# A Python Library for Exploratory Data Analysis and Knowledge Discovery on Twitter Data

Mario Graff[1,3] Daniela Moctezuma[1,2] Sabino Miranda-Jimnez[1,3]
Eric S. Tellez[1,3]

[1]INFOTEC Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación, Circuito Tecnopolo Sur No 112, Fracc. Tecnopolo Pocitos II, Aguascalientes 20313, México
[2]CentroGEO Centro de Investigación en Ciencias de Información Geoespacial, Circuito Tecnopolo Norte No. 117, Col. Tecnopolo Pocitos II, C.P., Aguascalientes, Ags 20313 México
[3]CONACyT Consejo Nacional de Ciencia y Tecnología, Dirección de Ctedras, Insurgentes Sur 1582, Crédito Constructor, Ciudad de México 03940 México

This work has been submitted to the Special Issue on Data and Information Services for Interdisciplinary Research and Applications in Earth Science of the Computers & Geosciences Journal

**Abstract**

Twitter is perhaps the social media more amenable for research. It requires only a few steps to obtain information, and there are plenty of libraries that can help in this regard. Nonetheless, knowing whether a particular event is expressed on Twitter is a challenging task that requires a considerable collection of tweets. This proposal aims to facilitate, a researcher interested in Twitter data, the process of mining events on Twitter. The events could be related to natural disasters, health issues, people's mobility, among other studies that can be pursued with the library proposed. Different applications are presented in this contribution to illustrate the library's capabilities, starting from an exploratory analysis of the topics discovered in tweets, following it by studying the similarity among dialects of the Spanish language, and complementing it with a mobility report on different countries. In summary, the Python library presented retrieves a plethora of information processed from Twitter (since December 2015) in terms of words, bigrams of words, and their frequencies by day for Arabic, English, Spanish, and Russian languages. Finally, the mobility information considered is related to the number of travels among locations for more than 245 countries or territories.

## 1 Introduction

Twitter is a well-known social network where users let others know their opinions, feelings, or any information in a succinct way. Users of the social network are both real people and organizations, connected due to their particular interests. These properties, and its worldwide popularity, make Twitter a vital data source for many research communities interested in capturing people's opinions. Due to users' culture of openness, most messages are indeed public, and these messages can be accessed through Twitter's API; please note that several requirements need to get access, but those are easier to comply with than other social networks. In Twitter, cyclic events such as weekends, Christmas, New Year, Valentine's Day, among others, naturally emerge by looking only at the tweets' frequency. For this reason, it can be said that Twitter reflects in some ways

the social behavior and its information have a high correlation with what is happening in the world [1]. Furthermore, the fact that people use social media to share and acquire information on any events gives an important clue into social media data's potential. Nonetheless, other relevant events are hidden in the overwhelming amount of information generated daily, and to retrieve them, a specific query is needed. In general, it is not straightforward to know whether the social event or phenomenon is being expressed in Twitter data.

The information collected from social network providers has been used to give information during a humanitarian crisis. For example, Facebook Disaster Maps[1] [2] provides access to dynamic maps that include location information, cell site connectivity, and phone battery charging. Social media, particularly Twitter, has been used to obtain situational information (see [3]), which is a post that can help persons make a decision or obtain information in an emergency [4]. Another use is the development of a surveillance system, e.g., to detect an outbreak of avian influenza [5]. It has also been used to assess the damages of earthquakes [6–8]. As can be seen, social media data has probed usefulness in many research problems. Imran *et al.* [9] provides a depth review of social media usage for emergency's analysis.

This contribution presents an open-source Python library, called *text models*,[2] that aims to help researchers and enthusiasts access aggregated Twitter as its data source, simplifying an event's data acquisition and processing. More detailed, we collected messages from the Arabic, English, Russian, and Spanish languages, some of them since Dec. 11, 2015. Our motivation to collect these languages is twofold, to cover distinct language families collecting one representative: Germanic (English), Slavic (Russian), Semitic (Arab), and Romance (Spanish) families; and to consider most spoken languages, according to The Ethnologue 200[3], English (1st), Spanish (4th), Standard Arabic (6th), and Russian (8th). In particular, we use the public Twitter's stream to listen to tweets with and without geotagged information. We process these messages into a set of useful per-day and per-country aggregated information such as vocabulary usage and user's mobility, among other useful information. We hope this information converts into useful knowledge allowing a better understanding of social phenomena that have resonance on Twitter. We also present two usage scenarios and code samples in this manuscript, one using text and another using mobility information to complement our library's description.

The rest of the manuscript is organized as follows. Section 2 describes several types of research using Twitter data. Section 3 provides an overview of the data collected from four languages over the globe. The analysis of text and mobility aspects are presented in Section 4. Finally, Section 5 is dedicated to summarize and conclude this manuscript.

## 2 Related work

Twitter and Facebook data have a lot of important usages. Current literature has demonstrated that social media helps examine events such as disaster management, mobility, public health, politics, furthermore pandemic situations.

For instance, in [2], a disaster maps with Facebook's data is presented. These maps are related to the Facebook population, movement, power availability, network coverage, and displacement. With these maps, the authors look to measure the population as the difference in the number of users based on pre-crisis levels to observe what areas are more affected by the crisis. For instance, what pair of places register major movements by users is registered, and those places where users charge their mobile phones. Moreover, the ability of network covering is another aspect recorded by Facebook Data. Events like a pandemic can also be analyzed using social media data, in [3] the COVID-19 epidemic is characterized, taking into account that people use social media to acquire and exchange information. Here, the Weibo data, the major micro-blogging site in China, was classified into seven types of situational information related to COVID-19. As situational information was considered, those related to 1) caution and advice; 2) notifications and measures

[1] `https://dataforgood.fb.com`

[2] Repository available at `https://github.com/INGEOTEC/text_models`

[3] `https://www.ethnologue.com/guides/ethnologue200`

been taken; 3) donations; 4) emotional support; 5) help-seeking; 6) doubt casting and criticizing; and 7) counter-rumor. For this, both content and user features were used to classify a major of unknown data based on 3000 COVID-19 related posts manually labeled into these seven categories. The predictive models' performance regarding accuracy was 0.54, 0.45, and 0.65 from Support Vector Machine, Nive Bayes, and Random Forest classifiers, respectively. Traffic phenomena are also analyzed using Twitter data. Ribeiro *et al.* [1] proposed a method to identify traffic events on Twitter, geocode them, and display them on a web platform in real-time. The method employs exact and approximate string matching to traffic event identification; for the string matching, the authors manually listed the most frequent terms used for traffic situations. Human mobility is another important event analyzed using Twitter data [10]. Jurdak *et al.* propose a proxy for human mobility using geotagged tweets in Australia. The population's mobility patterns were analyzed using a dataset of more than 7 million tweets from 156,607 users. The patterns found using Twitter data were compared with other technology types, such as call data records. McNeill *et al.* [11] proposed to estimate local commuting patterns using geotagged tweets. They found that Twitter's information is a good proxy to infer the local commuting patterns; even the bias imposed by the demographic of Twitter users is not significant. Natural disaster management, such as earthquakes, could also be analyzed by using Twitter information. In [7] is proposed a method based on tracking users' comments about earthquakes to generate an own Mercalli scale [4] computed only with Twitter data. Municipalities were the unit of analysis trough several features acquired at that level, such as number of tweets, average words, question marks, exclamation marks, number of hashtags, the occurrence of *earthquake* word, municipality population, among others.

As can be seen, event detection on Twitter is the task of discovering phenomena using the users' messages (tweets). Even without prior knowledge of the particular event we are looking for, we could automatically extract it. For instance, Kanwar *et al.* in [12] classify a set of popular events into different categories with more than 5 million tweets and more than 14,000 users in 10 months. Liu *et al.* [13] make a step forward, trying to discover the core semantics from events in social media. Another example is presented by [14], where Twitter's potential as a data source for analyzing urban green spaces is tackled.

Twitter has also been used for detecting and verifying real-time news events. Liu *et al.* in [15] propose a system for detecting news events and assessing their veracity on Twitter.

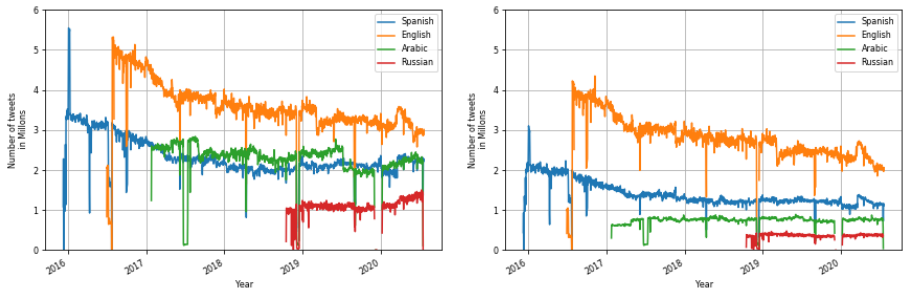
In conclusion, it can be said that there is much current literature related to the usage of social media information, mainly Twitter, trying to solve a vast amount of different problems. For this reason, one of the main contributions of this work is providing the appropriate tools to use Twitter as a data source to facilitate data management for any kind of investigation that researchers want to do.

# 3 Data Acquisition

For all languages, we use the Tweepy package to download public tweets[5]. There are two ways for this download process: one to obtain tweets without geotagged information and the other to retrieve geotagged tweets. In our case, the queries were designed to maximize the number of tweets collected for each language; to carry out this, the first query searches for the most common words for one specific language, i.e., the so-called *stopwords*, that means, for instance, in English, we look for words like *the, also, well, etc.*, in Spanish words like *el, los, como, etc.* using both accented and unaccented version, if it applies. This way, we can download the maximum number of tweets for a specific language because we use common tokens around 400 language-dependent words allowed by Twitter queries. Moreover, simultaneously, the language code provided in the streaming JSON file is used as a query parameter. Meanwhile, the second one also includes a

[4]Qualitative measure used to express the perceived intensity of an earthquake in terms of damages according to [7].

[5]Official site of the Tweepy package, available at `https://www.tweepy.org/`




(a) Tweets without geotagged information. (b) Geotagged tweets.

Figure 1: Number of tweets for Spanish, English, Arabic, and Russian languages.

geo-query that includes the geographic coordinates of the entire world, latitude and longitude values, and the language.

Figure 1 presents the number of tweets (in millions) that have been processed per day. Figure 1a presents the number of tweets retrieved without geotagged information; this was the first download process launched. From the figure, it can be observed the starting point of the collection for each language. Spanish was the first language being collected by us (December 12, 2015), followed by English (July 2, 2016), then Arabic (February 17, 2017), and the last one corresponds to the Russian language (October 31, 2018). From this figure, some particularities can be seen, such as the number of tweets for English and Spanish languages present a decreasing trend that stops around April 2017. This decreasing tendency may be caused by a substantial re-organization of the social media user distributions[6]. Please note that the decreasing slope is not present in the other languages since the collection started around or after April 2017. On the other hand, it can be observed that a bump arose in the English and Spanish languages around mid-February 2020, whenever the COVID-19 epidemic started in America.

Figure 1b complements the information presenting the number of geotagged tweets (in millions) per day. As can be seen, the curves' characteristics are similar in both figures being the difference that the number of tweets with location data is roughly half of the total number of tweets per language. The geotagged tweets also provide the country information. To have an insight of the distribution, in 2019, the majority of tweets have been produced in the United States (36%), this is followed by Saudi Arabia (7%), Great Britain (6%), Argentine (6%), Spain (5%), Russia (5%), Mexico (4%), Egypt (2%), Colombia (2%), and Canada (2%). The rest of the countries contribute to 25 % of the total of geotagged tweets produced.

## 3.1 Data's preprocessing and cleaning

The tweets collected are processed by day using the tokenizers of our previous development [16]. That is, the text is split into words and bigrams of words, as well as q-grams of 2, 3, and 4 characters, setting the text to lowercase, removing the punctuation symbols, users, and URLs. Space is replaced by symbol ∼, and, consequently, the words in a bigram are joined by it. Only those tokens that appear at least 0.01% of the retrieved tweets per day are included.

## 3.2 Mobility

The geotagged information, as georeferenced tweets, can serve as a proxy to estimate people's mobility. The process used to estimate it can be divided into two stages. The first one was performed only once, and the second one is responsible for actual mobility values. The first step computes geographic points, i.e., landmarks, that will provide the lowest resolution in the space. The idea is that each geotagged tweet will be associated with a landmark. The landmark set

[6]The rise of social media, `https://ourworldindata.org/rise-of-social-media`; revised in July 2019.

Figure 2: Set of geographic points used to calculate mobility

is a subset of the bounding-box centroids of all the tweets in the collection. The centroids kept appeared in more tweets than 1% of the number of days collected for the analyzed country. It is worth to mention that those geotagged tweets that provide a specific location are not considered in this process. The landmark set is depicted in Figure 2, where all the points belonging to a country have the same color; the result is that the countries can be observed in the picture. A final note regarding this algorithm, during the process of replacing the bounding box of the geotagged tweets with the centroid, it was found bounding boxes that included the whole country, so these bounding boxes are labeled to use them only on mobility analysis at the level of countries.

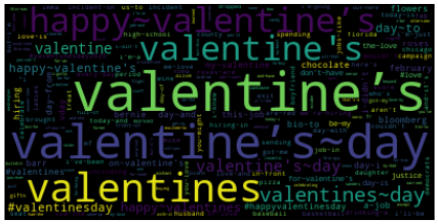
The second step corresponds to the count of the trips between the elements of the landmark set, previously defined. Each landmark is represented with a unique identifier. Then, each geotagged tweet is associated with the identifier of the closest element of the landmark set. In the case of the tweets, localization is a bounding box, and its centroid replaces the bounding box. Then the tweets are grouped by the user and sorted by time. The process iterates for all the tweets of a user; comparing the position of each tweet with the place of the following tweet; in the case, the length between the two locations is higher than 100 meters, then it is recorded as one trip from the first tweet identifier to the second tweet identifier. This process continues for all the users that published some message on that day. A limitation of this algorithm is that it does not consider trips that occur on different days. However, this decision was taken to favor efficiency in terms of time – the algorithm can be run in parallel at the period of a day – and memory since the number of users increases when one considers more days.

# 4 Analysis

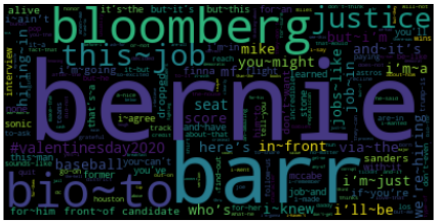
The analysis presented in this section is about two aspects, the first one, the text (tweets's content), the second one the mobility computed with the geographic information of tweet. These two types of data are the most useful and used for Twitter.

## 4.1 Text

Tweets are text messages limited to at most 280 symbols and contain the information that users share. Analyzing this information could give a huge insight into any event or social phenomena. Even only seeing the word frequencies over time may be helpful. In this sense, the word clouds are a useful tool to visualize these word frequencies easily. Figure 3 presents word clouds for the United States and Mexico on two different days. The word clouds were obtained by removing

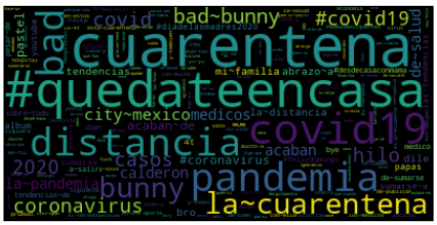



(a) United States word cloud of February 14, 2020. Valentine's Day.




(b) U.S. word cloud (2020-02-14) without tokens of previous years.


(c) Mexico's word cloud of May 10, 2020, it corresponds to Mothers Day.




(d) Mexico's word cloud (2020-05-10) without tokens of previous years.

Figure 3: United States and Mexico's word clouds obtained by removing q-grams, frequent tokens, and emojis.

the q-grams, and frequent tokens[7]. Emojis were also removed due to compatibility issues on the plotting library used to create the word cloud. Figures 3a and 3b show the United States' word cloud on February 14, 2020; whereas Figures 3c and 3d depict Mexico's word cloud on May 10, 2020, which corresponds to the celebration of the Mother's Day in Mexico.

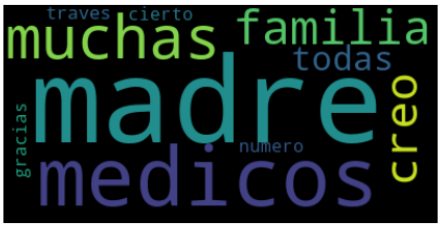
As can be seen in Figures 3a and 3c (left of the figure), word clouds represent only the holiday, and any other event is hidden behind it. That is, on Valentine's Day, the most common word is *valentine's* and its variations; on the other hand, on Mothers' Day, the most frequent word is *madre* (mother in Spanish) and its variations. A simple procedure, to retrieve other events present on holiday and probably not related to it, is to remove the tokens that appeared in the holiday's previous year. Removing the tokens of the previous years produce Figure 3b and 3d (right of the figure). In the United States, the topic was about *Bernie Sanders*, whereas in Mexico, the topic was about *COVID-19* with tokens regarding *cuarentena* (quarantine in Spanish). These examples analyze the tweet's content and frequencies at different dates, yet other phenomena and analysis can be performed without restrictions beyond data availability.

### 4.1.1 Topic Modeling

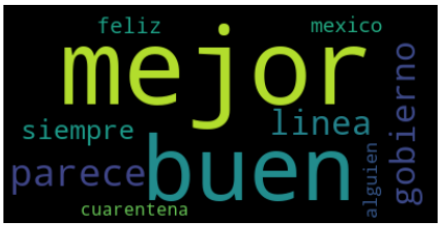
Another analysis carried out typically with text, in this case with the tweets, is topic modeling. To know what topics were included in a set of tweets, we use LDA (Latent Dirichlet Allocation) method [17]. LDA is a generative probabilistic model for data collections, such as text. LDA's idea is to group the text into topics, these topics have a representative set of words, and each document is then assumed to contain multiple overlapping topics. Each topic can be described by a set of words that hold the highest probability of association with that topic. We use Gensim[8]

[7]Frequent tokens are the vocabulary obtained from randomly selecting 5,000,000 tweets from the whole collection and keeping only those tokens that appeared at least 0.1% of the time.

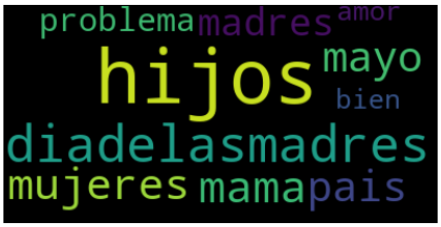
[8]https://pypi.org/project/gensim/



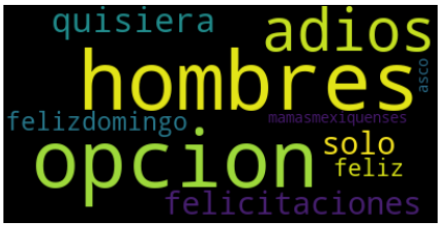

(a) Most representative words on topic 1




(b) Most representative words on topic 2




(c) Most representative words on topic 3




(d) Most representative words on topic 4

Figure 4: Word clouds of most representative 10 words in each generated topic on the date 2020-05-10 from Mexico.

library to compute the LDA model. In this sense, the *text models* library provides a vocabulary of terms with their frequencies. From that data, a corpus was generated to be the input of the LDA algorithm. This corpus is only the text repeated in its frequency indicated by *text models* removing emojis and q-grams.

As a result of topic modeling, four topics were generated only for 2020-05-10, which corresponds with the Mother's Day in Mexico. Figure 4 shows the clouds related to each one of these four topics. For simplicity, only the Spanish language from Mexico was considered. The figure shows words that could be related to Mother's Day; nevertheless, there are some words such as "m édicos" and "gobierno" (doctors and government in Spanish), which are odd in this particular day, but it is perfectly understandable by the pandemic circumstances.

The topics can also be visualized using a histogram. Figure 5 shows words and their corresponding weight and frequency for each topic. Please note that high-frequencies do not necessarily imply large weights for the LDA model. The topic analysis allows us to discover patterns or events on Twitter.

#### 4.1.2 Similarity between Spanish-speaking Countries

The different branches of a language are named a dialect. In each of them, diverse terms could be used for different things or concepts. The study of these dialects can help us understand cultural aspects among regions and their closeness [18]. For instance, Twitter posts can serve as a proxy to study the similarity between Spanish language variations.

Using the *text models* library, the tokens and their frequency, grouped by country, can be used to model, for example, the similarity of a particular language in different countries. Figure 6 depicts the similarity of Spanish-speaking countries, i.e., Mexico (MX), Chile (CL), Spain (ES), Argentina (AR), Canada (CA), Colombia (CO), Peru (PE), Australia (AU), Venezuela (VE), Dominican Republic (DO), Paraguay (PY), Ecuador (EC), Uruguay (UY), Costa Rica (CR), El Salvador (SV), Panama (PA), Guatemala (GT), Honduras (HN), Nicaragua (NI), Bolivia (BO), Cuba (CU), and Equatorial Guinea (GQ).

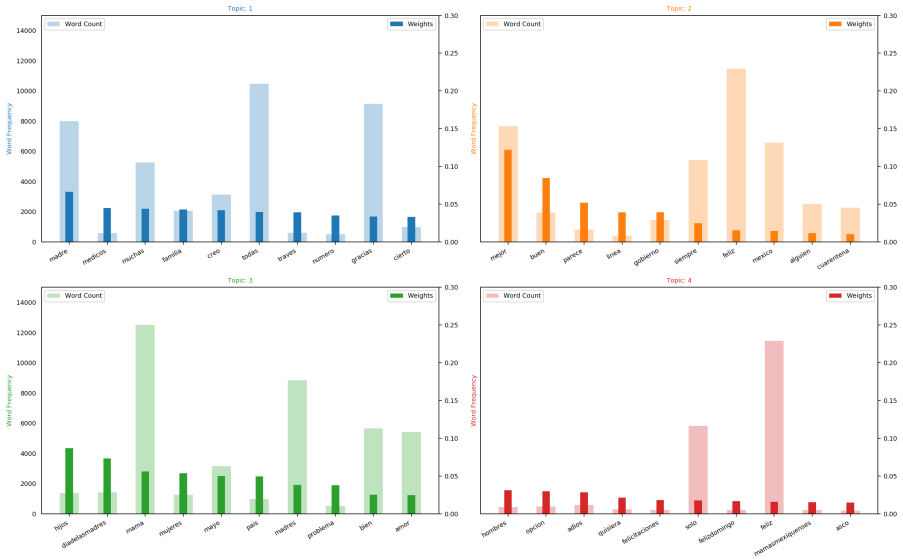
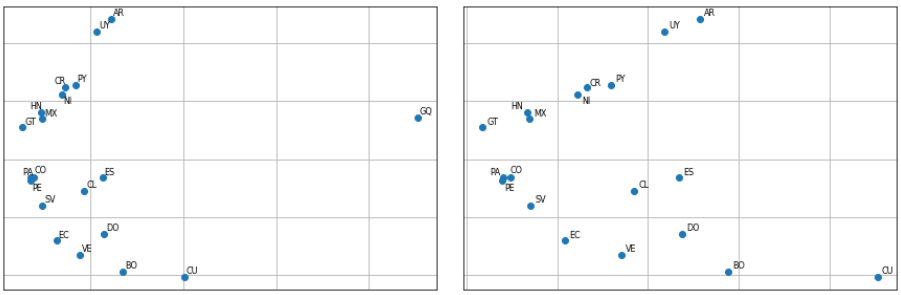



Figure 5: Keywords frequencies and their weights for each topic on date 2020-05-10 for Mexico's Spanish messages.



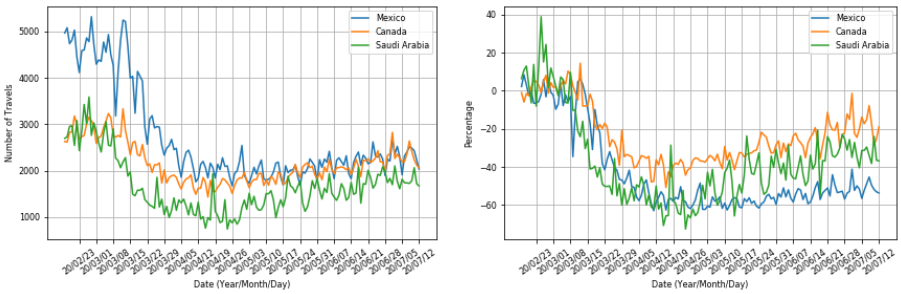

(a) Spanish-speaking countries. (b) Zoom of Spanish-speaking countries.

Figure 6: PCA's representation of Spanish-speaking countries computed on a matrix of the Jaccard similarity between words and bigrams of words among the countries.

The procedure followed was to create a vocabulary with words and bigrams (of words) of 180 days selected randomly from January 1st, 2019 to July 14th, 2020. The vocabulary was created for each Spanish-speaking country. Once the vocabulary is obtained, the Jaccard similarity is used to measure all the pairs' similarity. The preceding process creates a similarity matrix transformed with Principal Components Analysis (PCA) to depict each country in a plane.

Figure 6a depicts all Spanish-speaking countries; it can be observed that the most different country is Equatorial Guinea (GQ), it is worth to mention that Equatorial Guinea has three official languages, namely Spanish, French, and Portuguese.

Complementing the information, Figure 6b shows a zoom of the countries without Equatorial Guinea. From the figure, it can be observed that Mexico (MX) is closest to Honduras (HN) than to Guatemala (GT), which share a border. Argentina and Uruguay can be seen as a cluster; these countries share a border and are south of the continent. On the other hand, there is another cluster formed by Colombia (CO), Panama (PA), and Peru (PE). Geographically, Colombia in the middle between Panama and Peru. Another cluster is composed of Costa Rica (CR) and Nicaragua (NI), and close to this cluster is Paraguay, which is South America country. The rest of the countries cannot be seen as belonging to a cluster, and the one more different is Cuba in this figure. Then, it could be said that the geographic boundaries matter when it spoke of language similarities.




(a) Number of travels per day.

(b) Percentage of mobility per day.

Figure 7: Mobility on Mexico, Canada and Saudi Arabia using two measures.

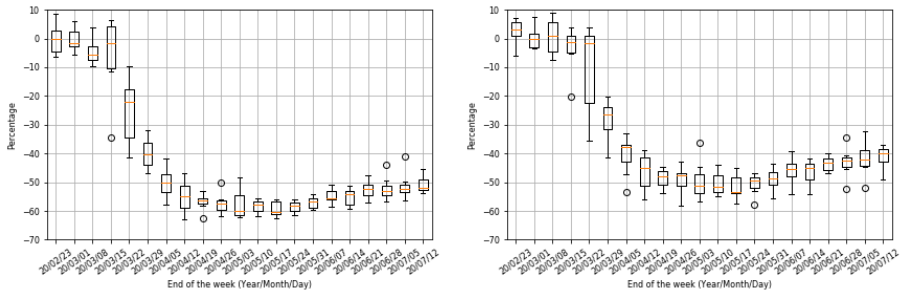
## 4.2 Mobility

In a particular region, mobility can be computed using the number of outward, inward, or inside travels, as well as the aggregation of all mobility events. Different measures can represent mobility; these could be, for example, the number of travels or the mobility trend seen as a percentage between a baseline and a specific date. The percentage of change computed as the number of trips into a rate by considering a baseline period. A plausible approach followed on different research works is to compute the average of mobility per weekday. Usually, the baseline period is previous to the period of interest. For example, Facebook Disaster Maps [2] uses 5-13 weeks previous to the crisis to compute the baseline, and the percentage is computed using the baseline information of the same weekday and hour period. Figure 7 illustrates the mobility in three countries (Mexico, Canada, and Saudi Arabia) using two measures, namely the number of travels (see Figure 7a) and the percentage of mobility (see Figure 7b) in 21 weeks where the baseline corresponds to the 13 weeks previous to the analyzed period.

From Figure 7, it is observed a decreasing trend in mobility. The trend started around March 12, 2020, for Mexico and Canada, and a couple of days before for Saudi Arabia. The lowest point appears one month later and then starts a positive slope. Saudi Arabia's slope is more acute than Canada's slope, whereas Mexico's slope is barely visible. It can be seen in Figure 7a a mobility pattern, to facilitate the reader all the dates correspond to Sundays. The pattern followed is that mobility has its lowest point on Sunday; it steadily increases until reaching the maximum value on Friday and decreases again to reach the minimum on the next Sunday. The weekday mobility pattern is used to compute the percentage, as shown in Figure 7b. From the figure, let us analyze the period before the decreasing trend started. Saudi Arabia's mobility is around zero percent; on the other hand, Canada's mobility is around -10%, whereas Mexico's mobility is -15%, with valleys reaching almost -40%. The behavior presented on Mexico's mobility highlight the necessity to use as baseline other structure instead of the weekday. That is, there are applications where it is desirable to have mobility around zero percent in order to identify more easily a positive or negative trend.

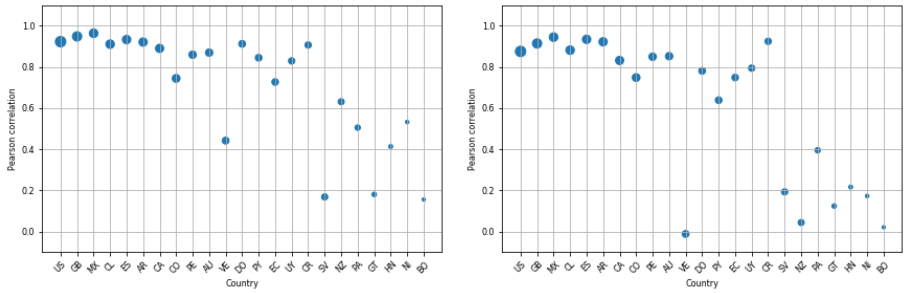
Using a statistic per weekday to compute the percentage relies on the fact that the mobility pattern depends on the weekday; however, there are situations where this is not the case, for example, the mobility of a long weekend Monday is more similar to a Sunday than to a standard Monday. Instead of identifying each day's particular characteristics and then developing an algorithm that considers these differences, one can abandon the idea that mobility can be grouped by weekday and use the mobility with a clustering algorithm to find the groups automatically.

The idea of automatically finding the groups is explored using the well-known k-means algorithm. That is, the mobility in the baseline period is used to train a k-means (where $k$ corresponds to the highest value of the Silhouette score obtained by varying $k \in [2, 7]$), then the percentage of mobility is computed as the rate between the closest centroid and the mobility value at hand. To facilitate the comparison between the two procedures to compute the percentage. Figure 8 presents boxplots of Mexico's mobility, where each boxplot is obtained with the mobility of a




(a) Computing percentage with weekday. (b) Computing percentage with k-means.

Figure 8: Boxplot of the Mexico's mobility on percentage.




(a) Mobility percentage using weekday. (b) Mobility percentage using k-means.

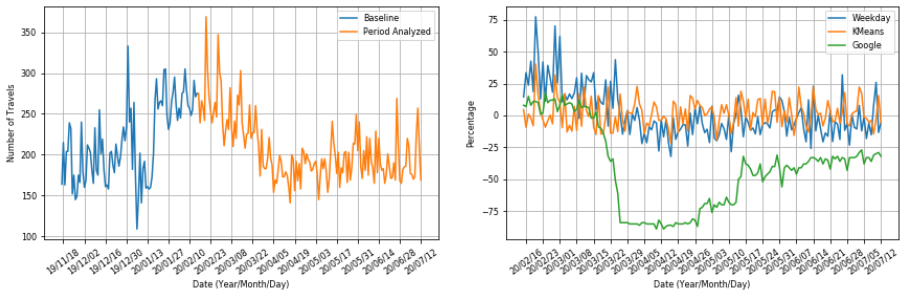
Figure 9: Correlation between Google's mobility report and the one calculated using Twitter data on different countries.

week, and the date is the last day of the week.

Figure 8a presents Mexico's mobility computing the percentage with the weekdays. Before March 22, 2020, it can be observed that the median is below -10%, and there are weeks where the minimum value is almost -40%. On the other hand, Figure 8b shows mobility when the percentage is computed with the centroids obtained with k-means. In the latter case, the third quantile is closed to zero percent before March 22, 2020. There is an outlier on the week ending on March 15, 2020, close to -30%. After March 22, 2020, the lowest valley is below -60% on Figure 8a, and the median is above -60% on Figure 8b. As can be seen, there is a difference of approximately 10% in the two procedures. This remarks on the importance of calculating and defining the baseline pattern of interest to observe the changes.

The mobility obtained from Twitter data can be compared to Google's mobility reported in different countries. Figure 9 presents the correlation, specifically Pearson correlation, between these two mobility measurements, on the United States (US), Great Britain (GB), Mexico (MX), Chile (CL), Spain (ES), Argentina (AR), Canada (CA), Colombia (CO), Peru (PE), Australia (AU), Venezuela (VE), Dominican Republic (DO), Paraguay (PY), Ecuador (EC), Uruguay (UY), Costa Rica (CR), El Salvador (SV), New Zealand (NZ), Panama (PA), Guatemala (GT), Honduras (HN), Nicaragua (NI), and Bolivia (BO). These countries correspond to the Spanish-speaking countries and English-speaking countries. Cuba and Equatorial Guinea are not part of the comparison due to the lack of enough data. The markers' size indicates using a logarithm scale, the median of the number of travel in the baseline, and this number of countries.

The mobility obtained from Twitter corresponds to the percentage of change computed with the median of the weekday in the baseline period (see Figure 9a, and the percentage computed with k-means on the baseline (see Figure 9b). From the figures, it is observed that the majority of countries present a correlation above 0.8. It is observed that the lowest correlations correspond




(a) Baseline and period of interest. (b) Google and Twitter with percentage.

Figure 10: New Zealand mobility.

to the countries with fewer travels, taking into consideration the number of travels. Nonetheless, Venezuela is the exception with a median of 473 travels, whereas El Salvador (SV) has a median of 248 travels. From El Salvador to Bolivia (with a median of 22 travels), the correlation is below 0.8.

Comparing the results presented in Figure 9a and 9b, it is observed that in the majority of cases, the correlation obtained using the median of the weekday is higher than using the k-means algorithm. There is a country where the difference is higher than the other countries, which correspond to New Zealand; the correlation obtained using the weekday's median is around 0.7, whereas using k-means is close to zero.

To analyze with more in-depth detail New Zealand's mobility, Figure 10 presents the mobility obtained from Twitter data as well as Google's mobility report. Figure 10a shows mobility using the number of travels, including the baseline period. On the other hand, Figure 10b presents the mobility computed by Google and the one obtained with Twitter using the two procedures to compute the percentage, namely weekday's median and k-means.

The baseline period, presented in Figure 10a, starts with mobility around 200 travels per day, then has the lowest movement, which is lower than the lowest mobility on the period of study. The study period has two peaks; a valley follows these, and, finally, there is a small increment. On the other figure, Google's mobility starts above zero, it is constant for around four weeks, and then it presents a deepest negative slope that settles below -75%. It is worth mentioning that the mobility behavior is similar in all the countries, albeit the percentage is different. Mobility's percentage using weekdays shows a small decrement when Google's mobility is at its lowest point, whereas the percentage using k-means is constant, around zero, in all the periods. The procedure used found in the baseline two clusters one for the valley, and the other for the rest of the mobility, these two clusters are the reason that the percentage on the period of interest is around zero, let us remember that the baseline has a lower mobility than the period of interest. On the other hand, the weekday median is barely higher than the valley on the study period, which produces a negative slope improving the correlation with Google's mobility.

# 5 Conclusions

We presented the *text models* library, which retrieves a plethora of information collected from Twitter in terms of tokens and their frequencies, and the mobility measured in different countries. The information is collected using the public API, where the query tries to maximize the number of tweets obtained in Arabic, English, Spanish, and Russian languages.

The library aims to facilitate a newcomer, the data mining of Twitter data serving as an exploratory data analysis tool to know whether a particular event is reflected on Twitter. The capabilities of the library are illustrated with different applications performed with the library. We started showing how it can be used to create word clouds on different languages and how the information from previous years can highlight events that, in other circumstances, are opaque by

a more popular one. The tweets content with the topic modeling technique is also done, and with this, it can be seen the most representative words in each topic generated. On the other hand, we exemplify how the library can analyze the difference between dialects. In particular, we presented a comparison between Spanish-speaking countries; the results where the results show that there are countries that it is important the geographic position, that is, the dialects are similar between neighbors, but this is not a general rule.

The mobility information is used to produce a mobility report on Mexico, Canada, and Saudi Arabia. It is described how the units used to measure the mobility can be transformed into a percentage, and two procedures are described to perform this transformation. The first one that corresponds to the use of weekday statistics has been previously used in the literature; however, using a clustering algorithm complements this idea tackling an issue seen at the beginning of the period analyzed. We also compared the mobility report obtained against Google's mobility report; the results show that in most countries studied, there is a strong correlation between them; however, some countries are uncorrelated.

In order to fulfill the aim of this contribution, in the appendix, we describe the code needed to replicate some of the applications, in particular, Figures 3a, 3b, 6. We consider that for a person familiar with Python, the work needed to replicate them is acceptable, and in most cases, a couple of lines are more than enough to produce the desired output.

# A Library usage

The appendix aims to illustrate the use of the *text models* library by replicating some of the applications presented in this manuscript. Specifically, Figures 3a, 3b, 6, and 7 are replicated using the library which correspond to the applications presented using the tokens and their frequency and the mobility information. The first step is to install the library; the mobility examples follow this, and the last part corresponds to the use of tokens.

## Preliminaries

*Text models* library can be installed by different means; however, the easiest way is to use the pip package installer. The following code installs the library using the command line.

```
pip install text_models
```

## Mobility

Once the *text models* library is installed in the system, it can be used to retrieve the mobility and the vocabulary used on different periods of time. To illustrate the library's use, let us replicate Figure 7, where the mobility is presented on the period contemplating from February 15, 2020, to July 12, 2020. The following code retrieved the mobility information on the specified period.

```
from text_models import Mobility
start = dict(year=2020, month=7, day=12)
end = dict(year=2020, month=2, day=15)
mob = Mobility(start, end=end)
```

Figure 7a presents mobility as the number of travels in Mexico, Canada, and Saudi Arabia. The following code lines compute mobility in all the countries. The first line counts the trips that occurred within the country and the inward and outward movement. The information is arranged in a *pandas' DataFrame* or a dictionary, depending on whether the *pandas*' flag is activated. The second line generates the plot for the countries of interest, i.e., Mexico (MX), Canada (CA), Saudi Arabia (SA).

```
data = mob.overall(pandas=True)
data[["MX", "CA", "SA"]].plot()
```

Figure 7b presents the information as a percentage. The percentage is computed using a baseline period, which corresponds to the 13 weeks previous to the event of interest. The baseline statistics can be computed using different procedures; in this contribution, two are described; one using the weekday and using a clustering algorithm, particularly k-means. The *text models* library has two classes; one computes the percentage using weekday information, namely MobilityWeekday, and the other using a clustering algorithm, i.e., MobilityCluster.

The following code lines compute the percentage using the weekday information; the code is similar to the one used to produce Figure 7a being the only difference the class used.

```
from text_models import MobilityWeekday
mob = MobilityWeekday(start, end=end)
data = mob.overall(pandas=True)
data[["MX", "CA", "SA"]].plot()
```

## Text

The other studies that can be performed with the library are based on tokens and their frequency per day segmented by language and country. We are in the position to replicate Figure 3a that corresponds to the word cloud produces with the tokens of retrieved from the United States in English on February 14, 2020. The first line of the following code imports the Vocabulary class, and the third line instantiates the class specifying the language and country of interest.

```
from text_models import Vocabulary
day = dict(year=2020, month=2, day=14)
voc = Vocabulary(day, lang="En",
                 country="US")
```

The tokens used to create the word cloud are obtained after removing the q-grams, the emojis, and frequent words.

```
voc.remove_qgrams()
voc.remove_emojis()
voc.remove(voc.common_words())
```

The word cloud is created using the library *WordCloud*. The first two lines import the word cloud library, and the library used to produce the plot. The third line then produces the word cloud, and the fourth produces the figure; the last is just an aesthetic instruction of the plot library.

```
from wordcloud import WordCloud as WC
from matplotlib import pylab as plt
wc = WC().generate_from_frequencies(voc)
plt.imshow(wc)
plt.axis("off")
```

As shown in the previous word cloud, the most frequent tokens are related to Valentines' day. A procedure to retrieve other topics that occurred on this day is to remove the previous years' frequent words, as depicted in Figure 3b.

```
voc.remove(voc.day_words())
```

The word cloud is created using a similar procedure being the only difference the tokens given to the class.

```
wc = WC().generate_from_frequencies(voc)
plt.imshow(wc)
plt.axis("off")
```

Figure 6 is created, taking randomly 180 days, from the period of January 1, 2019, to July 14, 2020. The *pandas* library has a function that handles date periods. Consequently, we use this library to gather the dates needed.

```
import pandas as pd
dates = pd.date_range("2019-01-01",
                      "2020-07-14")
dates = list(pd.Series(dates).sample(180))
```

Once the date is selected, we can retrieve the tokens on the specified dates, using data from the Spanish-speaking countries. The first instruction, from the following code, defines the Spanish-speaking countries. The second instruction retrieves the tokens and their frequency for each country.

```
countries = ['MX', 'CO', 'ES', 'AR',
             'PE', 'VE', 'CL', 'EC',
             'GT', 'CU', 'BO', 'DO',
             'HN', 'PY', 'SV', 'NI',
             'CR', 'PA', 'UY', 'GQ']
vocs = [Vocabulary(dates, lang="Es",
                   country=c)
        for c in countries]
```

As done with the word clouds, the q-grams and emojis are removed from the vocabulary.

```
for voc in vocs:
  voc.remove_qgrams()
  voc.remove_emojis()
```

The Jaccard similarity matrix is defined in set operations. The first instruction of the following line transforms the vocabulary into a set. It can be seen that the output is a list of sets, each one corresponding to a different country. The second instruction builds the matrix. There are two nested loops, each one iterating for the country sets.

```
tokens = [set(x) for x in vocs]
X = [[len(p & t) / len( p | t)
      for t in tokens] for p in tokens]
```

Each row of the Jaccard similarity matrix can be as the country signature, and to depict this signature in a plane, we decided to transform it using Principal Component Analysis (PCA). The following code transforms the matrix into a matrix with two columns.

```
from sklearn.decomposition import PCA
X = PCA(n_components=2).fit_transform(X)
```

Given that a two-dimensional vector represents each country, one can plot them in a plane using a scatter plot. The second line of the following code plots the vectors in a plane; meanwhile, the loop sets the country code close to the point.

```
from matplotlib import pylab as plt
plt.plot(X[:, 0], X[:, 1], "o")
for l, x in zip(countries, X):
  plt.annotate(l, x)
```

# References


[1] Slvio S. Ribeiro, Clodoveu A. Davis, Diogo Renn R. Oliveira, Wagner Meira, Tatiana S. Gonçalves, and Gisele L. Pappa. Traffic observatory: A system to detect and locate traffic events and conditions using Twitter. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-Based Social Networks, LBSN 2012 - Held in Conjunction with ACM SIGSPATIAL GIS 2012*, pages 5–11, New York, New York, USA, 2012. ACM Press.

[2] Paige Maas Facebook, Shankar Iyer, Andreas Gros Facebook, Wonhee Park Facebook, Laura Mcgorman Facebook, Chaya Nayak Facebook, P Alex, and Dow Facebook. Facebook Disaster Maps: Aggregate Insights for Crisis Response & Recovery. In *Proceedings of the 16th ISCRAM Conference*, pages 1–12, Valencia, 5 2019.

[3] Lifang Li, Qingpeng Zhang, Xiao Wang, Jun Zhang, Tao Wang, Tian Lu Gao, Wei Duan, Kelvin Kam fai Tsoi, and Fei Yue Wang. Characterizing the Propagation of Situational Information in Social Media During COVID-19 Epidemic: A Case Study on Weibo. *IEEE Transactions on Computational Social Systems*, 7(2), 4 2020.

[4] Sarah E. Vieweg. *Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications (Doctoral dissertation, University of.* PhD thesis, University of Colorado at Boulder, 2012.

[5] Samira Yousefinaghani, Rozita Dara, Zvonimir Poljak, Theresa M. Bernardo, and Shayan Sharif. The Assessment of Twitters Potential for Outbreak Detection: Avian Influenza Case Study. *Scientific Reports*, 9(1):1–17, 12 2019.

[6] Shalini Priya, Manish Bhanu, Sourav Kumar Dandapat, Kripabandhu Ghosh, and Joydeep Chandra. TAQE: Tweet Retrieval-Based Infrastructure Damage Assessment during Disasters. *IEEE Transactions on Computational Social Systems*, 7(2):389–403, 4 2020.

[7] Marcelo Mendoza, Brbara Poblete, and Ignacio Valderrama. Nowcasting earthquake damages with Twitter. *EPJ Data Science*, 8(3):1–23, 2019.

[8] Shalini Priya, Saharsh Singh, Sourav Kumar Dandapat, Kripabandhu Ghosh, and Joydeep Chandra. Identifying infrastructure damage during earthquake using deep active learning. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2019*, pages 551–552, New York, NY, USA, 8 2019. Association for Computing Machinery, Inc.

[9] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing social media messages in Mass Emergency: A survey. *ACM Computing Surveys*, 47(4), 6 2015.

[10] Raja Jurdak, Kun Zhao, Jiajun Liu, Maurice AbouJaoude, Mark Cameron, and David Newth. Understanding human mobility from Twitter. *PLoS ONE*, 10(7), 7 2015.

[11] Graham Mcneill, Jonathan Bright, and Scott A Hale. Estimating local commuting patterns from geolocated Twitter data. *EPJ Data Science*, 6:24, 2017.

[12] Sartaj Kanwar, Rajdeep Niyogi, and Alfredo Milani. Discovering popular events on twitter. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9790, pages 1–11. Springer Verlag, 2016.

[13] Weidong Liu, Xiangfeng Luo, Zhiguo Gong, Junyu Xuan, Ngai Meng Kou, and Zheng Xu. Discovering the core semantics of event from social media. *Future Generation Computer Systems*, 64:175–185, 11 2016.

[14] Helen Victoria Roberts. Using Twitter data in urban green space research: A case study and critical evaluation. *Applied Geography*, 81:13–20, 4 2017.

[15] Xiaomo Liu, Quanzhi Li, Armineh Nourbakhsh, Rui Fang, Merine Thomas, Kajsa Anderson, Russ Kociuba, Mark Vedder, Steve Pomerville, Ramdev Wudali, Robert Martin, John Duprey, Arun Vachher, William Keenan, and Sameena Shah. Reuters tracer: A large scale system of detecting & verifying real-time news events from twitter. In *International Conference on Information and Knowledge Management, Proceedings*, volume 24-28-October-2016, pages 207–216, New York, NY, USA, 10 2016. Association for Computing Machinery.

[16] Eric S. Tellez, Sabino Miranda-Jiménez, Mario Graff, Daniela Moctezuma, Ranyart R. Suárez, and Oscar S. Siordia. A simple approach to multilingual polarity classification in Twitter. *Pattern Recognition Letters*, 94:68–74, 2017.

[17] David M Blei, Andrew Y Ng, and Jordan@cs Berkeley Edu. Latent Dirichlet Allocation Michael I. Jordan. Technical report, 2003.

[18] Gonzalo Donoso and David Sánchez. Dialectometric analysis of language variation in Twitter. Technical report, 2017.