CS510 Data Science Exploration, Midpoint Report
Name: Qiushi Teng

Project Objective

In this project, I plan to compare how traffic has changed in the last four years using data from Portal. The system I plan to use is Apache Spark.

Project Approach

- System Setup

  I installed the VirtualBox software on my laptop, and inside VirtualBox, Ubuntu system is installed to run the Spark system. I now have the Spark system set up and run locally, and am currently learning how to perform different queries on the system on a small sample dataset.

  Right now I am able to read sample json and csv files, and use both Spark df command and Spark SQL (a temporary table or view needs to be created) to perform some basic queries , including displaying the table content, how to show particular columns or rows that meets some particular conditions, groupby and other aggregate operations. There are some other things I will continue to explore, such as how to deal with missing data.

- Portal Data

  Datasets that I plan to download for my project:

  - Data from 2017, 2018, 2019, and 2020 for yearly comparison
  - All months from 2019 for monthly comparison
  - Data from one week for daily comparison
  - Data for weekdays and weekend comparison

  for all of which with all week days included and resolution as 1 hour.

Team Structure

There is no update to team structure. I will continue working on my own.

Progress Made So Far

There are five milestones I listed in my project plan,

1.	Get the Spark system set up and ready for this project	Done
2.	Learn how to perform different kinds of queries using Spark	Done
3.	Gather data needed for this project	Working on
4.	Perform different queries for further analysis	Start Soon
5.	Analyze and compare how traffic has changed	Start Soon

The first milestone has been completed, and the second one is mostly done (I have tried most of the commonly seen query operations, just in case of some complex queries that need to be done.). I am now working on collecting data and will soon start to query on these datasets. Following that, I will analyze and compare the result from the queries.

Below is the list of queries and analysis that I think of,

- Compare average speed and volume from 2017 to 2020, also include total volume of each year for comparison
  (Updated: Above is my original plan. However, the dataset including one year data is too large to be downloaded, and the largest dataset that I am able to download is for one month. Therefore, I am going to pick a particular month from these four years to compare average speed, volume and total volume.)
- Pick all months from a particular year, compare how speed and volume change among different months
- Pick a particular week, compare how speed and volume change by weekdays, from Monday to Sunday
- Pick a particular week, compare speed and volume change hourly in a certain time period, such as from 8 am to 8pm, visualize hourly change and peak time for different days
- Compare how and if traffic has changed on weekdays and weekends for the last four years
- Pick month(s) that is(are) suitable for representing rainy seasons, and month(s) that works best for non-rainy seasons, compare if traffic is affected by the weather

These are the different queries and analysis I came up with now. I will start with the first one that compares how traffic has changed from 2017 to 2020. It is very likely that I will also perform queries that do a month-wise and week-wise comparison. Other queries also look interesting to me, but I will first get started and probably make some adjustments along working on this project, and may need to make some changes to some queries that I have right now.

Midpoint Meeting Time

July 21st, 1:20 PM