Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 52 (2021) 493–500

Transportation Research Procedia

www.elsevier.com/locate/procedia

23rd EURO Working Group on Transportation Meeting, EWGT 2020, 16-18 September 2020, Paphos, Cyprus

Understanding mobility patterns and user activities from geo-tagged social networks data

André Miguel Carvalho[a], Marta Campos Ferreira[1], Teresa Galvão Dias[1]

[a]*R. Dr. Roberto Frias, 4200-465 Porto*

Abstract

Social networks are strongly present in the daily life of modern society. Most people use these social networks to share information about their lives, their opinions, places they visit and their state of mind. Generally, these posts are composed of various information, being the location of the users location part of the data. The purpose of this work is to obtain the location of the posts and observe the users mobility pattern in the city of Porto, Portugal. This paper discusses the technologies available for obtaining the data, the social networks currently worth studying and their respective restrictions. It also explores new approaches to collect the data from the desired social networks, respecting all restrictions currently applied. The different software solutions developed for the social networks interactions are explored and described in depth. Subsequently, the necessary software for social networks is reviewed, the possible algorithms for data mining are discussed and its implementation is presented. Finally, the results obtained are interpreted and studied according to the characteristics of the city, tourism promotions and transport routes.

Keywords: Social Networks; Data Mining; Mobility Patterns; Clustering; Urban Planning.

1. Main Text

As the population grows exponentially around the world, most urban planning methodologies have also had to evolve to integrate the latest technologies and ensure that processes do not become obsolete. *Urban planning* is the subject responsible for designing, building and developing the land, air and water resources to better serve the population. It is also responsible for subjects like infrastructure, communications, distribution and public transportation. McGill The study of people's mobility provides information on public transport infrastructure and routes, police dis-

[*] Corresponding author. Tel.: +351 912149682.
E-mail address: up201802161@fe.up.pt

10.1016/j.trpro.2021.01.058

tribution and road development. Not so long ago, these studies were expensive and long. However, nowadays people are willing to share personal information through smartphone apps.

Location Based Social Networks (LBSNs) have grown over time to the point where users interact with them daily. Indeed, most users use them to post pieces of text or media which contains information regarding their behaviour. In particular, most posts present information related to the users locations. Ideally, these locations can be collected and studied to understand the most visited places within the city. Therefore, exploiting users posts in social networks can provide useful information on how to improve certain urban planning topics like infrastructure and public transportation grids.

This paper aims to understand the mobility patterns and user activities from data collected from social networks. This analysis was carried out for the city of Porto, Portugal. Porto has become one of the best destinations for tourism worldwide Europeanbestdestinations, and there is enough evidence that this growth would continue in the near future Silva et al. (2019) before COVID19. Moreover, there are no studies of this kind carried out so far for this city. In addition to the spatio-temporal analysis carried out, this paper also proposes a solution to collect data from social networks, respecting the limitations and restrictions imposed by data protection laws.

This paper is organized as follows: the next section describes the current state of the art by presenting similar work, their restrictions and how this work might solve previous problems. Section 3 describes the methodology used to perform the analysis. Afterwards the results obtained are presented and discussed in section 4, and section 5 presents the mains conclusions and suggestions for future work.

2. State Of The Art

Research with similar scopes has been previously conducted by other authors. The works of Bejar et al. (2016), Wu et al. (2018), Diamantini et al. (2017), Paldino et al. (2015) aimed to collect data from social networks and understand where people spent most time and how they moved.

Bejar et al. (2016) conducted a similar study for the cities of Barcelona and Milan. They collected the data from *Instagram* and *Twitter*. How they collected this data is not explained however, they were able to collect between 1.5 million to 6 million posts. Their work focused on discovering spatio-temporal patterns in the cities through location based networks and they discussed *Leader Clustering Algorithms*, *K-Means*, *FP-Trees* and *Affinity Propagation Clustering* but ended up using mostly *K-Means*.

The Wu et al. (2018) project aimed at building a network of tourism hotspots & clusterings. The city analysed in this research was Beijing and the main social network studied was *Flickr*. The authors were able to gather around 185,531 posts and mostly used the *Fast Search and Find of Density Peaks clustering algorithm*.

In Diamantini et al. (2017) the users watched the *Instagram* activity of the visitors of EXPO 2015. They were able to collect around 570,973 posts and their project focused on discovering the mobility patterns of users through process mining. To achieve this, the authors used *K-Means* techniques and *Infrequent Inductive Miner and Fuzzy Miner*.

The work of Paldino et al. (2015) focused on understanding the urban magnetism of cities through geo-tagged photos around the world. They were able to get a total of 90 million posts worldwide all coming from *Flickr*. For this analysis the authors used diverse clustering techniques.

Is important to notice that in these works the authors also started by gathering information from location based social networks. However, with the passing of time most of the methodologies mentioned have become obsolete or banned due to the new GDPR laws. Following the data collection, authors also used clustering techniques to understand which were the most visited places in the cities. At last, the authors proposed improvements to different subjects of the city in question.

This work aims to understand the mobility patterns and user activities from geo-tagged social networks data. This analysis was carried out for the city of Porto, Portugal, and there are no studies of this kind carried out so far for this city. Additionally, this paper proposes a solution to collect data from social networks. Nowadays, collecting data from social networks is very difficult and restricted, having been one of the major challenges encountered in conducting this analysis.

3. Methodology

Nowadays, with the spread of social networks, people voluntarily share personal information through smartphone applications. This information can be gathered, filtered and processed to target mobility aspects only and identify new possibilities. This section presents the methodology followed to collect, process and analyse data from social networks.

3.1. Data Collection Methods

Research for the state of the art began by exploring possible ways to obtain the data and from which platforms.. Works like Bejar et al. (2016), Wu et al. (2018), Valverde-Rebaza et al. (2016), Hammar et al. (2019) all used LBSNs for similar purposes as this work. Based on their experiences it was possible to conclude that the two best approaches for collecting the data are through *APIs* (application programming interface) Red Hat or *web scraping* Singrodia et al. (2019). In short, APIs provide methods for developers to communicate with the platforms. However, when a platform does not provide an API it is possible to develop a scraper that goes to the desired web pages and extracts the information needed. The main differences between these methods is that APIs are easier to integrate and to use, but the data obtained is limited to the platform creators limitations, while scrapers collect data without limitations but require an extensive software development. Although it was not used in previous works, other possible method for collecting the data is through web scrapers.

3.2. Data Sources

The first challenge comes from obtaining data from posts on social networks. As is well known, companies today must comply with strict rules to protect the users' information. These restrictions impose some limitations on the possible data obtained. Moreover, companies' market value is highly related to the quantity, quality and exclusivity of the data they have, therefore, so sharing too much data would make companies lose that competitive advantage.

Also, selecting proper data sources is essential in order to acquire quality data for meaningful results. On the papers mentioned above the authors analysed different social networks like: *Foursquare*, *Twitter*, *Instagram* and *Flickr*. Although these platforms were quite popular at the time, nowadays only *Twitter*, *Facebook*, *Instagram* and *Flickr* remain popular and therefore they were the chosen ones for studying. However, *Flickr* was removed because the posts are more focused on professional photography and not sharing places of interest. At last, *Facebook* also had to be excluded from this research, as it is only possible to search for people or events and not for city posts by location or hashtag. Therefore, social networks Twitter and *Instagram* were selected for data collection.

3.2.1. Twitter

Obtaining data from *Twitter* is fairly simple because they provide an API with all the operations needed. Their API integrates a *REST* protocol for communication. According to Codecademy, *REST* is an architectural style providing standards for communication between client and server side. These communications usually have four types:

- *GET* - Retrieve data from the server;
- *POST* - Create new data in the server;
- *PUT* - Update data in the server;
- *DELETE* - Remove data from the server.

The access to these operations is done with endpoints Smartbear. Most of the times there are libraries dedicated to making the communication with these accesses simpler. In this case, *Twitter4J* Twitter4J.

At last, it is important to mention that for this solution to work is necessary to create a developer account at *Twitter*. This step is to obtain the security keys provided by *Twitter* for accessing their API.

The *Twitter* solution is composed by two components: one for authentications and the other responsible for communications with the API. For the authentication process an *HashMap* W3school was created for storing the values associated to each security key. Afterwards, the *HashMap* is sent to *Twitter* for validation and the authentication token is returned.

Moving on, the authentication is passed for the component responsible for communication. In here the two parameters have to be defined: the hashtags/keywords to search and set the geolocations restrictions. With these parameters set, the request is sent for the proper *Twitter* endpoints and the tweets are returned.

At last, the data returned is trimmed to only the necessary values and the stored data is updated to add the new entries.

3.2.2. Instagram

The solution designed for *Instagram* differs from *Twitter* since there is no API. Unlike *Twitter*, this solution will be composed by only one component that executes all the scraper processes. The library used for scraping is called *JSoup* Jsoup. In short, *JSoup* provides functionalities for working with real-world HTML, fetching URLs data and manipulating it. The process for scraping *Instagram* data can be breakdown into the following:

- Extract all the HTML from the public page;
- Separate the data related to each post;
- Trim the and collect the first half of data for each post;
- Collect the second half of data from the individual post page;
- Store the data.

It is important to notice that the data is collected from two different pages: one with all the public posts published according to a hashtag and the specific post page. The process for working the data is done through JSON and strings manipulation. At last, the stored JSON is updated to add the new data.

GDPR. As is known, in the recent years GDPR have been gradually enforced to protect users data and privacy. In fact, due to this most companies simply stopped sharing users data or share them in a very protected and anonymous way. For this reason, all these posts collected belong to public profiles and the information for each post is limited.

3.3. Data Manipulation Process

The data mining process applied in this research is based in four major steps:

- Data cleaning - Removing inconsistent values from the data (i.e. errors, missing values, outliers, etc...);
- Data integration - Ensuring that data from different sources is stored according the established model;
- Data reduction - Reducing the data to only the necessary by removing unnecessary values;
- Data transformation and discretization - Create lighter version of the data if possible to improve performance.

The datasets for this project were composed by the username, the geolocation/address of the post, the time they were posted, the content and the link for the post. Then many variations were created according to the different conditions that were trying to analyse. The final inputs for the clustering algorithm were csv files composed by the geolocation points of the posts we wanted to analyse. Regardless of the approach used, the software must be running periodically to fetch and store data regularly. This can be achieved by deploying the software on a server or on a personal computer that keeps running the software. On the other hand, the software must ensure that it realises 'breaks' so that it does not exceed the maximum number of requests per minute, otherwise it gets blocked.

In this research two programming languages were used: *Java* and *Python*. *Java* is a class-based and object oriented programming language. *Java* allows developers to create well encapsulated code with high levels of cohesion enabling easy refactoring later on if needed. Also, *Java* already has libraries aimed at helping execute common functionalities. On the other hand, *Python* is a interpreted, high-level, general-purpose programming language. Just like *Java*, it is object oriented and possesses many libraries for handling common tasks. In particular, *Python* has libraries particularly dedicated to data mining and therefore becomes an essential tool for this work.

3.4. Data Mining

Data mining is the process that comes after collecting and storing the data. The language elected to handle this process was *Python* since it contains many libraries for handling data and machine learning.

The data mining files were split into two: one responsible for clustering and other responsible for data manipulation. The first step required was to convert the address (which came with the street names) into geopoints. Afterwards, the geocodes are saved in a csv file for better performance when looping through it. On the other hand, a series of scripts were written responsible for creating splitting the data according to weekdays or weekend and according to three different time schedules (morning, afternoon and night). At last, there is a script dedicated to identifying if the users are locals or foreigners. These identifications is highly subject to errors since the methodology was the following: collect all the available locations in each user profile and verify if more than 5 posts are done on the city of Porto. If this is true then the user is considered a local. This methodology can be easily wrong since a user can post many posts in Porto and still not be here however, due to the covid-19 situation it is possible to speculate that more posts were done in each user hometown.

The other file as said previously is dedicated to clustering. For clustering, the library *Scikit-learn* Scikit-learn was used in every operation. Two clustering techniques were selected in this project: *k-means* Stanford and *DBSCAN*. However, the work delved more into *k-means* since the results obtained were more appropriate.

There were support scripts developed for this entire work: one for loading all the required libraries, another for loading json files and one for saving json files. Afterwards, the first script created consisted in creating the map of Porto through *Folium* Folium which will be shared by the many operations. With all these support and shared scripts created the scripts responsible for handling clustering were developed. In a short way, for *DBSCAN* the points had to be passed into the function together with the minimum of samples to be considered a cluster and a plot. As for *K-means*, the points, the map and the number of clusters were passed to the function. The function would iterate through the points and create the clusters identified. For DBSCAN the number of minimum samples selected was 25 and the eps (maximum distance between each) was 0.002. As for KMeans the number of clusters selected was 20. At last, these clusters were marked on the map as circles and the result was stored in the computer.

4. Results And Discussion

The results and discussion section should start by analysing if the data collected has enough quality for providing realistic and reasonable results. The final data set had around 21,000 posts for *Instagram* and 27,000 posts for *Twitter*, from September 2019 to May 2020. Having in mind the current limitations when collecting data from *Twitter* and all the workarounds necessary to gather information of *Instagram*, it is safe to consider that the data obtained is enough to obtain good results.

4.1. Results Obtained

The first result set that will be presented is related to all the data. In the end, 20 clusters were obtained for the city of Porto. However, not all clusters shared the same strength and there were those that clearly stood out. In particular:

- Avenida dos Aliados - The central avenue of Porto.
- Norteshopping - A well-know shopping mall across the country and the largest in the northern region.
- Casa da Música - A worldwide famous concert hall.
- Serralves - A museum with a big garden on one of the most pleasant neighborhoods in the city

The only difference observed in this and further analysis between *Instagram* and *Twitter* is that the *Leça da Palmeira* cluster is eliminated and one is created in *Aeroporto Sá Carneiro*. Aside from this, the remaining clusters share an uniform distribution around them. However, since a large degree of data was concentrated downtown a separate analysis was carried out, focused only on the downtown area. The results obtained focused on the full dataset and the downtown dataset are presented in figure 1.

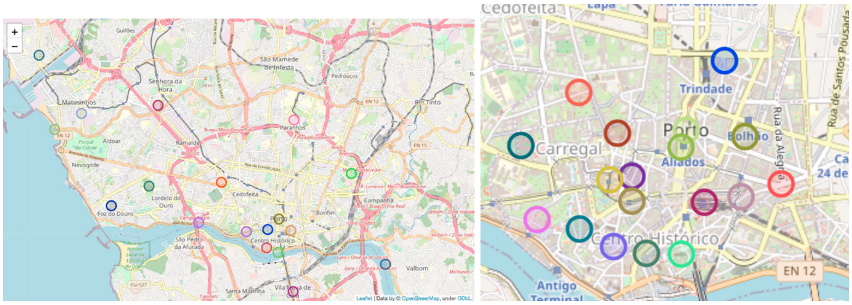

Fig. 1. Clusters obtained for all the data (left) and downtown posts (right and the map is zoomed in on Downtown area).

Analyzing figure 1, it is possible to observe three points that belong to the edge of the map. The first one in the northwest corner belongs to Leça da Palmeira, known for its beaches and restaurants. The south one belongs to the center of Gaia (the most populated city next to Porto). The last one, in the southeast side has small variations however it either belongs to the Freixo marina or the runners/walkway adjacent to it.

Moving on to the points contained in the city it is possible to observe two in the city of Matosinhos. As it is possible to observe, one of the clusters is on the beach and the other one close to it. These two clusters are expected since Matosinhos is the main beach for the city of Porto and the location is filled with interesting food spots. The blue one, is in Foz do Douro, a famous region for its large avenue close to the sea, its expensive/exclusive restaurants and good views. A cluster in a place like this is more than expected. Right close to it, there is a cluster close to Lordelo do Ouro. This cluster refers to the Jardim de Serralves, a large green still in the region of Foz with a museum that promotes original and interesting art expositions and movements. The one close to Senhora da Hora is related to the *Northeshopping* which is one of the biggest shoppings. At last, is possible to observe one cluster close to Cedofeita which is associated with Casa da Musica and one in Paranhos on the university campus. At last, there is one difficult to see close to the southwest side of the map. Just like Matosinhos this is another beach close to the city of Porto and also a runners track.

Overall, these clusters cover most of the city sight views, museums, historic places, beaches and famous restaurants.

4.2. Spatio-Temporal Analysis

The dataset was also split according to mornings, afternoons and nights. Most of the resulting clusters remain the same however, for the morning and afternoon dataset the posts around the beach and close to the Freixo marina disappear and others surge in the city. This is easily explained since people can be sharing more posts around their work place or college.

When analysing the dataset of weekend vs weekday the clusters are mostly the same. However, one major difference is that the clusters closer to the ocean move a bit more towards the beach on weekends. It is possible to assume that this happens since more people go to the beach on the weekend and on weekdays simply to walk close to it. But, is also possible that this is simply a deviation on the cluster algorithm.

When analysing the dataset dedicated for the weekday and ordered by morning, afternoon and night it was possible to conclude that the morning posts don’t have a single cluster on the beach and two new clusters appear on city centers around Porto. Since people spend their mornings on work or academic activities these moves make all the sense.

At last, for the weekend dataset partitioned according to morning, afternoon and night there are some considerable moves. First of all, the weekend morning data set loses almost all the clusters and only focus on tourist spots. It contains 13 clusters only on downtown and the other ones are focused on Serralves, Foz, Estadio do Dragão. Not even the beach sides appear at each of these ones. The best explanation for this is that more tourists post during the weekend morning while most of local population is asleep. Therefore, all the posts are extremely focused on tourist spots. For the weekend afternoon posts the clusters start to normalize but they still don’t reach places like Leça da Palmeira or the beaches in Gaia (two places not so famous as Matosinhos beaches). Only at the night the posts become more uniform and cover all the places for the full dataset except marina do Freixo. This might happen due to the fact that people share at night the posts with places they have been during the day. The figures 2 and 3 represent the variations of results for the different time schedules, tourists and residents. Due to the limitations faced when collecting the data we tried to ”discover” if a user is a tourist or a resident if more than a certain amount of posts in their profile was done in the city of Porto. Like said before, this approach isn't the most effective and we are aware of its limitation but in the context of this project it was the best we could do.

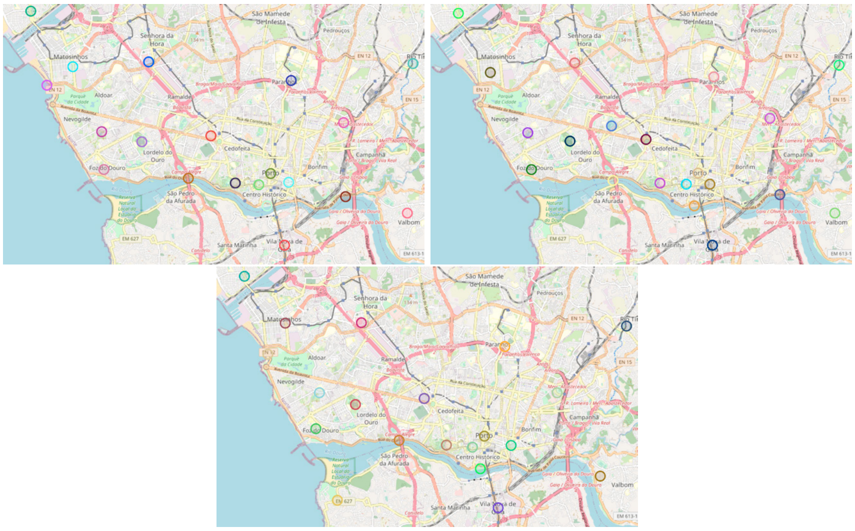

Fig. 2. Results comparison between morning (top left), afternoon (top right) and night (bottom).

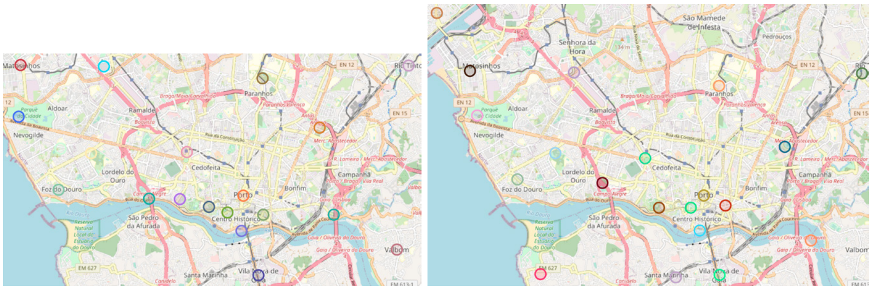

Fig. 3. Results comparison between residents (left) and tourists (right).

4.3. COVID-19

On the day 19th March 2020 the state of emergency was declared and effective immediately due to the COVID-19 pandemic faced by the whole world. During this period all commercial and leisure activities were closed down and people were suggested to stay at home. This situation lasted until the 3rd of May 2020. A different dataset was created for this period to study the impact of COVID-19 in the users behaviour. As expected, all the clusters on beaches and walkways disappeared. These were replaced with clusters either on the city of Porto or Gaia. People spent more time at home and therefore these were the reachable points for photos while respecting the state of emergency.

4.4. All The Maps Created

In order to comply with the number of pages for this paper all the maps with the different variations of the data set can be seen and obtained in this link (the files have to be downloaded and then opened in a web browser for working properly).

5. Conclusions And Future Work

First of all, this research proved to be possible to still use location based social networks for collecting data while being subject to many restrictions due to GDPR. Then, the data collected provided enough information and details for drawing meaningful conclusions about users mobility patterns.

With the many dataset created, it was possible to see how the city of Porto was exploited by the users according to each weekday or weekend and to different time schedules. It was possible to verify if people spend time on the places promoted by the city of Porto or if they spend time on places that had no previous promotion. Actually, at the end of this work is possible to safely say that the most visited places are downtown like *Torre dos Clérigos*, *Ribeira do Porto*, *Avenida dos Aliados*, *Super Bock Arena & Jardins Do Palácio de Cristal*, *Norteshopping*, *Casa da Música*, *Foz do Douro*, *Matosinhos Beach* and *Serralves*. These places are clusters in all the variations of the dataset no matter what. Also, it is possible to observe that at weekdays the posts are more concentrated in the city while at weekends they are more scattered covering more of the outskirts. This concentration is also visible in some schedules.

For future work it would be interesting to maybe develop a better formula for discovering if each user is a tourist, a local or a domestic tourist. Also, it would be extremely interesting to figure if all the public transportation routes cover these desired spots. At last, there was only one park which constantly appeared on the clusters (Parque da Cidade) however, Porto contains much more interesting parks that perhaps should be more visited.

References

Bejar, J., Alvarez, S., Garcia, D., Gomez, I., Oliva, L., Tejeda, A., Vazquez-Salceda, J., 2016. Discovery of spatiooral patterns from location-based social networks. Journal of Experimental and Theoretical Artificial Intelligence 28, 313–329. doi:10.1080/0952813X.2015.1024492.

Codecademy, . What Is Rest?

Diamantini, C., Genga, L., Marozzo, F., Potena, D., Trunfio, P., 2017. Discovering mobility patterns of instagram users through process mining techniques. Proceedings - 2017 IEEE International Conference on Information Reuse and Integration, IRI 2017 2017-Janua, 485–492. doi:10.1109/IRI.2017.69.

Europeanbestdestinations, . European Best Destinations. URL: https://www.europeanbestdestinations.com/.

Folium, . Folium. URL: https://python-visualization.github.io/folium/.

Hammar, K., Jaradat, S., Dokoohaki, N., Matskin, M., 2019. Deep Text Mining of Instagram Data without Strong Supervision. Proceedings - 2018 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2018 , 158–165doi:10.1109/WI.2018.00-94, arXiv:1909.10812.

Jsoup, . Jsoup. URL: https://jsoup.org/.

McGill, . About Urban Planning. URL: https://mcgill.ca/urbanplanning/planning.

Paldino, S., Bojic, I., Sobolevsky, S., Ratti, C., González, M.C., 2015. Urban magnetism through the lens of geo-tagged photography. EPJ Data Science 4, 1–17. URL: http://dx.doi.org/10.1140/epjds/s13688-015-0043-3, doi:10.1140/epjds/s13688-015-0043-3.

Red Hat, . What Is An Api? URL: https://www.redhat.com/en/topics/api/what-are-application-programming-interfaces.

Scikit-learn, . Scikit-Learn. URL: https://scikit-learn.org/stable/.

Silva, O., Amoêdo, N., Almeida, F., 2019. Urban Tourist Motivations in the City of Porto. Ottoman Journal of Tourism and Management Research 4, 445–462. doi:10.26465/ojtmr.2018339521.

Singrodia, V., Mitra, A., Subrata, P., 2019. A Review on Web Scrapping and its Applications. 2019 International Conference on Computer Communication and Informatics (ICCCI) , 376–380doi:10.1109/icgciot.2018.8753030.

Smartbear, . What Is An API Endpoint? URL: https://smartbear.com/learn/performance-monitoring/api-endpoints/.

Stanford, . KMeans. URL: https://stanford.edu/{%}7B{~}{%}7Dcpiech/cs221/handouts/kmeans.html.

Twitter4J, . Twitter4J. URL: http://twitter4j.org/en/.

Valverde-Rebaza, J., Roche, M., Poncelet, P., De Andrade Lopes, A., 2016. Exploiting social and mobility patterns for friendship prediction in location-based social networks. Proceedings - International Conference on Pattern Recognition 0, 2526–2531. doi:10.1109/ICPR.2016.7900016.

W3school, . HashMap. URL: https://www.w3schools.com/java/java{_}hashmap.asp.

Wu, X., Huang, Z., Peng, X., Chen, Y., Liu, Y., 2018. Building a spatially-embedded network of tourism hotspots from geotagged social media data. IEEE Access 6, 21945–21955. doi:10.1109/ACCESS.2018.2828032.