Balancing Time, Money, and Safety in Chicago Transit Decisions

Will McFadden

April 22, 2015

1 Problem Description

Both native Chicagoans and visitors to the city will often find themselves traveling on public transit through neighborhoods that can make them feel unsafe. To help prevent crime and to make people feel more safe, I'm proposing to build a tool that will evaluate the probability of criminal activity based on transit route and time of day. With this information, the user will be able to decide whether to take transit or call an Uber cab. Comparing to Uber prices gives an estimate of the cost of the next most practical option to replace public transit decisions.

The data will come from 3 sources: Transit route timing data will come from Google Maps Directions API. Uber fares and travel times will come from the Uber API. The data on transit ridership and locality based crime statistics will come from the city of Chicago.

I want to integrate these three datasets to display the safety of selected CTA routes and the pricing options for an Uber ride to the same destination. I plan to build a Google Maps Embed API illustration of the examined routes as the final deliverable.

2 Plan of Attack

The project development will move in 3 sequential stages: Data Collection, Building the Crime Evaluation, and Building the Web Interface.

2.1 Data Collection

The point to point CTA and Uber data will need to be generated because an existing dataset doesn't exist to my knowledge.

The Google Maps Directions API will be queried repeatedly to automate the route generation. The API is subject to usage limits.

- 2,500 directions requests per 24 hour period.
- 2 requests per second.

I hope to build the dataset over 2 weeks. This will allow me to build a set of 35,000 point to point directions. I will focus on routes starting and ending in Hyde Park, Lincoln Park, University Village, or Logan Square and branching to random locations throughout the rest city.

The same starting and ending locations will be queried using Uber's API. Uber's API gives estimated wait time and estimated fares. Unfortunately, Uber's terms of use prohibit storing data attained through their API. Therefore, any fare and wait time comparisons will have to be carried out in real-time.

2.2 Crime Evaluation

Using crime and ridership data from data.cityofchicago, I'll need to build a metric of relative danger for different routes. I'm currently considering the metric to be simply crimes/rider on a given route. However, other metrics might be better so I'm planning on aggregating crime and route data as a first step.

The crime data has 5 million entries, each containing GPS coordinates and a description of the crime and the location. The crimes will be split into those that actually took place at CTA stations in the description, as well as crimes on the street at locations of stops.

To connect crimes and routes for the data aggregation, I'll implement a MapReduce algorithm based on the following scheme:

2.2.1 First Map Function

All location data will be binned to the nearest city block, b. From route files, all route stops are associated with the nearest block, and a key pair for each block, (b, r), is released. Likewise, from crime files, every crime, c, will be sent to its corresponding block, (b,c).

2.2.2 First Reduce Function

In the reduce function, all crimes and routes that passed through a given block are aggregated. The function then releases a key pair (r, c) for every route and crime pair.

2.2.3 Second Map Function

The Identity

2.2.4 Second Reduce Function

This will simply aggregate crimes for a given route and output all crimes to a file for that route.

2.3 Web Interface

I hope to build a web interface for the results. I normally implement my frontends on Google App Engine (because it's free). I plan to illustrate the routes using Google Maps Embedded API. The data for the crime statistics and the alternative cost of taking an Uber will appear when you click on the route.

3 Deliverables

3.1 Mid-Quarter Update

At the mid-quarter presentation I plan to have built the CTA transit dataset. I will give a summary of the main findings and the simple comparison between transit costs for different times and locations throughout the city.

I will also have a better idea about the statistical methods I will use to compare different locations for safety.

3.2 Final Presentation

At the final presentation, I will walk the class through the web interface and show a few choice examples. I'd also like to build a few summary statistics that show the general trend in safety over different routes.

4 Evaluation

The crime prediction model will be evaluated against a subset of the data withheld from the model construction. The Web App can be evaluated based on its practicality.