MIS 4203 - Independent Studies

Privacy Preserving Data Analysis and Mining

R.A.D.K.Rupasinghe

2016MIS022

2018

Introduction

In this mini project, we will explore the potential privacy concerns regarding location data that is supposedly anonymous. We will use a modified version of NYC Taxi data and modified NYC complaints data.

1. Provide summary of the datasets

There are two data sets provided as Taxi Data and Complaints in New York City. Taxi Data contains 20200 records with the Pickup Time, Drop-off Time, Pickup Coordinates and Drop-off Coordinates. There are also other attributes which does not seem to be useful in the given task.

Complaint Data consist of 10,000 records of complaints along with the coordinates of the event which depicts the location, time of the incident, type of offense took place and type of premise. There are also some other attributes which define the complaint, but still seem not to be useful in analyzing the crimes.

1. Analyze dataset and identify find relationship between crime and taxi ride.

With the given two datasets it’s possible to identify that there are taxi data which seems to be related to the complaints in New York. This relationship can be built through the time and location of the taxi data and complaints. For a given taxi ride it’s possible to identify whether there is a complaint made at a near location. As the time reported through the complaints can vary, it’s possible to use a threshold time.

1. Using that relationship find the identity/location of criminal. For each crime you need to provide crime location and time, criminal pickup and drop-off location and time, information about taxi ride and short justification.

Following JSON file contains a complete list of crimes mapped to taxi data.



1. Show and list high crime areas in the New York City

Following list show’s top ten postal codes of NYC along with the number of crimes occurred in each area.

|  |  |  |
| --- | --- | --- |
| **City** | **Postal Code** | **No of Crimes** |
| Manhattan | 10016 | 104 |
| Manhattan | 10027 | 50 |
| Manhattan | 10029 | 44 |
| Queens City | 11434 | 33 |
| Jackson Heights | 11372 | 21 |
| Brooklyn | 11211 | 20 |
| Elmhurst | 11373 | 18 |
| Long Island City | 11101 | 16 |
| Brooklyn | 11207 | 15 |
| Brooklyn | 11217 | 14 |

1. Describe and methods and methodologies used in the analysis.

In order to analyze the data, both data files were loaded to a java program to read them and load only the necessary information from each data set. Then the complaints mapping to each taxi data were filtered using the time and location of each data entry. In order to do so, location information database was created using a reverse geo coding web service, which took inputs as latitudes and longitudes and provided the location information as output. Then the taxi rides which were between the time period of the complaints and the matching locations were selected as crimes related to each taxi data.

Complete project is uploaded to the following github location.

<https://github.com/kinathru/crime-analysis>

The project also contains the python script used to generate the map and the html file generated from the script.

1. Visualize your finding on New York map (You can use Python Map Visualization Library: Folium http://python-visualization.github.io/folium/). Please use different colors per each crime.

