COMPSCI 597: Big Data and NOSQL Databases

Alpha Release

Tanner, Reuben, Trisca, Gabriel

November 19, 2013

Copy of the dataset (load it on the onyx server and provide the location) (20 points)

We are working with four datasets:

Weather Daily weather summaries from the City of Chicago for the time period 2001 – 2013
Location: http://www.ncdc.noaa.gov/cdo-web/

Crime This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system.
Location: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

Socioeconomic Indicators Census Data - Selected socioeconomic indicators in Chicago, 2007 – 2011
Location: https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2

Public Health Statistics Selected public health indicators by Chicago community area.
Location: https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Selected-public-health-in/iqnk-2tcu

On `onyx`, the dataset can be found at
`/home/students/dtanner/bigdata/*`

Analysis of the data (20 points)

We are developing a predictive analyzer that, based on time, weather, location and other factors predicts what types of crimes will be committed in Chicago. This data set will be processed into a correlation matrix, and training examples for a classifier.

The first step in the analysis was to create a subsample of the data and see if there crimes that happened frequently enough to warrant further analysis. Once this was established, we proceeded to compare correlation between variables in the datasets to identify what crimes are correlated with a location or weather variable to later train a classifier that could predict crimes successfully.

There are many useful patterns that we found in this dataset, and the highly correlated variables are being compiled into a training set for a supervised learning algorithm.

Data cleansing issues that you identified and how did you address them? (10 points)

For the most part, the data was actually quite clean so there were very minimal issues; some typical unstructured data problems but nothing that was irrecoverable. At first, we thought one of the more critical pieces of data we would need is the latitude and longitude coordinates and that is true in many cases but in order to do analysis at a higher level, it seemed better to have a less granular judge of proximity. Through some research we discovered that Chicago is broken up in to approximately 75 "community areas"; unfortunately, the existing rows that had a community area code specified was quite small so our initial analysis largely consisted of figuring out which crimes occurred in which community area by reverse geocoding. Other issues we had were trivial: casing issues which were solved by normalizing all casing during preprocessing and our data was in csv form but there were some comma's within a single field in the set so we performed some extra reg-ex parsing fanciness to overcome it.

Code

The code can be downloaded by issuing:

```
git pull https://github.com/CS597/Eucleia
```