**Project Phase 3: GISCUP 2016 Challenge**

CSE 512, Spring 2017

By

Namrata Nayak (1212408767)

Srinivas

Varun

Venkat

**Problem Definition:**

Apply spatial statistics to spatio-temporal big data in order to identify statistically significant spatial hot spots using Apache Spark.

**Input**:

A collection of New York City Yellow Cab taxi trip records spanning January 2009 to June 2015. The source data may be clipped to an envelope encompassing the five New York City boroughs in order to remove some of the noisy error data (e.g., latitude 40.5N – 40.9N, longitude 73.7W – 74.25W).

**Output**:

A list of the fifty most significant hot spot cells in time and space as identified using the Getis-Ord statistic.

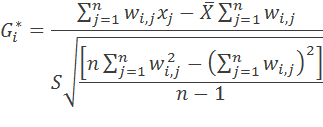
**Algorithm:**

We have two classes where; the class MathUtility.java deals with all the mathematical calculations for calculating Getis-Ord statistic and the other class HotSpots.java handles the steps required to calculate Getis-Ord statistics and list out top fifty hot spot cells.

We have taken a few constants like the minimum and maximum latitude and longitude as per the problem statement. Each point is of 0.01 units. The total number of cells is calculated as the multiple of number of (days\*latitude\_range\*longitude\_range). Also, the program takes the input file and the output file as the parameters.

The method ‘findtop50hotspots’ takes the input file and uses Map-Reduce on ech line to form the JavaPairRDD object which represents a point in the 3-dimensional cube. This cube is calculated based on the latitude and longitude average.**(explain more)**

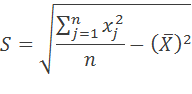
As per the formula for calculating the Getis-Ord statistics, we calculate the numerator and the denominator and then divide them to get the z-score:



(**explain)**

The covariance (S) is calculated by:

* Calculate the total cost: Summation of all the rides
* Calculate the average cost: total cost/total number of cells
* Calculate the covariance as per the formula:



Where,

S: covariance

Summation of xj2:  is sum squares of all the points in the space-time cube

N: total number of cells

X(bar): average cost

Once the z-score is calculated for each point, we create a priority queue and compare the**\_\_\_\_\_\_\_** for each point and store the timestamp, latitude, longitude and their z-score of the top 50 most significant cells in the output file.