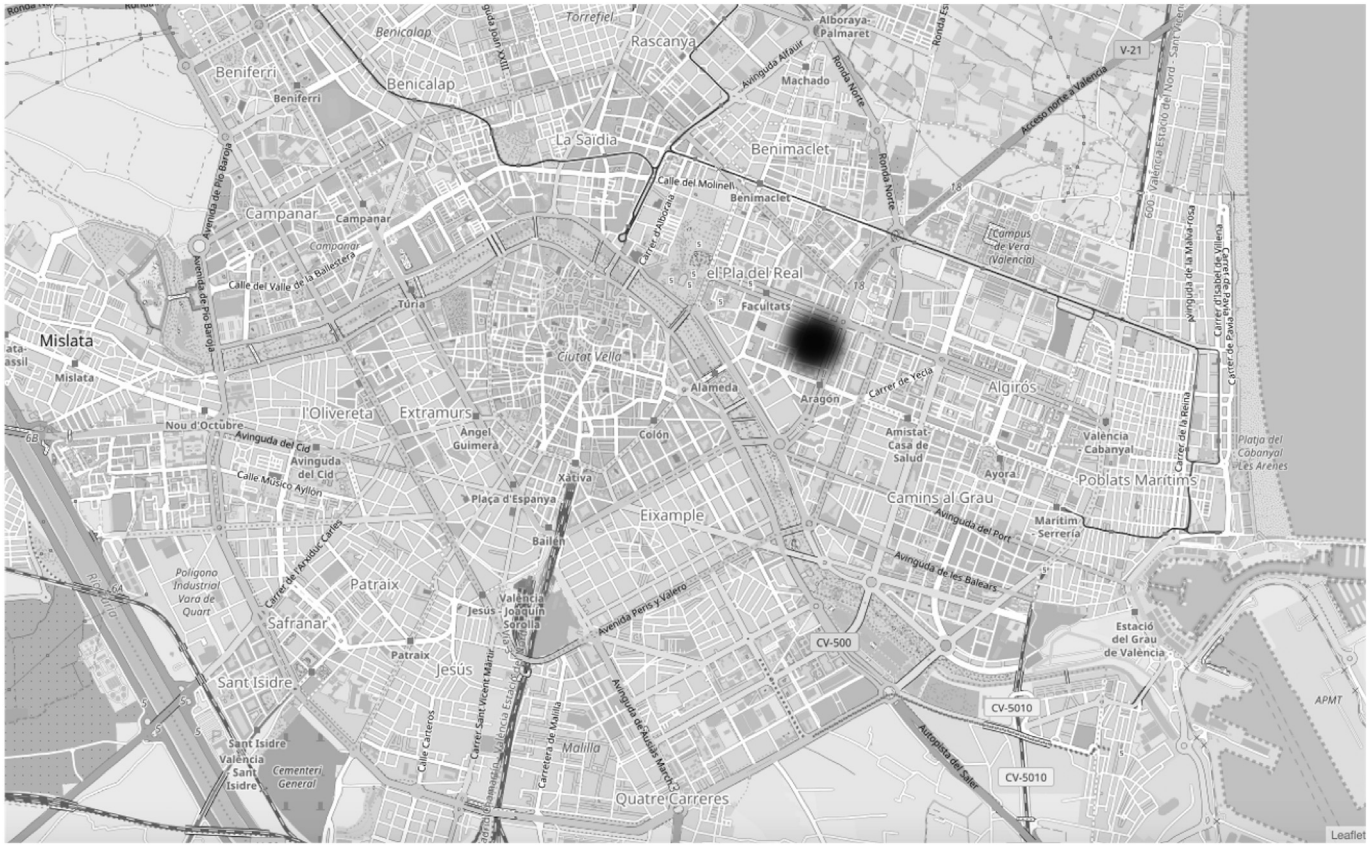

Fig. 13. Specific event map for the Valencia-Barcelona football match.

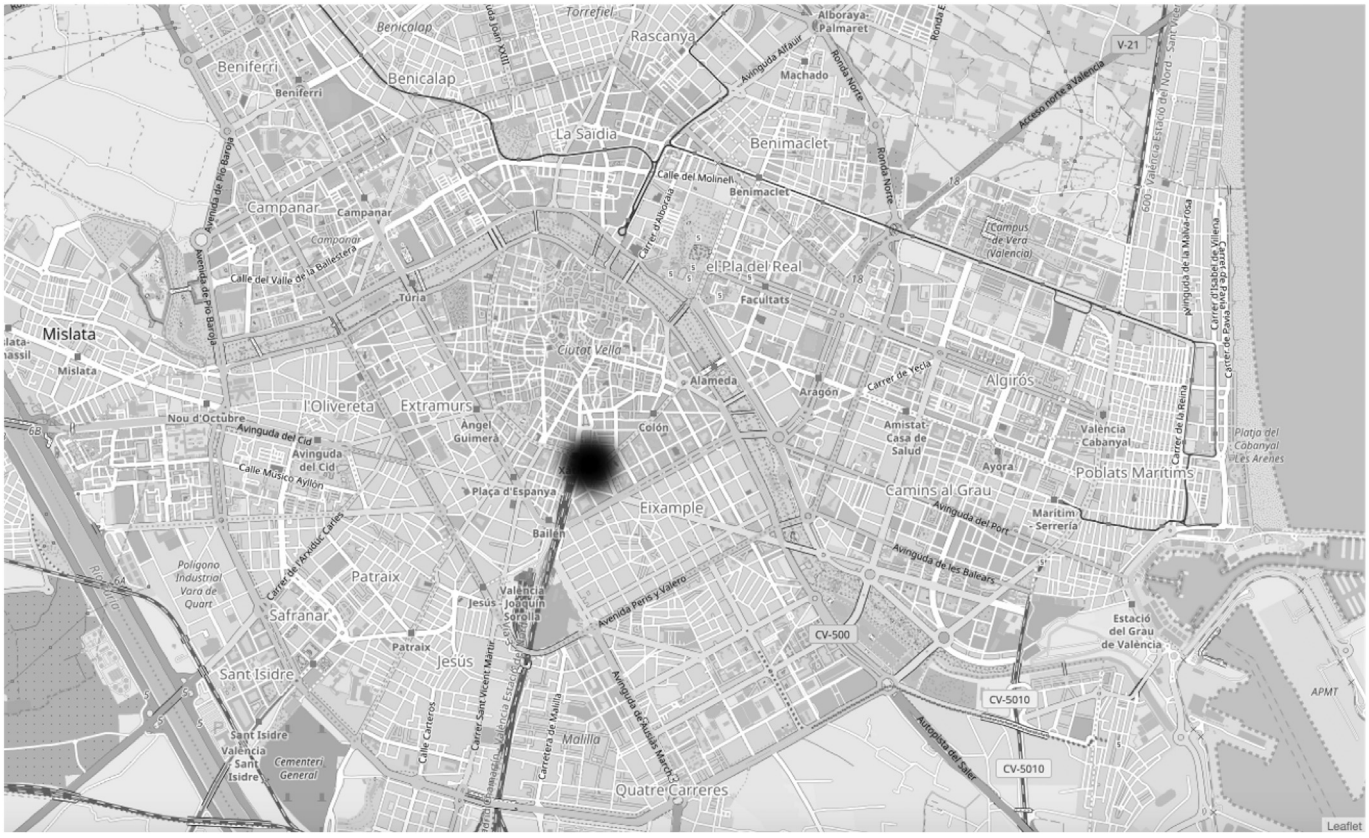

Fig. 14. Specific event map for the concert of Ricky Martin.

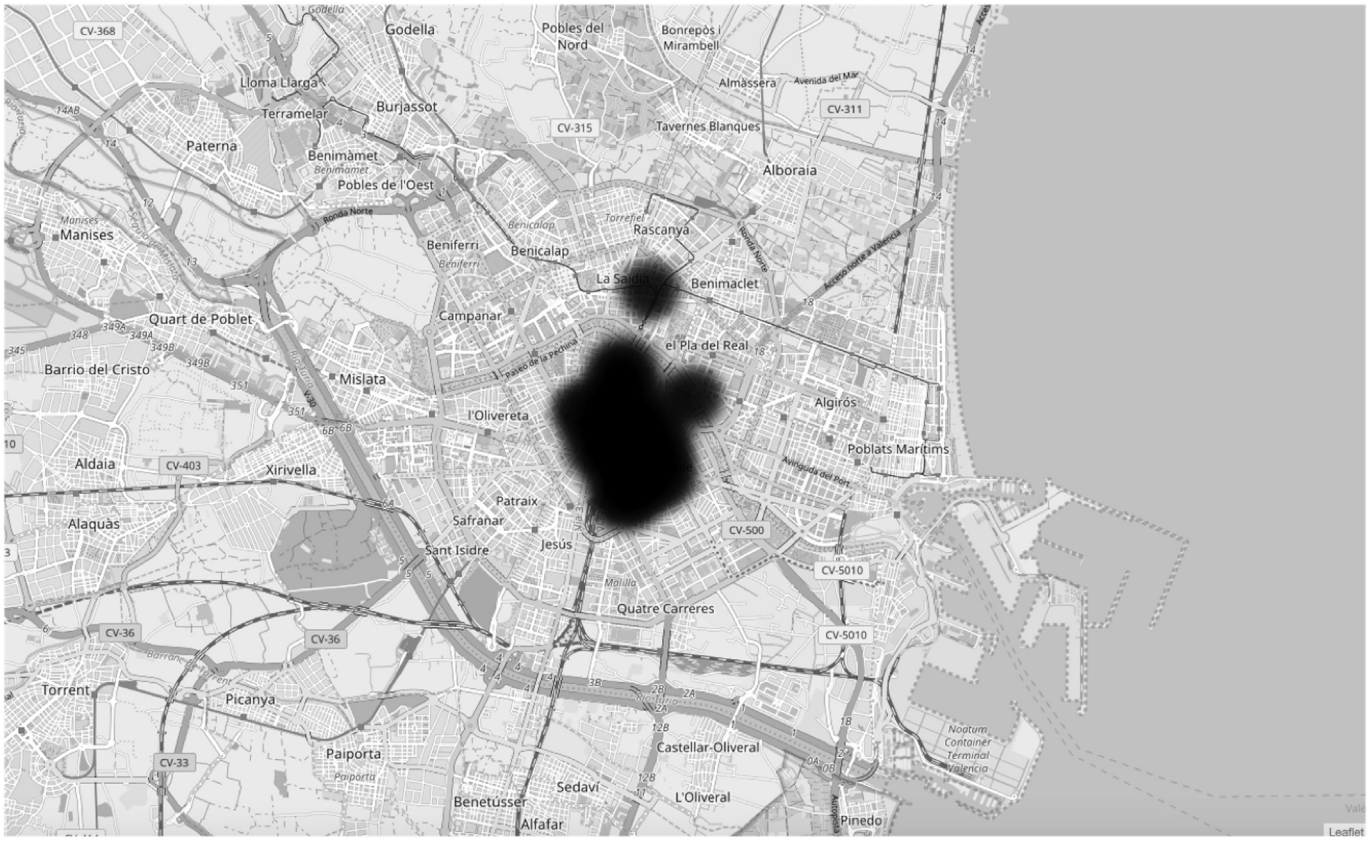

Fig. 15. Specific event map for one *Fallas* day.

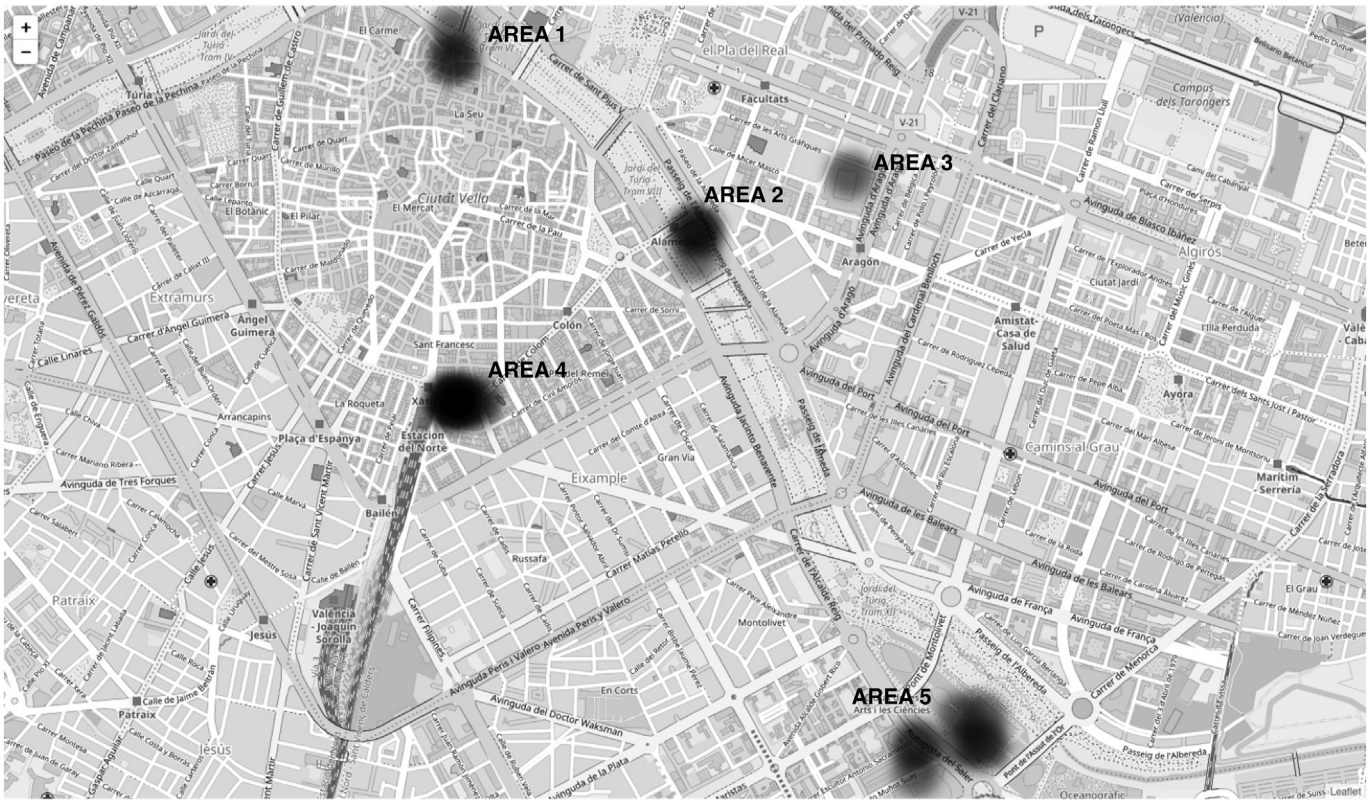

Fig. 16. Accumulated map of all the specific daily event maps of the studied year.

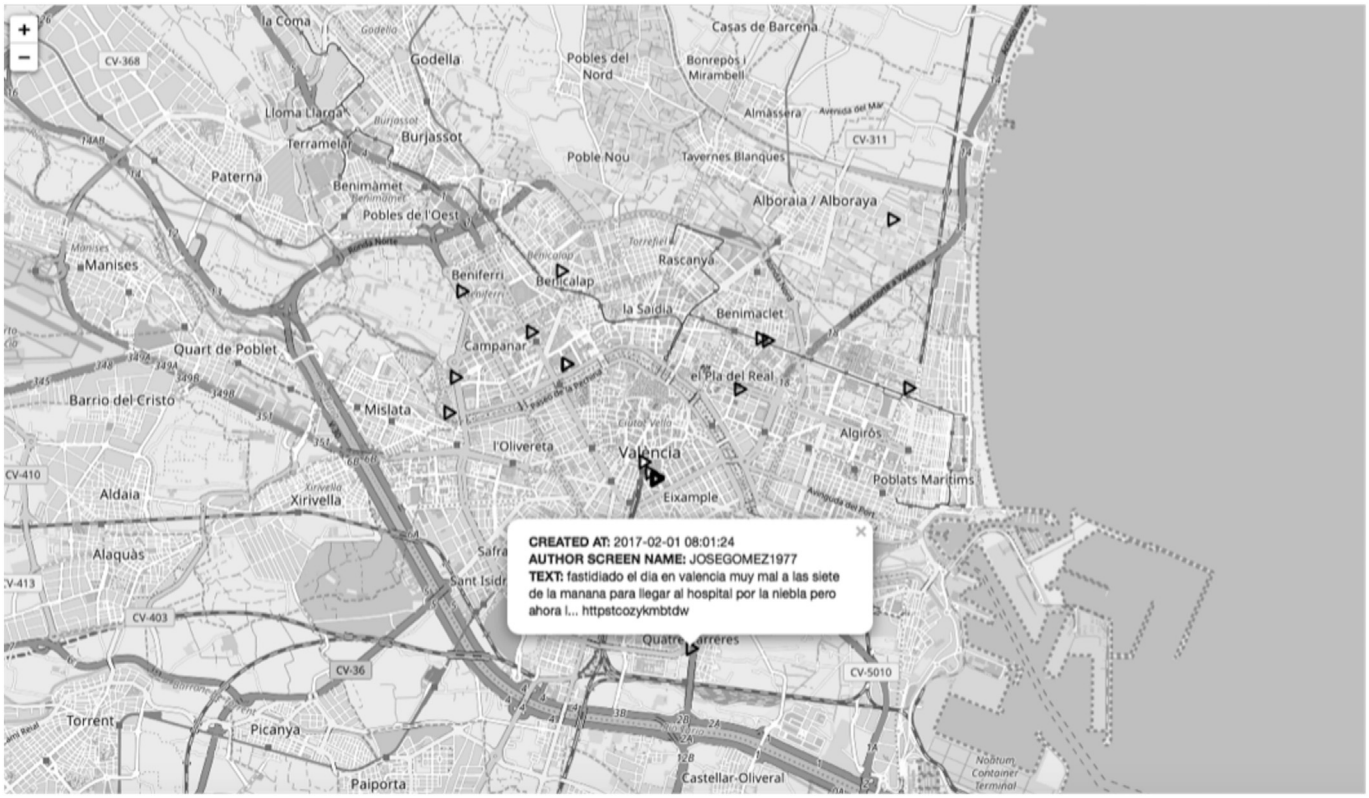

Fig. 17. Map of tweets location for 2017-2-1, where the word *fog* can be found in 7 of those tweets.

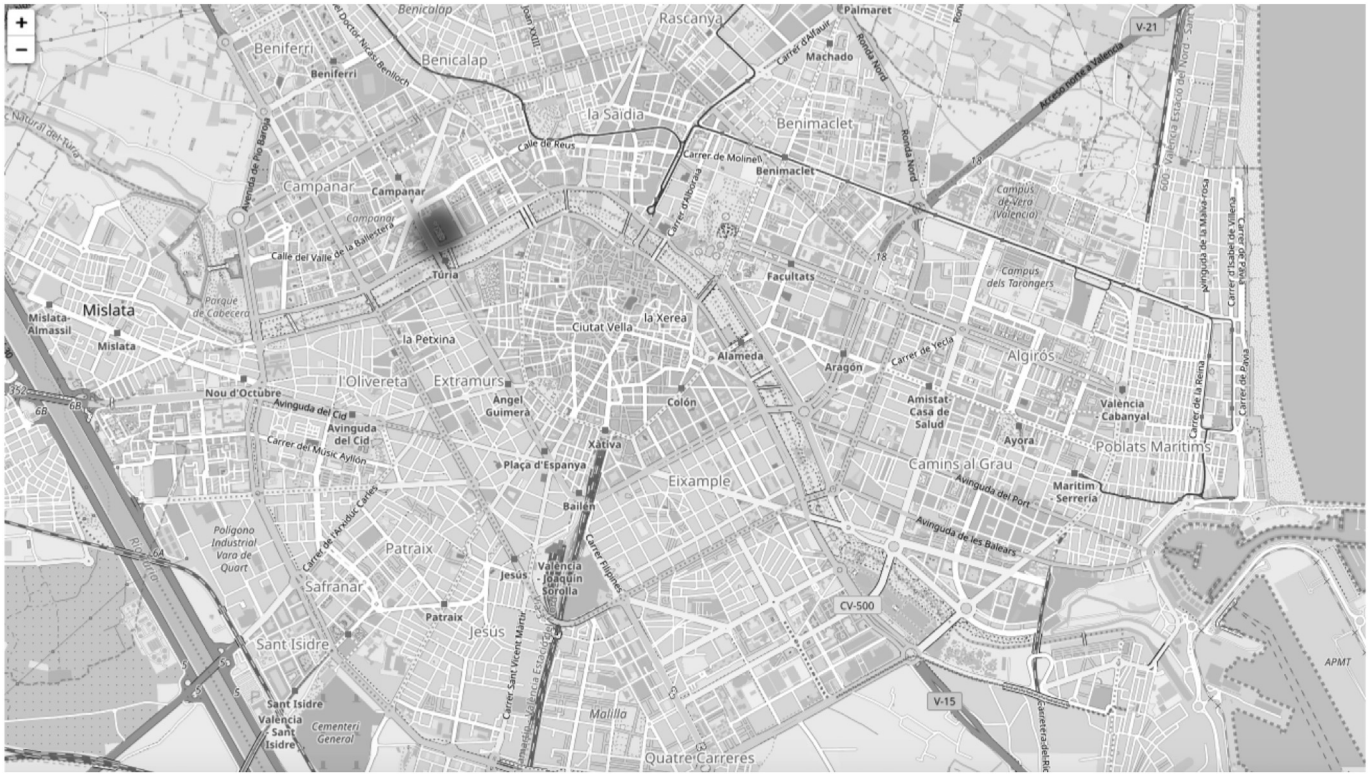

Fig. 18. Specific event map with positive sentiment for the signing of autographs by the contestants of a musical television show.

marathon, fireworks, and some fairs are located there.

–Area 3: Football stadium of the Valencia football club. This only generates specific event maps for important matches (Valencia vs. Barcelona or Valencia vs. Real Madrid).

–Area 4: Bullring of Valencia and a building with some radio stations. In this case these two buildings are close and generate specific event maps for concerts, in the case of the bullring, and special events in some of the radio stations, for example, an

important interview.

–Area 5: *Ciutat de les arts i les ciències de Valencia*, http://www.cac.es/en/home.html. The most touristic place in the city, where some concerts, festivals, and the start and end of the marathon occur.

To complete the proposed methodology, the previous procedure was repeated but considered only the tweets with negative and positive sentiment. Therefore, in addition to the specific event maps, specific event maps of positive and negative sentiments are generated. If the analysis is carried out together with the word count routine, it is possible to obtain information regarding events that have not previously appeared. For example, in Fig. 17, the map of points for 2017-2-1 is presented, where the word *fog* (translated from Spanish and Valencian) can be seen in 8 of them. This natural event has not generated a punctual map, but it emerged from the word count analysis. In Fig. 18, an event appears that did not emerge in the specific event maps analysis of all the tweets, which was based on the word count analysis (from positive tweets in this case). The event is an autograph signing in a commercial centre by the contestants of a musical television show.

In summary, the combination of the specific events maps, the specific event maps of positive and negative sentiments and the word count routine considering only those days with words repeated more than five times, generates a very specific and localized knowledge of the events that occurred in the city.

A detailed study of the events' locations will provide a clear idea regarding the infrastructure used by those events. All of the places founded in this study actually have a good urban capability to receive the specific events highlighted in our analysis, and in some cases, such as the marathon or fireworks, the traffic of the nearby streets is cut, and extra support by police and emergencies is activated; therefore, no other extra urban or human support in addition is needed. This means that the infrastructure of the city has been sufficient for the events that occurred in those sites, but it may not be for future events in the same places or different places. What is remarkable is that these places will most likely be the places for future events, so the city authorities should pay special attention to providing them with sufficient infrastructure.

4. Discussion

From this analysis, it has been seen that the most important tools to measure the pulse of the activity of the city seems to be the most used words and the specific map events. This characteristic is observed because daily heat and location tweet maps are too general to identify and distinguish between specific activities that occurred in most cases in the same places. Therefore, the only way to clearly understand what is occurring is to click on the location points of the tweets map and read the text of the tweets. However, a list with the top five words seems to show the activity notably well. A deep analysis of the list of top words for 365 days during the studied period showed that in general, if a day presents a list with more than two or three words repeated more than ten times each, it usually means something for this particular day. The reading of these words will help to know exactly what is occurring in the city for that specific day. Simultaneously, a look at the specific map events of that day and the specific map events with positive and negative sentiments, will help us locate the focal point of the activity if it exists.

5. Conclusions

Twitter activity can be used to take the pulse of the activity in the city of Valencia. These raw data should be collected, saved and analysed in a proper way to obtain valid conclusions. Big data tools have proven to be effective in this context. MongoDB is used in this research as a database, while Apache Spark is used as an analysis tool. The architecture developed is based in Python language, which has led to the development of self-contained software which, in the same process, performs the query in the database, runs the analysis and generates the appropriate output files and maps. Based on the resulting visualization of the analysis, it can be concluded that the top five words list together with the concepts developed of the *specific event maps and specific event maps of positive and negative sentiments* is the best way to have an idea about a particular activity and its location in the city. A deeper study for a specific day can also be performed by a look at the heat and tweet position maps where the information about a specific tweet (user name, date and text), which can be obtained by clicking on a tweet on the location map could be an important tool to distinguish among users or tweet creation time.

Based on the accumulation of the *specific events map and specific event maps with positive and negative sentiments* for every day of the analysed year, and the particular location of the highlighted events, we can conclude that city authorities should pay special attention to providing them with sufficient infrastructure, which has been enough so far but may not be enough in the future.

We acknowledge some limitations in this study. First, due to the low proportion of geotagged tweets, even though we can extract meaningful local events, only major local events can be revealed. Second, we used individual words as the analysis unit. Events described by phrases may improve the results, but due to the noisy and complex nature of social media messages, extracting meaningful and actionable knowledge from user generated content is nontrivial. However, the search for beneficial applications and services in regard to big social data have only just begun, especially in urban areas considering that by 2050, 66% of the world's population is projected to be urban (United Nations, 2014).

Acknowledgements

The authors would like to thank the comments and suggestions of the anonymous reviewers and the editor, which have helped to improve the original version.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Addair, T. G., Dodge, D. A., Walter, W. R., & Ruppert, S. D. (2014). Large-scale seismic signal analysis with Hadoop. *Computers and Geosciences, 66*, 145–154.

Asencio-Cortés, G., Morales-Esteban, A., Shang, X., & Martínez-Álvarez, F. (2017). Earthquake prediction in California using regression algorithms and cloud-based Big Data infrastructure. *Computers and Geosciences*. https://doi.org/10.1016/j.cageo.2017.10.011.

Bibri, S. E., & Krogstie, J. (2017). ICT of the new wave of computing for sustainable urban forms: Their big data and context-aware augmented typologies and design concepts. *Sustainable Cities and Society, 32*, 449–474.

Blanford, J. I., Huang, Z., Savelyev, A., & MacEachren, A. M. (2015). Geo-located tweets. Enhancing mobility maps and capturing cross-border movement. *PLoS One, 10*(6), e0129202. https://doi.org/10.1371/journal.pone.0129202.

Clark, J., Kearns, A., & Cleland, C. (2016). Spatial scale, time and process in mega-events: The complexity of host community perspectives on neighbourhood change. *Cities, 53*, 87–97.

Corea, F. (2016). Can Twitter proxy the investors' sentiment? The case for the technology sector. *Big Data Res. 4*(C), 70–74.

Correa, J. C., & Camargo, J. E. (2017). Ideological consumerism in Colombian elections, 2015: Links between political ideology, Twitter activity, and electoral results. *Cyberpsychology, Behavior and Social Networking, 20*(1), 37–43.

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). #Earthquake: Twitter as a distributed sensor system. *Transactions in GIS, 17*, 124–147. https://doi.org/10.1111/j.1467-9671.2012.01359.x.

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM, 51*(1), 107–113.

Deng, C., Lin, W., Ye, X., Li, Z., Zhang, Z., & Xu, G. (2018). Social media data as a proxy for hourly fine-scale electric power consumption estimation. *Environment and Planning A: Economics and Space, 0*(0), 1–5. https://doi.org/10.1177/0308518X18786250.

Enenkel, M., See, L., Bonifacio, R., Boken, V., Chaney, N., Vinck, P., ... Anderson, M. (2015). Drought and food security—Improving decision-support via new

technologies and innovative collaboration. *Global Food Security, 4*, 51–55. https://doi.org/10.1016/j.gfs.2014.08.005.
Figuereido, J. F. (2017). *Social Media text processing and semantic analysis for smart cities.* Thesis (PhD)Portugal: Universdade do Porto. https://arxiv.org/pdf/1709.03406.pdf.
Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144.
Gao, Y., Wang, S., Padmanabhan, A., Yin, J., & Cao, G. (2018). Mapping spatio temporal patterns of events using social media: A case study of influence trends. *International Journal of Geographical Information Science, 32*(3), 425–449.
García-Palomares, H. C., Henar Salas-Olmedo, M., Moya-Gómez, B., Condeço-Melhorado, A., & Gutiérrez, J. (2018). City dynamics through Twitter: Relationships between land use and spatiotemporal demographics. *Cities, 72*, 310–319.
Ghemawat, S., Gobioff, H., & Leung, S. (2003). The google file system. *Proceedings of the 19th symposium on operating systems principles, 19–22 October 2003, Lake George, New York*.
Huang, Z., Chen, Y., Wan, L., & Peng, X. (2017). GeoSpark SQL: An effective framework enabling spatial queries on Spark. *International Journal of Geo-Information, 6*(9), 285. https://doi.org/10.3390/ijgi6090285.
Huang, Q., & Xiao, Y. (2015). Geographic situational awareness: Mining tweets for disaster preparedness, emergency response, impact, and recovery. *ISPRS International Journal of Geo-Information, 4*, 1549–1568. https://doi.org/10.3390/ijgi4031549.
Khan, Z., Anjum, A., Soomro, K., & Tahir, M. A. (2015). Towards cloud based big data analytics for smart future cities. *Journal of Cloud Computating: Advances, Systems and Applications, 4*, 2. https://doi.org/10.1186/s13677-015-0026-8.
Kitchin, R. (2014). The real-time city? Big data and smart urbanism. *GeoJournal, 79*, 1–14.
Kousiouris, G., Akbar, A., Sancho, J., Tashma, P., Psychas, A., Kyriazis, D., & Varvarigou, T. (2018). An integrated information lifecycle management framework for exploiting social network data to identify dynamic large crowd concentration events in smart cities applications. *Future Generation Computer Systems, 78*, 516–530.
Lenormand, M., Picornell, M., Cantú-Ros, O., Tugores, A., Louail, T., Herranz, R., ... Ramasco, J. J. (2014). Cross-checking different sources of mobility information. *PLoS One, 9*, e105184. https://doi.org/10.1371/journal.pone.0105184.
Li, Z., Hu, F., Shnase, J. L., Duffy, D. Q., Lee, T., Bowe, M. K., & Yang, C. (2016). A spatiotemporal indexing approach for efficient processing of big array-based climate data with MapReduce. *International Journal of Geographical Information Science, 31*(1), 17–35.
Lim, C., Kim, K. J., & Maglio, P. P. (2018). Smart cities with big data: Reference models, challenges, and considerations. *Cities*. https://doi.org/10.1016/j.cities.2018.04.011.
Longley, P. A., & Adnan, M. (2016). Geo-temporal Twitter demographics. *International Journal of Geographical Information Science, 30*(2), 369–389.
Luo, F., Cao, G., Mulligan, K., & Li, X. (2016). Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago. *Applied Geography, 70*, 11–25.
Martí, P., Serrano-Estrada, L., & Nolasco-Cirugeda, A. (2017). Using locative social media and urban cartographies to identify and locate successful urban plazas. *Cities, 64*, 66–78.
Martínez, V., & Gonzílez, V. M. (2013). Sentiment characterization of an urban environment via Twitter. In G. Urzaiz, S. F. Ochoa, J. Bravo, L. L. Chen, & J. Oliveira (Vol. Eds.), *Lecture Notes in Computer Science: . 8276*. Cham: Springer.
Mersham, G. (2010). Social media and public information management: The September 2009 tsunami threat to New Zealand. *Media International Australia, 137*, 130–143.
Murthy, A. C., Douglas, C., Konar, M., O'Malley, O., Radia, S., & Agarwal, S.. Architecture of next generation Apache Hadoop MapReduce framework. Apache Jira. (2011). [online]. Available from: https://www.thesisscientist.com/docs/Others/95e16e87-4b94-444b-b619-10e823525ca3 (Accessed 1 September 2017) .
Neuhaus, F. (2013). New city landscape: Mapping twitter data in urban areas. *The Cartographic Journal, 46*, 25–30.
Olshannikova, E., Olsson, T., Huhtamäki, J., & Kärkkäinen, H. (2017). Conceptualizing Big Social Data. *Journal of Big Data, 4*(3).
Restrepo-Estrada, C., Camargo de Andrade, S., Abe, N., Fava, M. C., Mendionda, E. D., & Porto, J. (2018). Geo-social media as a proxy for hydrometeorological data for streamflow estimation and to improve flood monitoring. *Computational Geosciences, 111*, 148–158.
Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes twitter users: Real-time event detection by social sensors. *Proceedings of the 19th international conference on World Wide Web. 26–30 April 2010 New York* (pp. 851–860). . https://doi.org/10.1145/1772690.1772777.
Shaw, S. L., & Sui, D. (2018). Editorial: GIScience for human dynamics research in a changing world. *Transactions in GIS, 22*, 891–899. https://doi.org/10.1111/tgis.12474.
Surprenant, C.. Twitter big data [online]. (2012). Available from: https://es.slideshare.net/colinsurprenant/twitter-big-data (Accessed 12 January 2018) .
United Nations, Department of Economic and Social Affairs, Population Division (2014). *World urbanization prospects: The 2014 revision.* Highlights asdfNew York: United Nations (ST/ESA/SER.A/352).
Wallgrün, J. O., Karimzadeh, M., MacEachren, A. M., & Pezanowski, S. (2018). GeoCorpora: Building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science, 32*(1), 1–29.
Wang, R. Q., Mao, H., Wang, Y., Rae, C., & Shaw, W. (2018). Hyper-resolution monitoring of urban flooding with social media and crowdsourcing data. *Computational Geosciences, 111*, 139–147.
Wu, C., Ye, X., Ren, F., & Du, Q. (2018). Check-in behaviour and spatio-temporal vibrancy: An explanatory analysis in Shenzhen, China. *Cities, 77*, 104–116.
Wu, C., Ye, X., Ren, F., Wan, Y., Ning, P., & Du, Q. (2016). Spatial and social media data analytics of housing prices in Shenzhen, China. *PLoS One, 11*(10), https://doi.org/10.1371/journal.pone.0164553.
Wu, Y., Zhang, W., Shen, J., Mo, Z., & Peng, Y. (2018). Smart city with Chinese characteristics against the background of big data: Idea, action and risk. *Journal of Cleaner Production, 173*, 60–66.
Wu, L., Zhi, Y., Sui, Z., & Liu, Y. (2014). Intra-urban human mobility and activity transition: Evidence from social media check-in data. *PLoS One, 9*(5), e97010. https://doi.org/10.1371/journal.pone.0097010.
Yuan, M. (2018). Human dynamics in space and time: A brief history and a view forward. *Transactions in GIS, 22*, 900–912.
Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., ... Stoica, I. (2012). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. *NSDI'12 proceedings of the 9th USENIX conference on networked systems design and implementation. 25–27 April 2012 San José*https://www.usenix.org/system/files/conference/nsdi12/nsdi12-final138.pdf.
Zhi, Y., Li, H., Wang, D., Deng, M., Wang, S., Gao, J., ... Liu, Y. (2016). Latent spatio-temporal activity structures: A new approach to inferring intra-urban functional regions via social media check-in data. *Geo-Spatial Information Science, 19*(2), 94–105.
Zhou, L., & Wang, T. (2014). Social media: A new vehicle for city marketing in China. *Cities, 37*, 27–32.
Zhou, X., & Xu, C. (2017). Tracing the spatial-temporal evolution of events based on social media data. *International Journal of Geo-Information, 6*(88), https://doi.org/10.3390//ijgi6030088.