CICS 397A – Fall 2020

Homework 1

Due Tuesday, October 6th at 11:59pm ET

You are encouraged to discuss the assignment in general with your classmates, and may optionally collaborate with one other student. If you choose to do so, you must indicate with whom you worked. Multiple teams (or non-partnered students) submitting the same code will be considered plagiarism.

Code must be written in a reasonably current version of Python (>3.0). You are free to use Python's standard modules for data structures and utilities, as well as the pandas, scipy, and numpy modules if you really want.

Gather Ye Data

For this assignment, your task is to create your own data set using content found online. The data set you create can be drawn from any source that interests you — for example, you might create a data set of news headlines, movie ratings, sports scores, financial indicators, political polls, etc. Ideally, you will find something that will be suitable for analyzing in future assignments, but you can change data sources in the future if you'd like. You can obtain your data in one of two ways: using a public API or web scraping.

Option A: Use a Public API

For the first option, you'll use a publicly available API to download data, similar to the lab exercise we did with the Genius.com API. You are free to use any API you want. To help get you started, the link below has a number of data sources listed:
https://github.com/public-apis/public-apis

The interface for each data source is different, so you will have to consult the documentation to determine which endpoints will have the data you want, or whether you need to do something to authenticate. Sometimes understanding these docs can be tricky, so please reach out to the course staff if you need help.

Also note that you may need to make multiple API calls to obtain the data you need. For instance, a news source might have article headlines at one endpoint, and information about the author at another; to create a data table with both, you'll need to make a call for each and then "stitch" them together.

Option B: Web Scraping

If you choose this option, you'll obtain data found on a web page by parsing the html source and extracting the data you want. To do this, you can use the `scraper.py` code that we looked at in class. Alternatively, you can use a more full-featured set of tools such as Scrapy or Beautiful Soup. Make sure you abide by the Terms of Service associated with whatever website you're targeting.

Requirements and Grading

You must create a Python script called `get_data.py` which will produce your data file. Your script must be runnable from a Python environment (a terminal or within VSCode) with no command line arguments.

The data set you create should be representable as a single table with at least 200 instances (rows) and 4 attributes (columns). Depending on your data source, it might take some wrangling to assemble it into a cohesive form. Once together, the data should be exported to a CSV file called `data.csv` using Python's `csv` module or a similar utility. We will run your program and examine the CSV output for correctness.

What to Submit

You should submit:

- Your API or scraping code, contained in a file called `get_data.py`
- The data file that gets created by your code, called `data.csv`
- A `readme.txt`, containing
    - Your name(s)
    - A short (2-3 sentence) description of your data set.
    - Notes or warnings about what you got working, what is partially working, and what is broken.