COMP 598 Homework 4 – Bokeh Dashboard

30 pts
Assigned Oct 1, 2020
Due Oct 9, 2020 @ 11:59 PM

This is an INDIVIDUAL Assignment – each student's work must be their own, each student completes this assignment, there are no teams for homework 4.

In this homework, you are a data scientist working with the New York City data division. Your task is to develop a dashboard allowing city leaders to explore the discrepancy in service across zipcodes. You'll be using a derivate of this dataset:

- https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9

The original dataset is nearly 10 GB in size, so we're providing a trimmed down version… it's only 3.6 GB (when uncompressed)!! You can download it from the MyCourses. Trim down this dataset to only include the incidents that occurred in 2020 (for an added challenge, see if you can trim the dataset down using exactly one call to the grep command line tool).

For the remainder of this assignment, you should only work with the trimmed down dataset.

Task 1: Get Jupyter running on your EC2 (5 pts)

Setup Jupyter on your EC2 with password login support. Create a Jupyter notebook in which you have loaded the data file and printed out the number of distinct zipcodes in the dataset. You will submit the logs that Jupyter created when it booted up and handled you logging in via a browser. Note that an automated grading script will be used to evaluate your logs, so they must contain the log text produced when Jupyter successfully boots up AND when you log into the server.

output the logs you get after typing "jupyter notebook" in command line to jupyter.log, including logs of 1. boot 2. you logging in

- To capture the log data, redirect the stdout and stderr from your Jupyter notebook command into the file jupyter.log

Task 2: Bokeh dashboard (25 pts)

The goal of your dashboard is to allow a city official to evaluate the difference in response time to complaints filed through the 311 service by zipcode. Response time is measured as the amount of time from incident creation to incident closed. Build a bokeh dashboard which provides in a single column, the following:

Pre-compute stats such as avg for each month for each zipcode to make the response time less than 5s

- A drop down for selecting zipcode 1
- A drop down for selecting zipcode 2
- A line plot of monthly average incident create-to-closed time (in hours)
    - Don't include incidents that are not yet closed
    - The plot contains three curves:
        - For ALL 2020 data
        - For 2020 data in zipcode 1
        - For 2020 data in zipcode 2
        - A legend naming the three curves
        - Appropriate x and y axis labels

When either of the zipcode dropdowns are changed, the plot should update as appropriate.

Other details:

- Your dashboard should be running on port 8080. The dashboard name (in the route) should be "nyc_dash".

- The bokeh dashboard should authenticate any user who logs in with URL params username = “nyc” and password = “iheartnyc” (quotes not included). Failed authentications just need to fail to allow the user in (i.e., they don’t need to route the user to a login page that actually exists).
- On any change to either zipcode, your dashboard must update within 5 seconds.
- A design tip: there are WAY too many incidents in 2020 for you to be able to load and process quickly (at least quickly enough for your dashboard to meet the 5 second rule). The way to solve this is to pre-process your data (in another script) so that your dashboard code is just loading the monthly response-time averages for each zipcode … not trying to compute the response-time averages when the dashboard updates.

Submission Instructions

Your MyCourses submission must be a single zip file entiled HW4_<studentid>.zip. It should contain the following items:

- ip_address.txt – one line containing only the IP address of your EC2. This will be used by the TAs to visit your bokeh dashboard at http://<your-ip-address:8080/nyc_dash. Note: to receive credit, you must leave your dashboard running from Oct 13 @ 9 AM EST to Oct 15 @ 10 PM EST. During that time, the TAs will visit each dashboard and grade them.
- jupyter.log – the log capture described above for task 1.