Privacy in the Digital Age - CS 5436 / INFO 5303

NYC Open Data

Cornell Tech 2018

Ephraim Montag, em789 | Subhangi Agarwala, sa2265 | Svava Kristinsdottir, gsk72

Code Repository
Datasets Tagged

Topic

Every day, NYC collects information about people, places, and things. Much of this information is released to the public in open data. An example of a dataset would be calls made to 311, where the details of these calls are logged. Many of these datasets release identifiers or quasi-identifiers. Our hypothesis is that by using the different databases it may be possible to track people throughout their lives and determine people's day to day activity.

Methodology

Our project goal was finding a way to link the datasets within NYC open data. In order to do so, we wanted to find a methodological process that could be applied to any current existing datasets as well as any future ones. We decided to make tags that could be applied to the columns in each dataset and use those tags to make links between the datasets.

We used a 100 datasets through NYC open data. Our first task was deciding which tags to utilise. We decided to make it possible for multiple tags to be applied to a column in order to maximise the utility of each tag. We did not want to limit ourselves to one tag. This allows us to have general and specific tags. The table below shows the tags we used, their description and the justifications of why they were chosen.

As is shown in the table there is a wide variety of tags used to form possible links of the datasets. We manually reviewed each dataset and marked any columns designated as tags. We created one JSON file per dataset. The JSON files contained columns and any tags associated with that column. We then ran a python script to link the datasets based on the tags and to organize the results. A full list of datasets (json files) is presented with the code provided.

Tag	Justification
name	A name is a very good identifier and therefore very useful in possibly linking datasets.
streetaddress	A good partial identifier that could be used in conjunction with others to identify specific households/individuals.
zipcode	Zipcode on its own can help provide unique identifying information, as well as in conjunction with streetaddress.
longitude	A common column amongst many datasets which helps provide location data.
latitude	A common column that with longitude provides exact coordinates.

gender	A possible way to link data as it appears in multiple datasets.
dates	A good way to get a timeframe from one dataset to another by linking them together.
ethnicity	Can help link datasets and be used as a quasi-identifier.
profession	Can help be used to link datasets and determine trends based on profession.
email	A unique identifier that can be used to link across different datasets.
phone	Another identifier that can be used to link datasets and identify individuals.
location	One of the best identifiers and in conjunction with other data can be used effectively to build a profile on individuals. A broader category to encompass many different aspects of location that either don't fit into other categories or is used in conjunction with others to help build the best possible links between datasets.
licensenumber	License number which can be used to help link datasets with many datasets (involving vehicles for example).
licensetype	Used with license number to help link datasets.
time	Used to help build a timeframe for events. Time is used to build a timeframe for the datasets with the possibility that two events occuring at the same time in different datasets are linked.
persons_killed	A tag that might be able to help link crime related datasets or other datasets involving car incidents.
persons_injured	A tag that might be able to help link crime related datasets or other datasets involving car incidents.
unique_key	A key that might be used in one dataset directly connecting it to a case in another dataset. Further research would need to be done to see if the two were actually related with each other.
vehicle	Used to link different datasets related to transportation with a particular kind of vehicle.
jurisdiction_name	Used to link datasets recording accidents or other reports based on jurisdiction.
number	Used to identify data-type for a column.
percent	Used to indicate a numeric field is a percentage.

violation_type	Used to identify the type of violation associated with crime recorded by different datasets.
amount	Used to identify the quantity of something like amount in dollars.
institution_type	Used to identify the type of institution a record in an entity falls under.
text	Used to indicate that a column is descriptive in nature.
url	Url is used to to see if two or more Urls appear in different datasets.

Results

When we ran the code to detect common tags between datasets we saw some examples of linking between the datasets. Some of those examples are:

The datasets “311 Service Requests from 2010 to Present “ and “Medallion Drivers - Active” are linked at the columns closed_date and expiration_date through the tag ‘dates’.

The datasets “Bicycle Parking and Parking Violations Issued - Fiscal Year 2018” are linked at the columns borough and violation_location through tag the ‘location’.

The number of connections per tag is listed as follows:

Tag	No. of connections
TEXT	288
DATES	2315
TIME	186
URL	10
NAME	4201
ZIPCODE	899
LOCATION	**7191**
STREETADDRESS	3196
LATITUDE	900

LONGITUDE	900
NUMBER	2128
PHONE	103
GENDER	17
ETHNICITY	45
INSTITUTION_TYPE	1
LICENSENUMBER	45
LICENSETYPE	1
VEHICLE	21

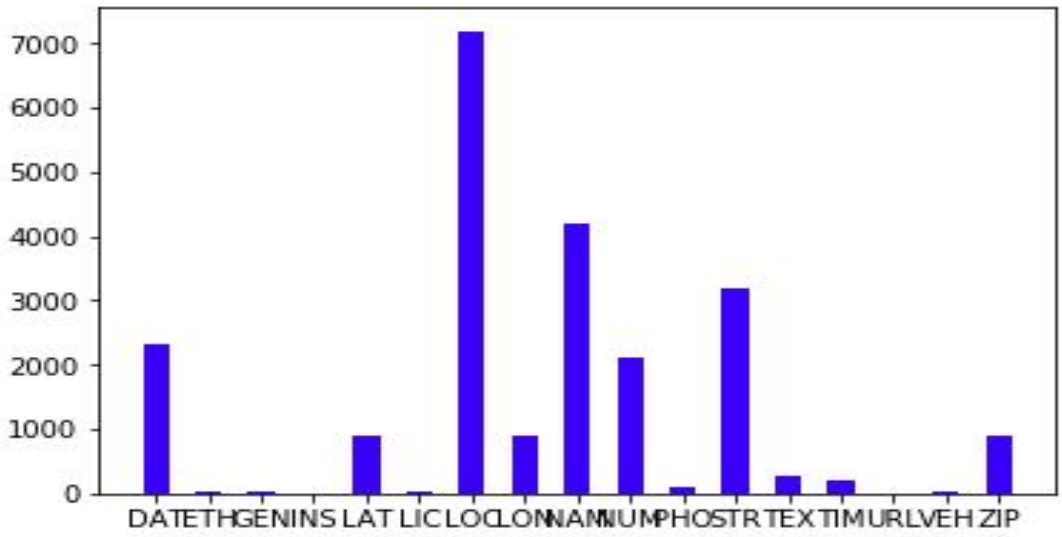

Fig. 1: Visual representation of the no. of connections per tag.

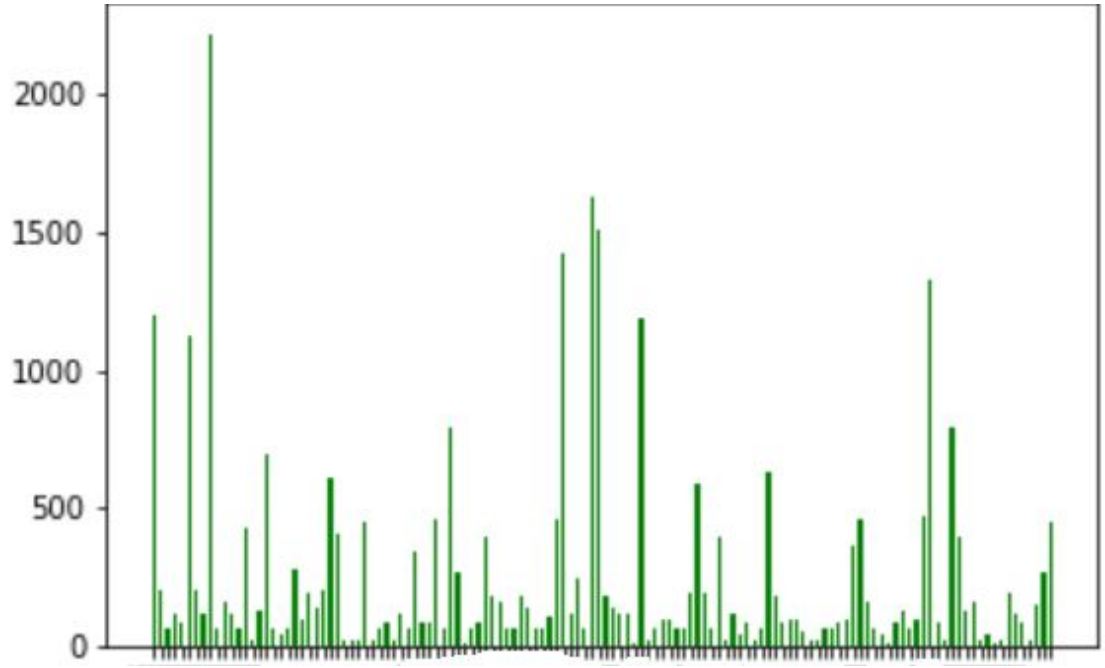

Fig. 2: Number of columns that could be connected

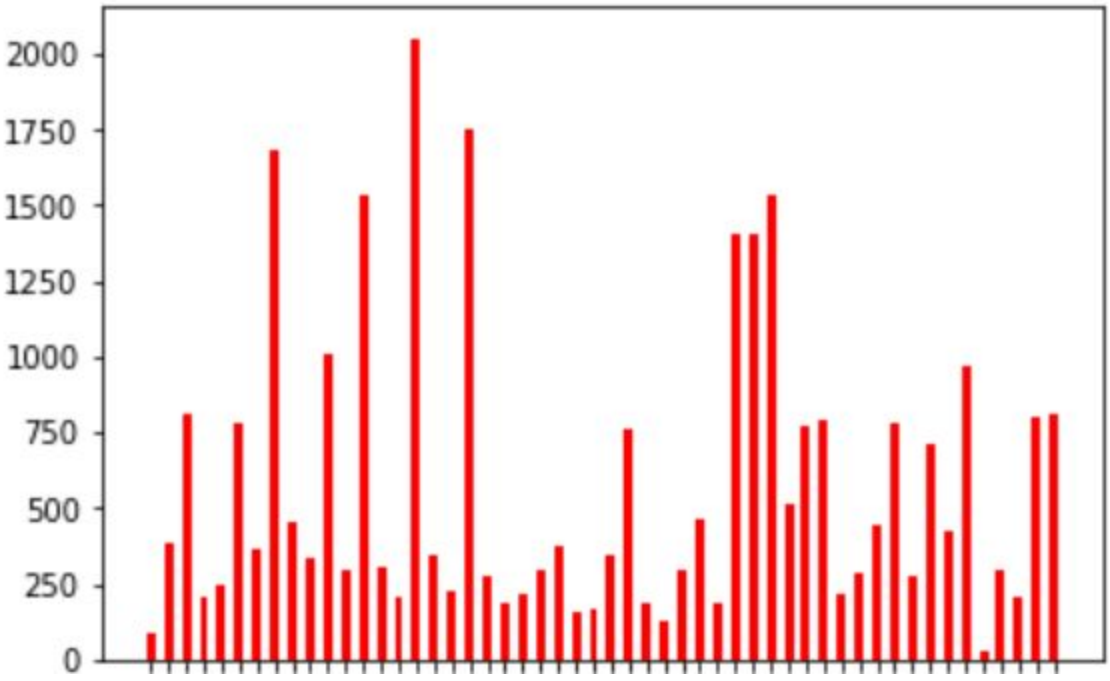

Fig. 3: Number of links per dataset

Discussion

Based on our results, it is quite clear that there are many different possible links between the datasets. Location is the largest link that is seen. Location is arguably the most private information that someone could have about someone else given that it can reveal much private information about individuals. However, in many datasets location of events is given freely. Based on the links between tags, we see that location data might be used to make connections to a specific individual. Coincidentally, the second most common tag is names. Names are a primary identifier. As mentioned this could provide specific insight into a specific individuals activity.

Anonymization can protect privacy, but sometimes identities can be inferred from anonymous data. To minimize privacy concerns, the collection of location data should be minimized. However, for location based services, the “principle of minimal data collection” is not clearly defined. It typically means collecting anonymous or pseudonymous data. However, as per prior research ostensibly anonymous location data can be traced back to it’s personally identifiable with the help of additional data sources as shown by Krumm in his paper “Inference Tracks on Location Tracks.”

Our experiment demonstrates that the location data is extremely sensitive and perhaps should require people’s consent before collecting location data. Location is an uncharted legal territory in the broader privacy debate. Many privacy advocates are trying to extend the current privacy framework to location data, but are far from success at this time.

Other than location data being sensitive it also shows that other data is also sensitive. Dates/Times, phone numbers, gender, vehicle license numbers are all stuff that can be linked from one dataset to another. With enough links possible trends and patterns can be formed perhaps even uniquely identifying people. If noise is not already added to the datasets, it should be to help protect individual privacy.

Limitations

Each group member had to manually review a third of the datasets. Because of lack of coordination between group members on choosing which tags to apply to which columns, some information may have slipped through the cracks. Also due to the manual nature of the work there is a chance of human error. We think that processing about 100 datasets is sufficient for this project, but another limitation may have been that more significant results could have been shown with a larger amount of data.

Works Cited

Krumm, J.: Inference Attacks on Location Tracks. In: LaMarca, A., Langheinrich, M., Truong, K.N. (eds.) Pervasive 2007. LNCS, vol. 4480, pp. 127–143. Springer, Heidelberg (2007)