CS145: Data Management and Data Systems

Stanford University, Fall 2021

Project 2: Visualizing Data
15% of Course Grade / 50 points

Due Date: Monday, November 1st, 11:59PM

Overview

In this project you will explore a public GitHub dataset using Colaboratory. Colaboratory is like a Jupyter notebook, but it has collaboration and integrations with BigQuery built into it. We will be using a subset of the BigQuery public dataset due to its size; the subset can be found here. We highly recommend pinning it to your sidebar for easy access.

You will be exploring the dataset in the provided notebook `project2.ipynb`. You can access the notebook from the course website, **make a copy of it in your own drive**, and begin the assignment.

Note that this part is intended to help prepare you for the last course project, where you will be running through an entire data cycle of querying, visualizing, and predicting on a dataset of your choosing. This is an individual project, but getting used to using Colab will aid you in Project 3 when you can work in pairs.

Get Setup With Colaboratory

Here is an overview of Colaboratory features and a brief guide for using BigQuery through Colaboratory. Before proceeding, make sure you have read and understood these support documents. To open a new notebook in Colab, you can go to *File > Upload notebook* and choose the file either from your computer or from Google Drive. You can also make a copy of an existing Colab noteboook by going to *File > Save a Copy in Drive ...* . Colab notebooks can be saved just like any other file to your own Google Drive account.

Note: You have to be careful with your BigQuery credits when running cells on your Colaboratory notebook, as Colab will not tell you how much data your query will use. We advise you to check your queries in the BigQuery interface first to see how much data it will consume; keep in mind that a query using about 5GB will cost ~ 2.5 cents.

Section 1: Understanding the GitHub Dataset (4 points)

To begin your exploration of the GitHub dataset, you will briefly investigate the various tables in the dataset and answer a few short questions. This will help you understand what tables to use for your visualizations in Section 2.

Section 2: Investigating Query Performance (8 points)

In this task, you will look at some queries that are pretty inefficient in terms of how many bytes need to be processed. You will think about what makes queries inefficient and how to make them more efficient. You'll also get a chance to look at how the query is optimized in BigQuery.

Section 3: Visualizing GitHub Data (38 points)

This task will be the bulk of Project 2. You will learn to create visualizations to help you understand and answer questions about the GitHub data. For this assignment, you will have to think about what data you should use to answer a question and what kind of visualization you should make to clearly convey your information. You will then analyze your visualizations to understand how certain features of a GitHub repository correlate to its popularity.

This part of the project will be valuable practice for Project 3.

Honor Code

This assignment is to be done individually. We encourage students to form study groups to complete the assignment, but the solutions to each assignment must be written independently. Be sure to list your collaborators on each part of the assignment at the top of the corresponding Colaboratory notebook.

We take the Honor Code seriously. Working in groups to discuss class concepts or a specific problem at a high level is OK, but the following would be considered honor code violations:

- Looking at the writeup or code of another student.
- Showing your writeup or code to another student.
- Discussing a problem in such detail that your solution is almost identical to another student's solution.
- Uploading your writeup or code to a public repository (where other students may be able to find it).

Submission Instructions

Once you have filled out the Colab notebook completely, you are ready to submit. If you collaborated with others to discuss findings at a high level or generate queries, make sure to add their names and SUNet IDs to the cell at the top of the Colab notebooks.

In total, you will be submitting two files, each to a separate Gradescope assignment.

To submit:

1. Download the Colab notebook as an iPython notebook - you can do this by going to *File > Download .ipynb*.
2. Create a PDF of your Colab notebook, **making sure that you have run all cells first.** Make sure you've closed the table of contents sidebar before you create the PDF so we can easily see your work and output.
    - In Google Chrome, you can do this by going to *File > Print* and then choosing *"Save to PDF"*. If your graphs are too large to fit in the PDF, you can try going to *More Settings*, setting "Scale" to "Custom", and choosing a smaller scale (try between 40 and 50).
3. Submit the PDF file to the **Project 2 - PDF** assignment on Gradescope.
4. Submit the iPython notebook to the **Project 2 - iPython** assignment on Gradescope.

Note: We reserve the right to deduct points from your project if you do not follow the submission instructions, if there are some cells which have not been run or with non-readable output, or if you have assigned your pages to the questions on Gradescope incorrectly. **Please read through your PDF document before you submit it and *ensure that all answers are clearly visible.*** Please also leave yourself enough time to do the assignment/submission, and go over your assignment in Gradescope to make sure it is correct!

You may resubmit as many times as you like; however, only the latest submission and timestamp will be saved, and we will use your latest submission for grading your work and determining any late penalties that may apply. Submissions via email will not be accepted.