CSE 512 Spring 2017
Project Phase 3 Requirement

Due Date: 17 April 2017

Description:

This phase will focus on applying spatial statistics to spatio-temporal big data in order to identify statistically significant spatial hot spots using Apache Spark. The topic of this phase is from ACM SIGSPATIAL GISCUP 2016.
The Problem Definition page is here:
http://sigspatial2016.sigspatial.org/giscup2016/problem
The Submit Format page is here:
http://sigspatial2016.sigspatial.org/giscup2016/submit

Special Requirement (Different from the GISCUP):

As stated in the Problem Definition page, in this phase, you are asked to implement a Spark program to calculate the Getis-Ord statistic of NYC Taxi Trip datasets.
To reduce the computation power need:
1. You only need to take the data of 2015 January Yellow Taxi (~1.8 GB in total).
2. Each cell unit size is 0.01 * 0.01 in terms of latitude and longitude degrees.
3. You should use 1 day as the Time Step Size.
4. You only need to consider Pick-up Location. You can ignore the Drop-off Location.
5. The code should be written in Scala or Java and be able to generate a executable Jar package.
6. You may use GeoSpark Maven Coordinates to call GeoSpark API and help your coding. You even can use GeoSpark source code as a starting point and do some changes on it.

Submission format (Different from the GISCUP):

1. A single folder that contains the original source code. Its name convention is like this “group1_phase3_source”. Please include a “README” file for any special instructions on how to compile the submitted code. Submission of the source code is mandatory to ensure originality of the submitted work.
2. A binary Jar package compiled from your source code. Its name convention is like this “group1_phase3.jar”. The submitted jar will be invoked using the following syntax: ./bin/spark-submit [spark properties] --class [submission class] [submission jar] [path to input] [path to output] http://spark.apache.org/docs/latest/submitting-applications.html
3. A CSV file contains top 50 hot spots (cells). Its name convention is like this “group1_phase3_result.csv”. You need to sort the top 50 hot spots in the descending order of z score. Its format should be as follows:

cell_x, cell_y, time_step, zscore

4. A PDF report that describes your algorithm in any document format. Its name convention is like this “group1_phase3_report.pdf”.

All the four items should be put into a folder such as “group1_phase3” and compressed into a single ZIP file. This ZIP file should be submitted to Blackboard.

Evaluation formula (Different from the GISCUP):

We use the Jaccard similarity coefficient of the reference result from TA (R) and the candidate result from your group (C). Your group score of Phase 3 will be:

*Phase 3 score = 100 * Jaccard(R, C)*

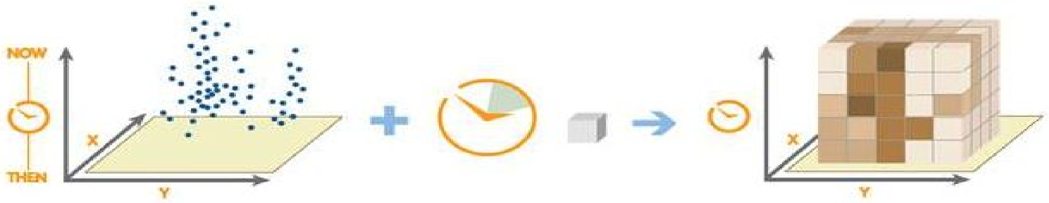
R and C contain the top 50 hot spots (cell IDs). Compilation errors and other errors may also lead to point deduction.

The following example shows the definition of Jaccard similarity coefficient: Consider two sets A = {0, 1, 2, 5, 6} and B = {0, 2, 3, 5, 7, 9}. How similar are A and B?

Jaccard(A, B) = A ∩ B / A ∪ B = {0, 2, 5} / {0, 1, 2, 3, 5, 6, 7, 9} = 3/ 8 = 0.375

Hints:

1. A higher Getis-Ord z score means this cell is hotter.
2. If a cell has more taxi trip records, that means this cell is hotter and has higher z score. The number of trips in this cell is the attribute value of this cell.
3. This spatial neighborhood is created for the preceding, current, and following time periods (i.e., each cell has 26 neighbors). For simplicity of computation, the weight of each neighbor cell is presumed to be equal. You can treat it as 1. If two cells are not neighbors, their weight is 0.
4. You can use GeoSpark 0.6 or later to do spatial aggregation and further generate this cube (Photo from GISCUP website).


NOW
THEN
X
Y
X
Y