COMPSCI 597: Big Data and NOSQL Databases

Project Proposal

Temporal and Spatial Characterization of Service Requests from Historical Data in Washington DC

Tanner, Reuben and Trisca, Gabriel

October 10, 2013

1 Identify the dataset that you wish to analyse

We will be working with the 311 Service Request dataset from Washington DC. This dataset contains locations and attributes of service requests received by the Office of Customer Service Operations of the Office of Unified Communications (OUC) through the Mayor's Call Center (311)[1].

The Service Request dataset was selected because it is already geocoded (avoiding the expensive reverse-geocoding operation on thousands of addresses), and also because the different service requests have unique identifiers that will make any clustering operations much easier to perform.

2 Mention the type of backend system that you would use

- The dataset will be loaded into HDSF
- Hadoop will be used for processing and grouping the different service requests
- Mahout will be used to cluster data
- Google Maps to visualize the data

3 Articulate the type of analysis you would conduct

The primary objective behind collecting this data was to provide timely and accurate information for decision-making[1]. Our intention is to analyse it further to achieve the following:

1. Create a map that characterizes and predicts the service requests on different areas of the city
2. Identify the geographical areas where the city is spending more resources to maintain
3. Infer correlation between different service requests based on proximity or temporal correlation.

4 Provide Metrics for your systems evaluation

- The system can provide an overview of areas where certain service requests where reported on a temporal basis.
- The system reports the areas with the most unique service requests

- The system identifies service requests that are most likely correlated due to geographical proximity or other reasons.
- The system can incorporate new data and generate visualizations on arbitrary datasets that conform to a pre-established schema.

5 Outline a tentative timeline and projected number of hours

- Randomly sample the data to understand its geographical distribution and the relative frequency of service requests **(10 hours)**
- Investigate applicable clustering/regression algorithms on top of Mahout **(15 hours)**
- Implementation of ML algorithms to produce the results outlined on #3 **(20 hours)**
- Visualization of produced data **(15 hours)**

References

[1] 311 Service Requests *Office of Unified Communications (OUC)*. http://data.dc.gov/Metadata.aspx?id=4 Accessed on October 9^{th} 2013