Portfolio Report

Gunda Sai Ganesh Reddy   
School of Computing and AugmentedIntelligence(SCAI)  
Arizona State UniversityTempe, USA  
sgunda3@asu.edu

# Introduction

The amount of data created in the modern world is enormous, thanks to increased internet speed and technological breakthroughs. And observational data may be found virtually anyplace. These datapoints can be used by enterprises by obtaining them from user mobile assets. The collected data is highly important in terms of real-time analytics. To analyze this massive amount of data efficiently, we require specialized algorithms and frameworks. Organizations may utilize data point analysis to their advantage and increase income. To address these challenges, we chose two problems to solve utilizing data processing techniques.

The first problem is to discover all the restaurants in a particular place, and once that location is provided, we need to display the restaurants that are within a reasonable range of the present location. This may be utilized by restaurants or businesses to assist their consumers in getting to their location quickly and easily. It also assists businesses in growing their company by giving relevant eateries that belong to the same category and are within the user's reach. The second issue we tackled was performing a geographical hot spot analysis on temporal data utilizing distributed processing techniques in the Apache spark framework, followed by computing the distance measure to determine the commonalities of hot zone spots. The estimated G score assists us in determining the distance between the cell locations. This data is critical in lowering air pollution since it informs the authorities about which locations are most popular at what times. It also helps us find a way through to firms so that they can compete with their competitors by making travel easier for riders and supplying more taxis at busy hours. These two problems can take a quite an ample of time to do the analysis if gone through conventional way of Data mining. So, data processing played a important role and reduced the load time for doing the analysis.

# Description of Solution

To gain a thorough grasp, we will go over each of the subproblems addressed in this project step by step. The first challenge includes two tasks to complete: (I) find businesses by city and (II) find businesses by location. The second problem similarly contains two tasks that we must do. Hot zone analysis (I) and hot cell analysis (II). Going through each of these functions step by step.

## **Find Businesses By City**

The function accepts three arguments: the city to be searched for, the output location file, and the collection to be used. I went through the restaurant object collection and changed everything to dictionary format. Then, will iterate through the dictionary and search for the collection type of restaurant. So, the resulting dictionary will contain just the restaurants that we are looking for. The findings are saved in the output file, together with all pertinent information such as the restaurant's name, entire address, city, and state.

## **Find Businesses Based On Location**

The function accepts five arguments: the categories to search for, the location, the maximum search distance range, the location of the output file, and the collection to search for. Then, we will loop through the resulting dictionary, gathering just the eateries that meet our description. Now we need to determine if that eatery is within driving distance of our present location. I used the Haversine formula to determine the distance between two eateries. Which uses cosine and sine angles to calculate the distance between the latitude and longitude of the current location and the restaurant location. The resulting distance will be in radians, which will be converted to miles later. If the distance is within the range, it is then tested. If so, the restaurant’s information is copied to the output file.

Graphical user interface, text, application

Description automatically generated

Figure 1. Compute Distance Function

**Note**: angle needs to be in radians

For the second subproblem, we will do a geographical hot spot analysis using taxi trip data from New York. It is real-time data collected between 2009 and 2012. We're talking about two ways here. We'll go over each of them step by step.

## **Hot Zone Analysis Approach**

We are given two data sets, one point and one rectangle. We must first compute the hotness of the dataset's rectangle. The point data from the New York taxi trip data that identifies the pickup and zone data are used as input for this hot zone study. The result is a rectangular string in ascending order of hotness. The number of dots represents how heated the rectangle is.

**Steps**:

* Load the rectangle and point dataset
* Combine the datasets with the condition to determine if the point in the point dataset exists inside the rectangle bounds.
* Choose the rectangular strings in ascending order of hotness.
* Store the results in the excel sheet.

## **Hot Cell Analysis Approach**

We must now apply spatial statistics to temporal data from the New York taxi dataset. I utilized Apache Spark to identify major hotspots. Then I wrote a Spark Application to compute the Getis-Ord for the Dataset. Before proceeding with the analysis, I assumed a few restrictions. Like

* Input temporal Data will be in Monthly format
* Each cell of the point dataset is of 0.01 in terms of latitude and longitude.
* The step size is assumed to be 1 day
* We have used only the pickup location

Now we will go through analysis step by step.

1. Load the data set from the source
2. Determine the cell value: We will next give cell coordinates to the pick points, with x and y coordinates indicating latitude and longitude. Whereas the Z coordinate represents the time of pickup.
3. Next, we define the maximum and minimum for each of the x, y, z coordinates
4. We will now run a series of range queries to get the mean, standard deviation, and G-score of the cells.
   1. Selecting only the cells which are within range
   2. Removing duplicates like x, y, z is same as y, x, z and z, y, x
   3. Finding the cumulative sum of cells
   4. Finding the mean and standard deviation of the cells
   5. Calculating the G-score of the cells

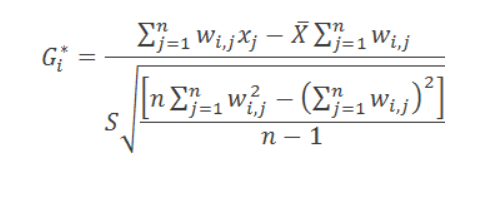


Figure 2 Getis-Ord statistic

Where

Xj = attribute value for cell j

Wi,j = Spatial weight between i and j

n = total number of cells

Schematic

Description automatically generated

Figure 3 Mean and Standard Deviation

The monthly point data from the NYC taxi trip dataset is used as input for the hot spot study. The result is the coordinates of the top 50 hot cells, ordered descending.

# Results

We will now examine the outputs produced by each of the previously stated challenges. When the Find Business Based on City function with the arguments Tempe, Output.txt, and data is called. It displays the three eateries in Tempe City.

Text

Description automatically generated

Figure 4 Find Business On City Function Output

It shows the list of restaurants in Tempe city. I have designed the test cases to check the reliability of the function. The following three test cases are primarily focused on:

* Does the function accept three arguments
* Is the function writing the data to correct location
* Is the function returning the correct number of results

The outcome is accurate if all three requirements are met! I tried the function using the provided and additional test cases, and I received a perfect score. This can be used by the search engines to display the restaurants in a location requested by the user. For Find Business Based on Location function is passed with the option ['Buffets'], [33.3482589, -111.9088346], 10, 'output loc.txt', data is supplied, the restaurant that corresponds to the buffet category and is less than 10 miles away from the present location is returned.

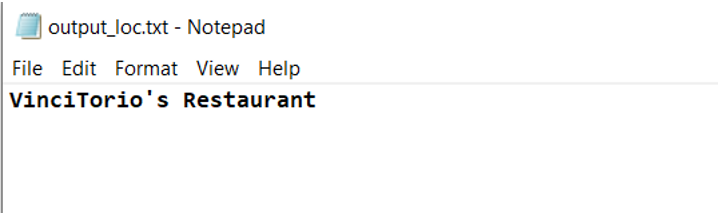


Figure 5 Find Business Based On Location output

Also, for the second sub problem I have designed the test cases to check the reliability of the function. The following three test cases are primarily focused on:

* Is the function accepting five parameters
* Is the function writing data to correct location
* Is the function returning right number of results

The outcome is accurate if all three requirements are met! I tried the function using the provided and additional test cases, and I received a perfect score. This function has significant business ramifications. This function, where we are offered recommendations of the desired category restaurants within our range, may be seen even in our daily usage. It can assist businesses in growing by acquiring new customers.

Let us now discuss the implications of resolving the problem addressed in the second project. The hot zone analysis provides information on hot zones in a geographic place. To show this, we used the New York trip dataset, on which Hot Zone analysis yielded a rectangle string of hotness order in decreasing. The amount of data points within the rectangle's bounds is used to calculate the level of hotness.

A picture containing table

Description automatically generated

Figure 6 Hot Zone Analysis Output

The output shows the count of the rectangles- number of datapoints in the rectangular region. This can be used in the transportation sector to get an understanding of user business with their company and weather organization to understand the hot zones of the country. The spatial analysis of the temporal data from the New York trip dataset yields the second subproblem of this study. The dataset of NYC taxi trips' Getis-ord is calculated using the spark application that was constructed. The top 50 hot cells' coordinates are displayed for the monthly point data of New York City taxi trip data input, in descending order.

Calendar

Description automatically generated with low confidence

Figure 7 Hot Cell Analysis Output

The output is the coordinates of the top 50 hot cells sorted in a descending fashion. This is an extension of the previous problem where we are just getting the generic overview of the system on which part has the hot zone. To get an accurate understanding of the latitude and longitude of the hot zone of the place we can employ this function. Since it lets the government know, which areas are busiest when, this information is vital to reducing air pollution. By facilitating travel for riders and providing more cabs during peak hours, it also assists us in getting through to businesses so they can compete with their rivals. In this way Data processing at scale techniques can help the business grow their revenue and increase customer satisfaction.

# Lessons learned

These projects broadened my awareness of the value of big data in today's environment and how businesses may utilize these algorithms to enhance income. For the first subproblem, discovering restaurants, I utilized a NoSQL database. It helped me understand how simple they are to utilize without a schema and increased the project's viability and convenience of usage. UnQLite, a JSON-based document storage system, is utilized for this project. It's a little overhead, but it's quick and really beneficial. Because of UnQLite, it is now easier to understand and debug the database. This project provided me with a practical grasp of how well-known websites and applications use the user's current location to list local eateries. Understanding the spherical law cosines has been greatly aided by learning about the methods used to compute the geographical distance between two sites on Earth. Using latitudes and longitudes, I calculated the shortest path between two geometrical sites on Earth using the haversine formula. This formula has many practical advantages, but the formula's single drawback is that it assumes the earth to be a perfect spherical.

While going through the second project I used the Apache Spark and Hadoop to do the handle the large datasets. The distributed framework of Apache Spark has been quite helpful. I have deployed the Spark application using Java and generated the executable Jar file which now can be used in any system to perform the analysis. Initially I thought of using the map reduce but after knowing the advantages of Apache Spark over the map reduce, I was prompted with Spark framework. The speed of execution by Spark is unparallel it performs analysis hundred times faster than Map reduce. The ease of use of spark is also one of the quite advantages factors with its developer friendly API which can be directly used. Never had any real time understanding of temporal data. This project helped understand how companies use these kinds of data to grow their business. It made me realize the importance of every data point that’s been collected from mobile assets by organizations. The hot cell analysis approach where we do the spatial statistics on the temporal data, I was faced with a challenge on how to group the point dataset of rectangle, but I did find a approach of using the Getis ord statistic. It did help me get the z-score and p -values which later tell me whether the cluster is high or low spatially.

# References

1. Nikitopoulos, P., Paraskevopoulos, A.I., Doulkeridis, C., Pelekis, N. and Theodoridis, Y., 2016. BigCAB: Distributed hot-spot analysis over big spatio-temporal data using apache spark (GIS Cup). In Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographical Information Systems.
2. http://sigspatial2016.sigspatial.org/giscup2016/problem
3. http://www.movable-type.co.uk/scripts/latlong.html