Team Ottawa

Deliverable 2 - The Query Engine

Contributors:

Alex Thiele
Mark Waters
Anirudh Sodhi
Tivaashan Varathanathan

1. What data fields will you be using from the crime data sets? Remember that some data fields might not be available in the same form or at all in every year.

The Coordinate field to search for records matching a specific Geo, Violation and Statistic.

The Value field: for multiple statistics, we would want to convert this value to the same data type so that output can be displayed with consistent formatting on a graph or other means.

2. *What years of data will your system cover? Just the years in the crime data set (1998 2015) or will you be getting data for other years from other data sources?*

Initially we will just be using the years 1998 to 2015, but if we have extra time at the end and we all agree to, we might create another year to add to the data set. However, this would require either finding additional pre-configured CSVs, or creating our own, and may not be feasible in regards to the deadline.

3. How will you be organizing your team to look at the data? Who will be responsible for various data and questions?

Our team will work together during the algorithm planning phase, so all team members have adequate input on the direction of the project. Coding will be done individually or in pairs, but the team will focus on designing the program to correctly answer one question at a time.

Once we are satisfied with our progress, we will add to the program to answer another question, and will complete as many questions as possible within the development timeframe.

4. List the types of questions that you will initially be trying to answer. Organize them into various categories based on type of answer, amount of data needed to find the answer, how the answer will best be presented, etc.

Raw values:

How many occurrences of a specific violation took place in a particular location?

-One violation in one location

-One violation in several locations

-A group/category of violations in one location.

-A group/category of violations in several locations

Categories are predetermined sets of violations coded into the program that may be used to give a broader perspective on the data. Example categories include: violent crime, fraud, drug offenses, etc.

For each of the above questions, the user will specify a single year within the allocated range for each round of output.

A bar graph would be the ideal output for this category of questions. Alternately, the user may request the numbers in text outputted to a file instead.

Crime Per Capita:

Is a particular crime more prevalent in one area, or another?

Building on the results from the previous set of questions, the user can choose to see the data in terms of how common each crime is, rather than the raw number of occurrences.

The previous data would be calculated, but then the raw number would be divided by the population from the census data, giving a relative value that can be directly compared with that of other locations.

This would be useful for comparing the rate of crime in a small town with that of a large city.

A bar graph or numbers in a text file would still be ideal for this category of output.

What is the relationship between a particular crime and household income?

Additional data from Statscan would be used to cross-reference with a particular crime and location.

Percentage Change:

Did a particular crime increase or decrease in occurrence?

The user will select one crime, one location, and a year range. Scripts from the previous questions would be reused for this mode of output, with the addition of repeating for the specified number of years and outputting the data in a line graph.

Ranking of the most dangerous cities in a specific province (or provinces in Canada)

- Results can be output in a list form.

Geographical <u>concentration</u> of a specific violation (or all violations).

- Results could be organized in a pie chart if the program is considering all types of violations.
- Results could be given in raw text output if a single violation is specified. (For example, “How many assaults occurred in this location and year? Answer: 1,000”)

5. Estimate the number of Perl scripts you will need to do the following:

Around four scripts will be used for the entire program.

a. Collect the data from the crime and census files
    - The first script should retrieve information from the user to determine what kind of data is required. The program will ask if the user wants data on one, or several locations, which locations, which specific violation, and which type of visualization they would like the data displayed in (if applicable).
    - The next scripts would collect data from the crimes CSV, and another script would collect data from the census CSV. If another dataset is required, such as income, a third script would retrieve this data. For each dataset a separate script is implemented.

b. Organize the data into an “answer”
    - Tables or numbers or answers such as “Yes”, “No”, “increasing”, etc.

        - One script for all of the mathematical calculations that would need to performed
    - Visualizations (line graphs, bar graphs, pie charts, etc.). The smaller the number of scripts the better - look for commonality in the processing and look for opportunities to aggregate the data, ie, you might create files that contain summarized/aggregated data for some fields for some time periods. For example, you might create a file that contains the murder data for a year where each line represents a geographic area.
        - One script for the generation of any type of graph or chart