Toward a Privacy-aware Data Collector for Economic and Urban Analytics

Miguel Nunez-del-Prado, Bruno Esposito, Ana Luna
Universidad del Pacífico
Av. Salaverry 2020
Lima - Peru

{m.nunezdelpradoc, bn.espositoa, ae.lunaa}@up.edu.pe

Abstract

Nowadays, there are a mature set of tools and techniques for data analytic, which help Data Scientist to extract knowledge from raw heterogeneous data. Nonetheless, there is still a lack spatio-temporal historical dataset allowing to study everyday life phenomena, such as vehicular congestion, press influence, the effect of politicians comments on stock exchange markets, the relation between food prices evolution and temperatures or rainfall, social structure resilience against extreme climate events, among others. Unfortunately, there are few datasets combining different sources of urban data in order to carry out studies of phenomena occurring in cities (*i.e.*, Urban Analytics). To solve this problem, we have implemented a Web crawler platform for gathering a different kind of available public datasets.

1 Introduction

Providing citizens with free access to raw data is one of the new global trends. These data, generated on a daily basis, could come from different sources such as governmental entities, NGOs, companies or Public Administration Entities, social networks, newspapers, congestion services, *etc.* Therefore, the data format must be a standard to make easier the access, use, generation of information and sharing. Thus, it is crucial that governments and private organizations, which have valuable data in their systems, servers, and databases make available these datasets for common benefits but taking into account citizen's privacy.

Unfortunately, Latin American countries and in particular Peru lacks of historical data available to citizens. In May 2017, the government published Legislative Decree[1] to create the National Authority for Transparency and Access to Public Information; and, strengthen the Regime of Personal Data Protection. This first step would allow not only greater transparency on the part of certain institutions but also the possibility of the citizens of becoming a partner and author of solutions that could improve the life quality of our society. Peru is beginning to generate an open data culture and developing an open data portal at the national level[2]. Nonetheless, some phenomena need fine-grained data. For instance, in a research paper (Srivastava, 2017), the author highlights the need to collect segregated data of urban poor for inclusive urban planning. The huge scarcity of segregated data does not allow to make a comprehensive understanding of their vulnerabilities. Segregated data of urban poor are essential for inclusive planning and to build sustainable cities. Undoubtedly, there are many benefits of having segregated data of urban poor in urban planning, not only for inclusive planning but also to understand the vulnerability, to know the contribution of urban poor in urban economy and to prioritize actions.

Our main contribution is presenting an alternative to gathering daily basis generated data (from public sources for storing, organizing and sharing) to perform Urban Big Data Analytics under a privacy aware framework in a developing country such as Peru. The information collection has followed a sanitization process, which assures the identity safety of citizens and brands located in Peru. The aim of this platform is to provide an urban dataset for studying different phenomena in urban environments such as urban planning, on-

[1] Legislative Decree 1353: "Decreto legislativo que crea la autoridad nacional de transparencia y acceso a la información pública, fortalece el régimen de protección de datos personales y la regulación de la gestión de intereses"

[2] Sistema Nacional de Información Ambiental: sinia.minam.gob.pe/

line emergency detection, vulnerability, climate change, resilience and even poverty.

The present paper is organized as follows: Section 2 describes the related works, Sections 3, 4 and 5 detail the framework architecture, the collected data and data statistics of some datasets, respectively. Then, Section 6 shows an application of Urban Analytics. Finally, Section 7 concludes our work.

2 Related works

Open Data promotes innovation thoughts societal participation with the use of the data. Such datasets include measurement data from city-wide sensor networks on smart cities as well as from citizen sensors. In the current section, we present some efforts for data collection to tackle urban planning problems, to detect emergencies and to show city insights in real time.

Concerning data collection for urban planning, (Rathore et al., 2016) propose a smart city data collection platform. This platform gathers information about floods, water usage, traffic, vehicular mobility traces, parking lots, pollution, social networks and weather from smart homes, smart parking, vehicular networking, water & weather and environmental pollution monitoring systems. The authors use the collected information for urban planning decision-making. Nevertheless, in (Santos et al., 2017), authors claim a need to give some context to this kind of measurements. Thus, they proposed the Human-Aware Sensor Network Ontology for Smart Cities (HASNetO-SC) to describe knowledge associated with data collection from city-wide sensor networks with an appropriate level of contextual metadata for data understanding. Therefore, they implemented the architecture for data collection in an urban metropolitan area in Brazil. Consequently, the platform opens the possibility that citizens, who have a little to no knowledge about the collection environment and the collected data, to access and process the information.

About emergencies detection, the work of (Xu et al., 2016) proposes a mechanism to gather information from *Weibo*[3] about urban emergency events. The platform discovers What, Where, When, Who, and Why of a given emergency event from Weibo comments. Thus, to complete these pieces of information, the platform relies on *Social Sensors* and *Crowdsourcing* layers. The former receives textual data from Weibo users. The latter extracts the basic elements of an emergency event (what, when, where, who, and why) to provide information for rescue services or decision making.

Regarding urban data gathering, which is an essential element of modern cities, a great challenge appears, such as data volume, velocity, data quality, privacy, and security, among others. In the paper (Panagiotou et al., 2016), authors describe the development of a set of techniques that aim at effective and efficient urban data management in real settings in Dublin city. The solutions were integrated into a system that is currently used by the city. The system can detect multiple types of incidents, each one focusing on a different input source. Hence, the solutions can identify events by analyzing in real-time GPS trajectories, data coming from sensors installed in junctions, or textual information coming from social media. Authors developed analysis modules to forecast load so that they can manage efficiently the volume and velocity. Besides, they combine information to infer events from data anomalies. Noisy data and erroneous measurements are also dealt. Moreover, machine learning algorithms are used to identified relevant tweets avoiding in that way low-quality data. Another real-time data collection application can be found in London where the information can be viewed in a dashboard[4] . More details on this work can be found in the reference (Gray et al., 2016), where its main idea is to understand the city dynamics in a better way. The system gathers data from different third party entities and open data platforms given as *CSV* file, *JSON* object or an *HTML*.

At last, there is also a multi-source dataset of urban life in Milano and Trentino (Barlacchi et al., 2015). The authors put together different data sets, such as spatial grid, social pulse, telecommunications, precipitations, weather, electricity, news, and social pulse. The scientists locate spatially all the data in a grid, which allows comparing datasets generated in various companies using different standards. The idea behind this dataset is to make a testbed for different solutions to urban problems like energy consumption, mobility planning, tourism and migrant flows, urban structures and interactions, event detection, urban well-

[3]Weibo website: tw.weibo.com

[4]London city dashboard: citydashboard.org/london

being, etc.

In the next section, we detail the framework proposed in the present effort.

3 Data gathering framework

In the present section, we describe the different components of the architecture of the Web crawler. Figure 1 depicts the different parts of the architecture, where each part is responsible for a given task as follows:

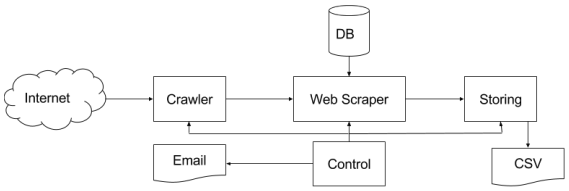

Figure 1: Data gathering framework

Crawler: this artifact is responsible for reading the Uniform Resource Locators (*URLs*) from the database to download the target web pages from different websites.

Data base: this data base engine stores a list of *URL* provided by the user and the rules for reading and extracting relevant fields.

Web Scraper: it receives the downloaded web pages and the extracting rules to parse the web pages for gathering the needed fields of a given web page.

Storing: generate the Comma Separated Vector (*CSV*) files for each treated web site.

Control: verifies the aforementioned parts (*i.e.*, Web Scraper, Crawler and Storing) are alive and are able to perform their task without problems.

The process begins with the manual creation of a list of target *URLs* and the rules needed to extract relevant fields from these. Then, those information are stored in the *Data Base*. The target *URLs* are chosen by the user. The data extraction starts with the *Crawler* reading the (*URLs*) from the *Data Base* to download target Web pages. Then, the *Web Scrapper* receives a set of Web pages as input. Consequently, it generates an index I of all the Web pages in each registered Website. Next, the *Web Scrapper* extract the data from the different gathered web pages, listed in the index I, using the particular structure settled for a given website. It is worth noting that websites structures for extracting data are stored in the database beforehand. For implementing this part of the crawler, we relied on the *Scrappy* library[5].

Extracted data is temporally stored in the database to generate a Comma Separated Vector *CSV* file at the end of a day. Therefore, we store the datasets in a daily basis to build a historical repository. Finally, the last part of the framework is the *Control* mechanism that verifies the state of the crawler, scraper and file generation to notify by email if something goes wrong while gathering data from the different Websites.

In the next section, we detail the different datasets collected by the platform.

4 Collected datasets

In this section, we detail the variables of the collected datasets, which are available[6]. It is important to note that a sanitization process was performed over the datasets. Therefore, we consider as sensible information people's and brands' names, which are pseudonymized and erased, respectively. Concerning the opinion dataset, the comment field is sanitized by erasing stop words and sorting words alphabetically to reduce the impact of a De-anonymization re-link attack (Gambs et al., 2014). Consequently, these sanitization processes are carried out to prevent a privacy breach. Please note that the sanitization process is performed off-line. Thus, the sanitization does not limit the extraction of useful information.

It is possible to extract unstructured datasets from the websites targeted (i.e. news from newspapers, satellite imagery, etc.). In this work, we present two non-tabular datasets: newspapers and social networks. Nonetheless, a tabular structure has been applied to them for easier readability.

In the following paragraphs, we described each of the nine categories of data sources as well as the datasets in each category.

Beauty consists of a description, date, price, category and cosmetics products (*c.f.* Table 1).

Climate category contains data from monitoring stations, atmospheric pollutants and radiation

[5]Scrapy: scrapy.org

[6]BITMAP Urban Analytics: bitmap.com.pe/urbands.html

category	date	price	title
mujer	2016-08-04	69.0	Noir de Nuit
mujer	2016-08-04	0.0	Orianité

Table 1: Sample of beauty dataset.

provided by the National Meteorological and Hydro-graphic Service of Peru (SENAMHI).

Table 2 shows atmospheric pollutants (CO, NO, NO_2, NOX, O_3, PM_{10}, $PM_{2.5}$, SO_2) acquired from several monitoring networks distributed in Metropolitan Lima and the Province of Lima. It also presents the date and time of the measurement.

CO	NO	NO2	NOX	O3
0.6	13.9	19.3	33.2	3.5
PM10	PM2.5	SO2	date	hour
51.6	34.0	2.5	2016-08-04	00:00

Table 2: Sample of Climate dataset.

Table 3 specifies the set of meteorological data (humidity and temperature) of different monitoring stations. It also reports the station name where it was measured, the date and the UTM location of the record (department, district, latitude and longitude and altitude).

altitude	district	date	hum
3928	AYAVIRI	16-08-04	9
lat	lon	temper.	type
14°52'22"	70°35'34"	19	Met

Table 3: Sample of Station dataset.

Table 4 describes the data corresponding to solar radiation levels in different regions in Peru.

arequipa	cajamarca	cusco	puno
8.0	7.0	9.0	9.0
date	ica	junin	tacna
2016-08-04	6.0	9.0	4.0
lima	moquegua	piura	
2.0	8.0	8.0	

Table 4: Sample of UV Radiation dataset.

Markets category reports maximum and minimum prices, description of different products of the first necessity of three different markets of supermarkets and suppliers in Lima.

Table 5 describes the product data for sale in markets. This dataset records the name, category, minimum price, maximum, average and date.

title	type	min_price
acelga	acelga	4.0
max_price	avg_price	date
5.0	4.5	2016-08-04

Table 5: Sample of a Market dataset.

Medicament category comprises prices of medicines provided by the Ministry of Health of Peru (*c.f.*, MINSA). Table 6 shows the registered drug dataset containing the condition of the drug, address, technical director, pharmacy name, price, name, country, and date of manufacture.

Condition	address	director
Con receta	C. Cordoba 2300	Luis Guia
manufacturer	date	Working hours
Hersil	2016-08-27	L-V 9:00-20:20
name	country	regulation
Bot. Pharmalys	Peru	NG1279
price	register	phone
2.5	NG1279	2660488
holder	location	
Hersil	Lince - Lima	

Table 6: Sample of Medicament dataset.

Newspapers category contains news from different print media in Peru. Table 7 describes newspapers dataset composed of the publication date, author, and the section of newspaper.

content	date	author
long text	2014-02-14 13:33:47	journalist name
section	location	title
mundo	mundo/eeuu	Edward Snowden

Table 7: Sample of written press dataset.

Real Estate Market comprehends prices for houses, apartments, and offices sales and rental nationwide. Table 8 shows real estate data grouped in columns detailing whether the state of the property is a sale or rental, describes the property, and its address. Additionally, the longitude and latitude of the property are provided.

title	section	description
text1	alquiler	text2
location	area	price
Rep. Panama	320m	$5000
lon	lat	date
-77.032584	-12.0431	2016-08-06

Table 8: Sample of Real Estate Market dataset.

Social Networks category contains geo-referenced comments from people on different topics and social relations.

Table 9 shows opinions of various users, the language in which the opinion was made as well as the region of origin. It is worth noting that user id was pseudonymized using a hash function. In the same spirit, comments were sanitized to reduce re-identification risk.

user id	timestamp	language
1059254686	1476728629010	es
lon	lat	region
-77.0364	-12.0513	Lima, Pe.
comment		
alza bar cafe centro el futuro gifs los puño		

Table 9: Sample of Opinion dataset.

Table 10 represents the friendship links between users of the social network.

user id1	user id2
1059254686	1059254367
1059254686	2259254876

Table 10: Sample of Social links dataset.

Stock Market category contains two datasets for money exchange rates and stock exchange markets. The former contains the different historical exchange rates, from Soles to other foreign exchange as shown in Table 11.

currency	buy	sell	date
Swiss franc	3.346	3.686	2016-08-04
Euro	3.684	3.819	2016-08-04

Table 11: Sample of money exchange dataset.

The latter dataset also contains the transactions of the Stock Exchange of Lima. This dataset contains the price of the last transaction, opening price, purchase, sale, company, dates, amount, the amount of the stock, operations, and price variation. These variables are detailed in Table 12.

pre	open	sale	company	date
7.8	7.7	7.56	Alicorp	2016-08-04
currency	amountNeg	mnemonic	noShare	noOper
S/	1 325 647	ALICORC1	173 677	16
sector	segm	last	var	sale
IND	RV1	7.56	-3.08	7.7

Table 12: Sample of Stock Exchange dataset.

Transportation contains main avenues traffic jams and domestic as well as international departure and arrivals flights at Jorge Chavez airport in Lima, Peru.

Tables 13 and 14 detail datasets of both alerts and congestion, respectively. These datasets contain information of some points in Lima city. It details the street, city, date, latitude and longitude coordinates of the alerts or the level of congestion. Table 13 also indicates the level of traffic, the node where the traffic level is reset as well as the speed of the traffic.

street	city	date
Av. Los Frutales	La Molina	2016-08-23
latitude	longitude	traffic level
-12.071628	-76.964632	2.0
node	speed	
Calatrava	4.719	

Table 13: Sample of Jams dataset.

In the case of Table 14, the types and subtypes of alerts are gathered in addition to the above-mentioned data.

Street	city	date
Av. Aviación	San Borja	2016-08-23
latitude	longitude	type
-12.086116	-77.003996	JAM
subtype		
JAM HEAVY TRAFFIC		

Table 14: Sample of Alerts dataset.

Table 15 shows the dataset for both arrivals and departures flights from Jorge Chavez airport in Lima, Peru. The name of the airline, the city of origin or destination, the status of the flight, the belt in which the suitcases are delivered. The estimated and scheduled time, and the door and flight number are also reported.

Airline	city	state
Peruvian	Cusco	Landing
belt	date	estimated time
4	2016-07-22	10:50
scheduled time	door	flight
11:00	3	210

Table 15: Sample of Airport traffic dataset.

Table 16 shows the size of the data, the number of attributes and the number of records per data set. On the other hand, Table 17 synthesizes the characteristics linked to data types, headers, and temporal space granularity. Concerning temporal granularity range from 1.02 to 6.84 minutes. With regard to spatial granularity, there are seven datasets georeferenced with UTM coordinates (*i.e.*, latitude and longitude).

Data set	DataFrame	Data points	Attributes	Size (Mb)
Beauty	Beauty	22,190	4	1.2
Stock Market	Money ex.	997	4	0.0
Stock Market	Stock ex.	7,253	16	0.8
Weather	Ate	2,952	10	0.2
Weather	Campo-de-marte	3,291	10	0.2
Weather	Carabayllo	2,796	10	0.1
Weather	Stations	291,839	11	19.3
Weather	Huachipa	2,322	10	0.1
Weather	Puente-piedra	2,968	10	0.2
Weather	Radiacion UV	3,181	11	0.2
Weather	San-borja	3,180	10	0.2
Weather	S. J. Lurigancho	2,565	10	0.1
Weather	S. M. d Porres	2,890	10	0.2
Weather	Sta. Anita	3,233	10	0.2
Weather	V. M. del Triunfo	3,195	10	0.2
Real estate	Real estate3	415,225	9	172.6
Real estate	Real estate1	159,665	9	107.3
Real estate	Real estate2	162,736	9	140.7
Medicines	Medicines	555,7353	14	1,311.6
Markets	Commerce1	136,433	6	12.3
Markets	Commerce2	490,582	5	37.1
Markets	Markets	11,178	6	0.7
Opinion	Opinion	6'979,829	7	239
News.	Newspapers1	1'560,134	6	1,129.6
News.	Newspapers2	13,392	6	23.4
News.	Newspapers3	27,060	6	11.4
News.	Newspapers4	40,451	6	16.1
Transport	Arrival	726,624	9	39.3
Transport	Departure	679,260	9	41.2
Transport	Alerts	351,690	7	33.3
Transport	Jams	1'643,277	8	187.5

Table 16: Summary of different data sets size.

In the next section, we describe some statistics of our datasets.

5 Dataset Statistics

In this section, we detail the statistics of the different dataset in the described categories in Section 3.

Beauty.- Table 18 shows the most interesting characteristics of the beauty dataset, described in Table 1. On the one hand, two of the four attributes of the table contains categorical values (discrete values) for which the mode is the most important. These values are "wrinkles" and "Essential greasy skin" for the *category* and *article* attributes, respectively. On the other hand, we have a numerical attribute that is the price, of which we have the mean, median, standard deviation, minimum and maximum values. It is important to note that, no attribute contains null values (*c.f.*, NAs).

Dataset	Temporal Granularity	Geo referenced	Spatial granularity[1,2]	Types
Beauty	1.02	False	None	float, str, date
Money ex.	3.19	False	None	float, str, date
Stock ex.	1.16	False	None	float, int., str, date
Ate	1.18	False	None	float, TSTP, date
Cpo-de-marte	1.18	False	None	float, TSTP, date
Carabayllo	1.18	False	None	float, TSTP, date
Stations	1.16	True	UTM	float, int., str, date
Huachipa	1.18	False	None	float, TSTP, date
Puente-piedra	1.18	False	None	float, TSTP, date
Radiacion UV	1.16	False	Region	int., date
San Borja	1.18	False	None	float, TSTP, date
S. J. Lurigancho	1.18	False	None	float, TSTP, date
S. M. de Porres	1.18	False	None	float, TSTP, date
Sta. Anita	1.18	False	None	float, TSTP, date
V. M. del Triunfo	1.18	False	None	float, TSTP, date
Real estate3	3.13	True	LL	float, str, date
Real estate1	2.52	True	LL	float, str, date
Real estate2	2.57	True	LL	float, int., str, date
Medicines	1.17	False	District	float, int., str, date
Commerce1	1.2	False	None	float, str, date
Commerce2	1.2	False	None	float, str, date
Markets	3.09	False	None	float, str, date
Opinion	-	True	LL	float, TSTP, str
Newsp.1	3.89	False	None	str, date
Newsp.2	1.18	False	None	str, int., date
Newsp.3	1.35	False	None	int., str, date
Newsp.4	6.84	False	None	int., str, date
Arrivals	1.21	False	None	TSTP, str, int., date
Departures	1.17	False	None	TSTP, str, int., date
Alerts	1.16	True	LL	float, str, date
Jams	1.16	True	LL	float, int., str, date

[1] UTM is the Universal Transverse Mercator coordinate system.
[2] LL is the Latitude - Longitude coordinate system.

Table 17: Spatial-temporal granularity summary and data types of different datasets.

N	variable	type	mean	median	std
0	category	str	-	-	-
1	date	date	-	-	-
2	price	float	935	45	35152
3	article	str	-	-	-
N	mode	min	max	NAs	%NAs
0	arrugas	-	-	0	0
1	-	-	-	0	0
2	0	0	1400000	0	0
3	Essential cutis graso	-	-	0	0

Table 18: Statistics of beauty dataset.

Climate.- We describe the characteristics associated with climate data. Two datasets will be described: 1) data from meteorological stations and their measurements (Table 19); and, 2) data on pollutants by districts (Tables 20). Other datasets are not described due to lack of space.

In Table 19, we have a large number of cate-

N	variable	type	mean	median	std
0	altitude	integer	2143.82	2485.00	1560.81
1	department	str	-	-	-
2	district	str	-	-	-
3	station	str	-	-	-
4	date	date	-	-	-
5	humidity	float	62.60	62.00	226.54
6	latitude	str	-	-	-
7	longitude	str	-	-	-
8	province	str	-	-	-
9	temperature	float	17.03	16.70	41.39
10	type	str	-	-	-
N	mode	min	max	NAs	%NAs
0	3812.00	0.00	5192.00	0	0
1	-	-	-	291839	1
2	-	-	-	291839	1
3	CABO INGA	-	-	272508	1
4	-	-	-	0	0
5	100.00	5.00	45059.00	43820	0
6	12 46' 17.86"	-	-	0	0
7	75 0' 44.52"	-	-	0	0
8	-	-	-	291839	1
9	20.80	-30.80	4974.20	37258	0
10	Meteorologica	-	-	272508	1

Table 19: Statistics of stations dataset.

gorical attributes and represent the characteristics of the monitoring stations. However, each of these monitoring stations measures two meteorological characteristics, which are represented by numerical values. These two attributes have characteristics such as the mean, median, standard deviation, mode, and minimum and maximum values.

In contrast, the pollutant data by districts (*c.f.*, Table 20) contain a large amount of numerical data. As already mentioned in the previous descriptions, some of them are the mean, median, standard deviation, mode, minimum and maximum values.

N	Var.	type	mean	median	std
0	CO	float	1.299	1.2	7.659
1	NO	float	46.657	40.1	31.31
2	NO2	float	18.084	16.5	9.719
3	NOX	float	64.632	58.95	35.193
4	O3	float	7.661	5.4	9.329
5	PM10	float	115.786	107.25	55.978
6	PM2.5	float	35.145	28.9	27.485
7	SO2	float	11.274	9.6	10.499
8	date	date	-	-	-
9	horas	TSTP	-	-	-
N	mode	min	max	NAs	%NAs
0	1.4	0.0	410.9	64	0.02
1	27.0	1.2	263.4	708	0.24
2	19.6	0.1	164.3	720	0.24
3	63.9	0.8	328.7	708	0.24
4	0.5	0.3	198.3	9	0.0
5	94.5	0.0	948.0	10	0.0
6	0.0	0.0	203.0	518	0.18
7	8.7	2.7	353.3	290	0.1
8	-	-	-	0	0.0
9	-	-	-	1	0.0

Table 20: Statistics of pollutants dataset.

Markets.- Here we describe the statistics of a market and super market that sell groceries. This category was described in Table 5.

Table 21 shows statistics about product. In the variable title, the most popular product is *potato*. The most mentioned type is *red-headed onion*. The minimum price has an average value of $2,427PEN$ and varies between $0.57PEN$ and $13.00PEN$. The maximum price is on average $3.057PEN$ and varies between $0.71PEN$ and $14.00PEN$. Finally, the average price is $2.746PEN$ and fluctuates between $0.61PEN$ and $13.5PEN$.

N	Var.	type	mean	median	std
0	title	str	-	-	-
1	type	str	-	-	-
2	min_price	float	2.427	1.5	2.339
3	max_price	float	3.057	2.0	2.828
4	av._price	float	2.746	1.75	2.525
5	fecha	date	-	-	-
N	mode	min	max	NAs	%NAs
0	PAPA	-	-	0	0.0
1	CEBOLLA CABEZA ROJA	-	-	0	0.0
2	2.0	0.57	13.0	0	0.0
3	2.0	0.71	14.0	0	0.0
4	1.25	0.61	13.5	0	0.0
5	-	-	-	0	0.0

Table 21: Statistics of market dataset.

Newspapers.- Table 22 shows the most frequent content that is the *text1*, which is a given news (we prefer not to share the content due to the lack of space). Then, the most cited author, section, and title are *Carlos Battle*, *executive zone*, *world / Current* and *5 tips for a startup to survive*, respectively. It should be noted that the percentage of missing values of content, author, and location are 81%, 82% and 95%.

Real Estate.- This dataset contains three different datasets. We only show one table due to lack of space. Described data are rents and/or real estate sales, the data are mostly categorical.

In Table 23 we have only two attributes with numerical values associated with the location of the property. These data have statistics such as mean, median, standard deviation, mode, minimum and maximum values. On the other hand, we have another group of characteristics with categorical values. Among them, the more frequent section

N	variable	type	mean	median	std
0	content	str	-	-	-
1	date	date	-	-	-
2	author	str	-	-	-
3	section	str	-	-	-
4	location	str	-	-	-
5	title	str	-	-	-
N	mode	min	max	NAs	%NAs
1	texto1	-	-	1264992	0.81
2	Carlos Batalla	-	-	1275035	0.82
3	zona-ejecutiva	-	-	0	0
4	mun./act.	-	-	1480648	0.95
5	5 cons. para que una startup sobre.	-	-	0	0

Table 22: Statistics of newspapers dataset.

variable is *renting*. We can also see that properties of $100m^2$ are the most "offered", using the attribute area. Regarding the price, the most frequent value is $900. It is important to note that, although we have a large amount of null data (*NAs*), these represent a rather low percentage due to a large amount of data available.

N	variable	type	mean	median	std
0	title	str	-	-	-
1	section	str	-	-	-
2	description	str	-	-	-
3	location	str	-	-	-
4	area	str	-	-	-
5	price	str	-	-	-
6	longitude	float	-77.01	-77.03	0.07
7	latitude	float	-77.01	-77.03	0.07
8	date	date	-	-	-
N	mode	min	max	NAs	%NAs
0	Alquiler de...	-	-	0	0.0
1	alquiler	-	-	0	0.0
2	Rento...	-	-	2524	0.02
3	Ubicacin...	-	-	27912	0.17
4	100 m	-	-	2790	0.02
5	US$ 900	-	-	0	0.0
6	-77.03	-77.76	-76.13	0	0.0
7	-77.03	-77.76	-76.13	0	0.0
8	-	-	-	0	0.0

Table 23: Statistics of real estate dataset.

Transportation.- We describe the statistics of the datasets related to the terrestrial transportation mode, As far as the dataset of air transport is concerned, we do not show them due to lack of space. For land transport, Table 24 and 25 report the statistics on alerts and congestion, respectively.

Table 24 shows the most reported street, which is *Av. Javier Prado*. In the *city* variable, that corresponds to the district, San Isidro is the most reported district. More precisely in the coordinates

N	variable	type	mean	median	std
0	rue	str	-	-	-
1	city	str	-	-	-
2	date	date	-	-	-
3	latitude	float	-12.087	-12.09	0.008
4	longitude	float	-77.01	-77.016	0.036
5	subtype	str	-	-	-
6	type	str	-	-	-
N	mode	min	max	NAs	%NAs
0	Av. Javier Prado Este	-	-	0	0.0
1	San Isidro	-	-	0	0.0
2	-	-	-	0	0.0
3	-12.091415	-12.1	-12.067	0	0.0
4	-77.003755	-77.072	-76.949	0	0.0
5	JAM_HEAVY_TRAFFIC	-	-	28659	0.08
6	JAM	-	-	0	0.0

Table 24: Statistics of alerts dataset.

$-12.091415, -77.003755$ where alerts of type *jam* and subtype *jam heavy traffic* are often reported.

Table 25 shows the street and the node where most congestion is reported *Av. Circunvalacin del Golf Los Incas* in the *la Molina* district with an average traffic level of three on a scale of one to five. Finally, it shows the average speed of 2.59 Km/h.

N	variable	type	mean	median	std
0	calle	str	-	-	-
1	ciudad	str	-	-	-
2	fecha	date	-	-	-
3	latitud	float	-12.084	-12.086	0.009
4	longitud	float	-76.995	-76.993	0.035
5	nivel de trafico	integer	3.286	3.0	0.713
6	nodo	str	-	-	-
7	velocidad	float	2.597	2.369	1.302
N	mode	min	max	NAs	%NAs
0	Av. Circunvalacin del Golf Los Incas	-	-	0	0.0
1	La Molina	-	-	0	0.0
2	-	-	-	0	0.0
3	-12.076636	-12.112	-12.061	0	0.0
4	-76.963088	-77.081	-76.941	0	0.0
5	3.0	1.0	5.0	2	0.0
6	Av. Circunvalacin El Golf Los Incas	-	-	0	0.0
7	2.025	0.139	12.289	14288	0.01

Table 25: Statistics of jams dataset.

Stock Market.- This category has two different datasets, which are money exchange and stock market. The former is detailed in Table 26. The latter dataset is not described due to the lack of space in the present work.

N	variable	type	mean	median	std
0	currency	str	-	-	-
1	purchase	float	2.823	3.315	1.134
2	sell	float	2.891	3.397	1.432
3	date	date	-	-	-
N	mode	min	max	NAs	%NAs
0	Swiss franc	-	-	264	0.26
1	0.031	0.025	4.513	146	0.15
2	0.034	0.0	4.827	264	0.26
3	-	-	-	0	0.0

Table 26: Statistics of money exchange dataset.

Concerning the money exchange dataset (*c.f.*,

Tabla 26). We have the half of variables categorical (currency and date) and the another half numerical (purchase and sell). It is worth noting that the mode of the currency attribute is *Swiss franc*. Please note that there are some null values (no data), which are represented by the *NAs* and the percentage is given by *%NAs*.

In the present section, we have not described the statistical metrics of two data categories, which are medicaments and social networks because of lack of space. In the next section, we propose an example of Urban Analytics to show the potential of our datasets.

6 Traffic congestion application

To show a possible application of our datasets, we take the *Congestion* (*c.f.,* Table 13) and *Alerts* (*c.f.,* Table 14) datasets of the *Transport* category to analyze traffic jams in a given district of Lima. Consequently, we filter the congestion reports and alerts of Lince district. Then, we select all records produced in this district. Finally, we extract a CSV file containing traffic data to analyze it.

In the present example, we make a visual analysis of congestion to show the enormous potentiality of our dataset. Subsequently, we rely on Qlik[7] to depict Figure 2.

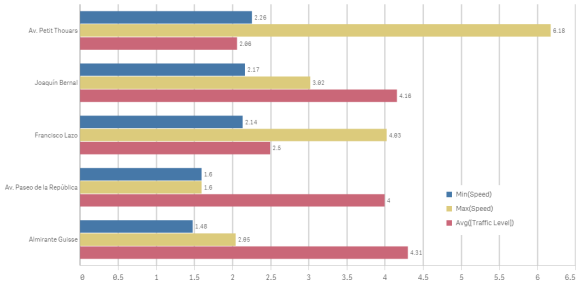

Figure 2: Minimum, maximums and averages speeds in different streets and avenues of Lima

Figure 2 shows traffic level in red, minimal and maximal speed in blue and yellow, respectively. This analysis was done for five different avenues and streets. As we can see, the lower the speed, the higher the traffic level. Another interesting fact is the reported maximal and minimal speed in *Paseo de la Republica* Avenue, which has the same value (*i.e.,* $1.6Km$) meaning a high congestion in this avenue.

[7]Qlik: qlikid.qlik.com

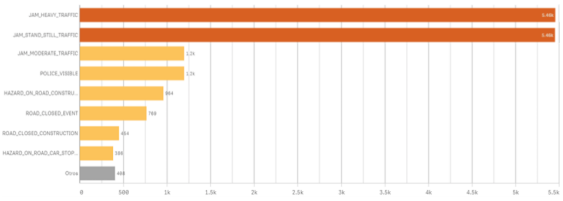

Figure 3: Alerts

Figure 4 shows the distribution of the 17700 alerts gathered with our platform. We note that *Jam heavy traffic* and *Jam stand still traffic* are the most reported alerts with 5500 alerts each one.

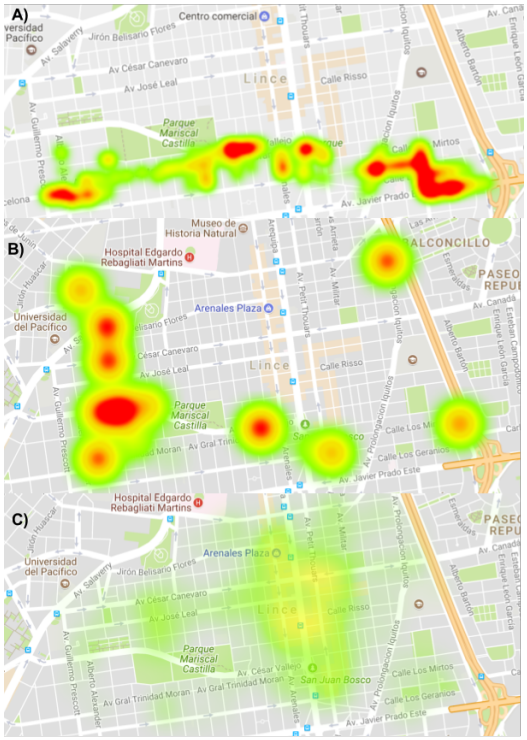

Figure 4: Heatmap of reported alerts (top), real estate (middle) and business locations (bottom) in Lince district.

Finally, using Fusion Tables[8], we can draw a geolocated heatmap of reported alerts and jams in Figure 4A. As we can see, the most reported and congested part is the road exchange between *Javier Prado* Avenue and *Paseo de la República* highway at the bottom right the figure. We have analyzed this segment of the city to study the traffic level between this important road exchange and the location of our university (*i.e.*, Universidad del Pacífico) in the upper left part of the figure.

A simple analysis is already interesting but

[8]Fusion tables sites.google.com/site/fusiontablestalks

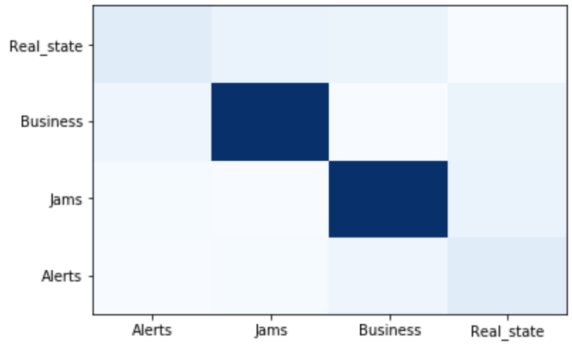

Figure 5: Heatmap of reported alerts (top) and jams (bottom) in Lince district.

crossing datasets could reveal useful insights. Therefore, we use datasets from real estate and business locations. The former is a dataset described in Table 23. The latter is a private dataset containing the code, latitude, and longitude of businesses. Relying on these datasets, we measured the influence of real estate and business locations over congestion. Figure 5 a heat-map (where light and strong blue mean short and large distance, respectively) of the distance among the datasets locations. It is possible to see that alerts are influenced by real estate and business location and jams, only by real estate locations.

7 Conclusion

In the present work, we have described the architecture of a Web crawler platform to gather information about nine different categories of datasets to make urban analytics. The main contribution of this work is the provision of information to the scientific community and policy makers for analyzing and studying social behavior and urban phenomena in a developing country such as Peru. We have collected data following a privacy-aware structure. Sensible information about the citizens or brand-names has been sanitized. These datasets have enabled us to implement a Knowledge Tier Platform for Graph Mining (Nunez-del Prado et al., 2016) and perform urban analytics (Di Clemente et al., 2017) or study urban resilience (Abbar et al., 2016).

In the future, we plan to extend the crawler to collect more information from new public available Web sites. We also plan to make a privacy risk analysis of the described datasets.

References

Sofiane Abbar, Tahar Zanouda, and Javier Borge-Holthoefer. 2016. Robustness and resilience of cities around the world. *arXiv preprint arXiv:1608.01709* .

Gianni Barlacchi, Marco De Nadai, Roberto Larcher, Antonio Casella, Cristiana Chitic, Giovanni Torrisi, Fabrizio Antonelli, Alessandro Vespignani, Alex Pentland, and Bruno Lepri. 2015. A multi-source dataset of urban life in the city of milan and the province of trentino. *Scientific data* 2.

Riccardo Di Clemente, Miguel Luengo-Oroz, Matias Travizano, Bapu Vaitla, and Marta C Gonzalez. 2017. Sequence of purchases in credit card data reveal life styles in urban populations. *arXiv preprint arXiv:1703.00409* .

Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. 2014. De-anonymization attack on geolocated data. *Journal of Computer and System Sciences* 80(8):1597–1614.

Steven Gray, Oliver O'Brien, and Stephan Hügel. 2016. Collecting and visualizing real-time urban data through city dashboards. *Built Environment* 42(3):498–509.

Miguel Nunez-del Prado, Edgardo Bravo, Miguel Sierra, Miguel Canchay, and Isaias Hoyos. 2016. Knowledge tier platform for graph mining in (smart) cities. In *Proceedings of Symposium on Information Management and Big Data*.

Nikolaos Panagiotou, Nikolas Zygouras, Ioannis Katakis, Dimitrios Gunopulos, Nikos Zacheilas, Ioannis Boutsis, Vana Kalogeraki, Stephen Lynch, and Brendan OBrien. 2016. Intelligent urban data monitoring for smart cities. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pages 177–192.

M Mazhar Rathore, Awais Ahmad, Anand Paul, and Seungmin Rho. 2016. Urban planning and building smart cities based on the internet of things using big data analytics. *Computer Networks* 101:63–80.

Henrique Santos, Vasco Furtado, Paulo Pinheiro, and Deborah L McGuinness. 2017. Contextual data collection for smart cities. *arXiv preprint arXiv:1704.01802* .

Ambey Kumar Srivastava. 2017. Segregated data of urban poor for inclusive urban planning in india: Needs and challenges. *SAGE Open* 7(1):2158244016689377.

Zheng Xu, Yunhuai Liu, Neil Yen, Lin Mei, Xiangfeng Luo, Xiao Wei, and Chuanping Hu. 2016. Crowdsourcing based description of urban emergency events using social media big data. *IEEE Transactions on Cloud Computing* .