Kapil Bastola

Jan 29, 2017

DS 670

Data Set Loading and Data Summary

**Data Set Loading**

The dataset I was assigned for the capstone class is the Road Traffic data from CityPulse dataset collection. This dataset consists of data from traffic in a city called Aarhus in Denmark. This dataset is collection of traffic data between two points for certain duration of time. There are four different datasets available in the CityPulse website for different durations. The first dataset is from February 2014 to June 2014, second dataset is from August 2014 to September 2014, the third dataset is from October 2014 to November 2014 and the fourth dataset is from July 2015 to October 2015. The datasets are available in two formats. One is a zipped file with collection of csv files while the other format is semantically annotated format. For the purposes of this paper I am using the zipped dataset with raw .csv format files. CityPulse website also has a .csv metadata file which provides additional information regarding where the data was collected from.

First I downloaded the first dataset for February 2014 to June 2014 and unzipped the files. This created a folder called traffic\_feb\_june with 449 csv files in my Downloads directory. I then used zeppelin to read the dataset. Using Spark, I combined all the files in the directory mentioned above into one table called “roadtraffic.” Then I registered the table as an SQL table so that I could run some queries to see what the dataset looks like and obtain data summary.

**Data Summary**

The first query I ran was to see 15 rows of datasets. This showed me what different columns are in the dataset and what the format for data in each column is. The dataset has 9 columns. These are status, avgMeasuredTime, avgSpeed, extID, medianMeasuredTime, TIMESTAMP, vehicleCount, \_id, and REPORT\_ID. I will expand on each of these columns below. Next I wanted to see how many rows are in the table. There are more than 13.5 million rows in the table. In order to understand some of these columns further I ran some more queries.

The **status** column has “OK” in all of the first 15 lines that I pulled. I wanted to see if there are any other values for this field. I looked at all the distinct values in the status column and saw that “OK” is the only value in the whole column.

**avgMeasuredTime** column has numerical values. The two sensors on two points of the road measure how long it took a vehicle in seconds to reach the second point from the first point. This field gives us the mean of total time taken in seconds by different vehicles to reach from the first point to the second point for each reading.

**avgSpeed** column also consists of numerical values. This column provides the average speed of vehicles between the two points in kilometer per hour (kmh).

**extID** column consists of 3-4 digit numerical values. Initially I was not certain what this column represented so I looked at all the distinct values in the column. I saw that they are sequential numbers and there are total of 449 distinct values in the table. Since we had 449 total files in our dataset, extID is a unique identifier for each file.

**medianMeasuredTime** column also consists of numerical values and has similar values as the avgMeasuredTime in the first 15 data points we looked at. This column gives us the median of total time taken in seconds by different vehicles to travel between the first and the second point for each reading.

**TIMESTAMP** column consists of date and time values and gives us the date and time of each reading.

**vehicleCount** column consists of numerical values. For each reading there are multiple vehicles passing between the two points. This column gives us the number of vehicles that travel between the two points during the readings.

**\_id** column consists of numerical values as well. In the first 15 rows, there are 6 digit numerical values that are all different. I wanted to count the distinct number of values in the \_id column and saw that the number of distinct values in this column is equal to the count of rows in the table. This suggests that \_id is the unique identifier for each row of data.

**REPORT\_ID** column consists of numerical values as well. In the first 15 rows, all the values were same in this column. Hence, I looked at the count of distinct values and found that there are 449 total unique values in the column. This suggests that this is an identifier for each file as well. In the metadata provided in the website I see that there is a column with same name and the values as the report\_id in the dataset. Hence, report\_id can be used to join the data set with the metadata file to obtain more information on each of the reading that took place.

The metadata file is a single .csv file with more information on the data streams. It has 449 rows of data implying that each row corresponds to each file in the dataset. The metadata file has information on where exactly the two points were. It contains information like street, city, latitude, longitude, postal code, and country for the two points. Apart from this it also contains ext\_id and REPORTID. These columns were present in the data set as well. Upon further look, ext\_id in the metadata file did not match the ext\_id present in the datasets but the REPORTID were same in the both file. So we can use the REPORTID in the metadata file with the REPORTID in the dataset to join the two tables if the need arises.

I did the initial analysis and data evaluation using only the first dataset in the website which is for Feb 2014 – June 2014. The website has a link to download individual files as well and when I looked at an individual file for other durations, the format of the data is similar to the format in the data files of the data set I downloaded.

**Reference:**

1. Dataset Collection. (n.d.). Retrieved January 29, 2017, from http://iot.ee.surrey.ac.uk:8080/datasets.html