CS236 Spring 2018: Project Description

Andres Calderon

May 26, 2018

1 Introduction

The main goal of this project is to understand and implement spatial queries using Simba[1]: an Apache Spark framework extended with spatial capabilities.

2 Description of datasets

The first dataset (POIs.tar.gz) are Point-of-Interest (POI) locations around Beijing. The dataset was extracted from OpenStreetMap[2] and it collects diverse amenities and shops around the city. It shows the identification, longitude, latitude and a brief description of each POI. The second dataset (trajectories.tar.gz) are trajectories of moving objects around Beijing. It was extracted from the GeoLife[3] Project. It collects the trajectory identification, object identification, longitude, latitude and time of each measurement.

3 What you need to implement in this project

The project has three parts. First, you will get familiar with Simba and its R-tree implementation. You will use a sample of the trajectories dataset (around 10% of the data) to create an R-tree index. Then you should be able to extract the bounding box of the points at each partition which represent its MBR and plot the associate rectangle (you will find the `mapPartitions(`*func*`)` function quite useful). You can see an example here. You should play with different values of partitions (by setting the `simba.index.partitions` parameter) before to choose the one you want to plot.

You can use any tool to plot the rectangles and points. There is no need to use any GIS software or web mapping libraries but that would be a plus!. If you are planning to

[1] http://www.cs.utah.edu/~dongx/simba/

[2] https://www.openstreetmap.org/

[3] https://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/

use any of these option please note that the projected coordinate system used for these datasets is the New Beijing / 3-degree Gauss-Kruger CM 126E (EPSG code: 4799, more details here).

For the second part, you will answer the following spatial queries:

1. Retrieve all the restaurants located inside the 5^{th} road ring of the city. Have a look at this article to get an idea about the transport road network of Beijing. We will simplify the query using a rectangle with coordinates (-339220, 4478070) and (-309375, 4444725).

2. The Tiananmen Square is an interesting place to visit in Beijing. The Beijing Tourist Agency would like to know how many people (average per hour) are around 2 Km of it. The coordinates of the Tiananmen Square are (-322357, 4463408). Focus your query just on workdays (Monday to Friday).

3. Using the same area as query 1, let's divide it in for quadrants as it is shown in figure 1. Use the midpoint to find the limits of each quadrant. How many trajectories start in a quadrant but end in a different one? How many trajectories start and end in the same quadrant?

4. Let's select a sample in the trajectories dataset for all the points happening between February and June. For each point in your sample count how many other points are in a radius of 100 meters. Show in a table the location of the top-20 points with more points around them.

5. During workdays, which are the top-10 more popular POIs in 2008 and 2009 individually. Count as popular those POIs with objects around 100 meters. The more objects around, more popular the place. Show the results in tables aggregating the results by month.

For the third part, you will take the last two queries (4 and 5) and measure the Scaleup (execution time using increasing number of cores) and Speedup (execution time using N cores over execution time using 1 core). For 4, you should use different values for the radius of search (from 100 to 500 meters). For 5, you should change the distance to each POI (from 100 to 500 meters). Find the more popular POIs for 2008 and 2009 together, do not worry about any aggregation.

4 What you should submit

You must write a report with your answers, findings and/or problems. Along with your report (in **PDF** format, please). You also have to submit your (**well-documented**) source code, along with a **README** file that explains how to reproduce your findings.

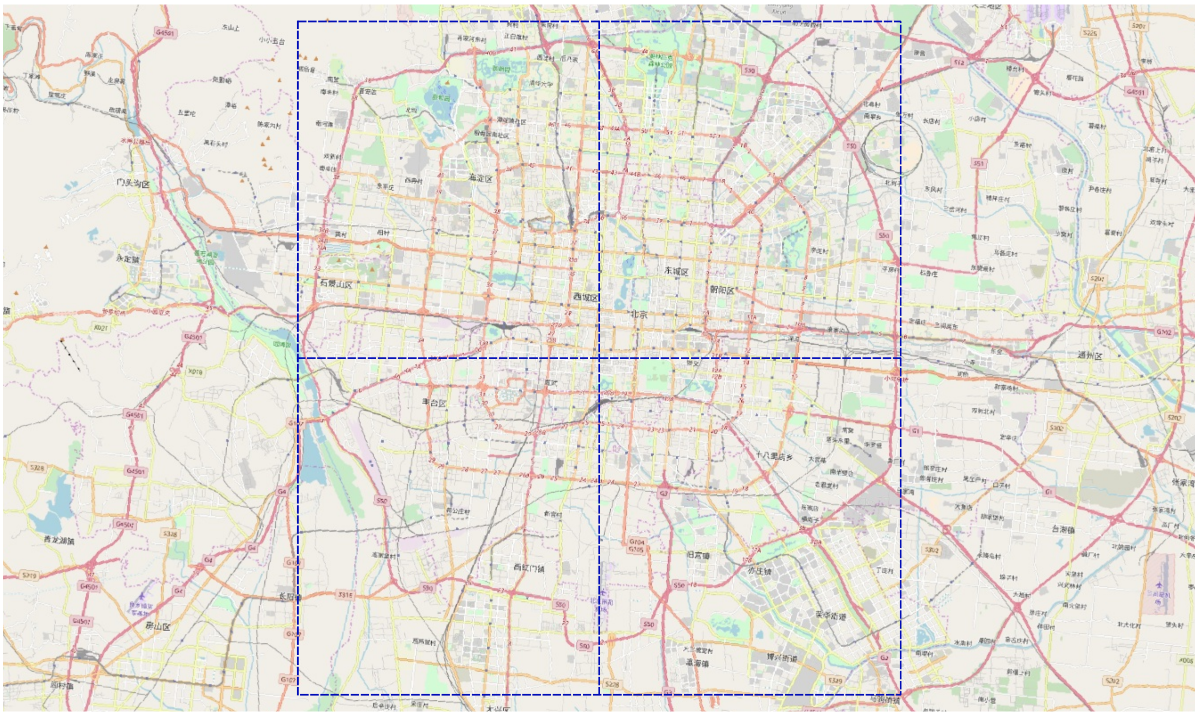

Figure 1: Quadrants for trajectories dataset.

5 Deadline for the project

The deadline for this project is the last week of instruction. The project is to be done in groups of **two** students. Please submit a tar.gz file named “username1_username2” (username=your NetID) to acald013@ucr.edu with the subject “[**CS236 Final Project**]”.