DSCI 510: Principles of Programming for Data Science

Fall 2022

Homework 3

Data Collection

Due: 12/2/2022 23:59 PM PT

Task:

- Please submit a zip package of:
  - A PDF, Markdown, or Jupyter notebook file answering the following 1 question. 5 points.
  - Data collection code. 5 points.
  - A small sample of the collected data. (No more than 10 MB). 5 points.

1. What data did you collect? How many different sources did you use? How many data samples did you collect? (Similar to Homework 2 Q2, but now you should answer based on your collection progress, rather than just your plan). 5 points.

Rubric:

1. The data should:
   a. Specify the general nature of data, e.g. weather, hotel price, airfare, etc. 1 point.
   b. Specify exact data sources, e.g. a link to the dataset, a link to an API, or a link to a website to parse. 1 point.
   c. Have 2 or more data sources. 1 point.
   d. The data source should be large and recent. It should contain at least 100 samples (recommends 1000 samples or more), and is no older than 10 years from now. 2 points.

2. The data collection code should:
    a. Show your progress on data collection. You don't have to finish all the work by then. 4 points.
    b. Have a link to your github repository. If it's a private repository, you need to add mail@g1eb.com and yuzhongh@usc.edu as your collaborator in the repo settings. 1 point.
3. The sample data should:
    a. Have a description about the original data format, e.g. For structured datasets, specify its format. For API, specify what keys you are collecting. For Web parsing, specify what elements/attributes you are parsing. 2 points.
    b. Have a description about the saved data format. e.g., you save them as a CSV/JSON/XML file and describe its format; you save them in SQL/NoSQL database and describe its schema. 3 points.