Home Work Assignment 3-B

Out of the given topic, I have selected crime related twitter tweets being posted in and around the Albany region. The twitter data collected has tweets from central Albany, Troy, Guilderland, Schenectady, Rensselaer and region within 10000 miles of radius.

Topic : Crime related twitter data

The motivation for doing this project was primarily an interest in undertaking a challenging project in an area of anomalous pattern detection, a subfield of data mining. This project will gives us an opportunity to learn, how to analyze multivariate time series data for anomaly detection. We are basically focusing towards analysis of twitter data related to crime. The idea behind selecting this topic is that crime is a situation which keeps on changing with increasing distance from the area where it took place. Considering the twitter data related to crime will help us in identifying those tweets which are anomaly and thereby helping us assessing the strength of the policies drafted by various governmental organization for controlling the crime rate. Furthermore, it will also help us identifying which crime (violations, felony or misconduct) was maximum in the state and is, policy related to the crime with highest rate.

Queries used in the topic :

we have taken 1000 tweets covering an area of 10000 miles in and around the city of Albany.

**query = 'crime OR domestic abuse OR kidnapping OR **
**murder OR rape OR robbery OR assault OR terrorism AND **
(commit OR accused OR claim OR bail)'

1. All the tweets containing either crime, domestic abuse or kidnapping
2. From the above identified tweets we search for tweets related to crime
3. After the second step, we are interested identifying all the tweets in which there is a crime and the accused is arrested.
4. Finally, we are mainly interested in identifying those tweets in which the crime is related to rape.

We are using these queries because we are interested in finding all the tweets which are related to "increased sexual crime rate" in and around the region of Albany.

We have collected all the tweets related to crime, in which some tweets are out of context as it does not contain any terms from query. Hence, we can say our data is unbiased. We have calculated the random sample of tweets based on the geographical coordinates of the region in which we are interested.

Our claim that out data is unbiased can be shown by the results of our execution :

Number of D tuples 1000
N: 729
M: 341
A: 30
B: 522
C: 532
API Recall: 0.467764060357
Quality Precision 0.0879765395894
Quality Recall 0.0276752767528

Since the quality recall is very less than API recall, we can say that the data is reasonable.

I haven encountered a problem in the assignment, which is that when I ran my code with max ID and since ID, still I was not able to remove much of the redundancy in the data. I applied sorting based on the twitter id, still no improvement was noticed.s