PH1975 Capstone Project

Total: 100 pts

Design and implement your own Python program that can do the following:

Q1. A scraper module that can collect paper title, author list, publication time, and abstract from PUBMED for the keyword **"HIV"** within a pre-specified time window **01/01/2020 – 08/30/2020**, and the retrieved data should be saved in the CSV format.

Hint: you may use BioPython and an example of this can be found on:
https://stackoverflow.com/questions/40433080/searching-pubmed-using-biopython-and-writing-to-csv
(20 pts)

Q2. A database module that can import the CSV file to SQLite to build a database automatically. Then implement SQL code to query the publications by author's name (i.e., input an author's name and find out and return all his/her publications). **(20 pts)**

Q3. A visualization module that can **i)** read the CSV file; **ii)** show the number of publications in each month; **iii)** generate and visualize the summary statistics for the publication numbers per month, including mean, SD, range, median, 1st to 3rd quartile; and **iv)** visualize the trend of the publication numbers over time (by months). Be creative about visualization (e.g., you can create your own dashboard). **(20 pts)**

Q4. Implement a demo using Jupyter Notebook (that is, you will submit a .ipynb file and all other Python code files so our TAs can run the demo code step by step and check the output). **(20 pts)**

Q5. Finally, write a report with the following sections to describe and summarize your work: **(20 pts)**

> **Section 1. Program design.** Please describe your ideas, draw your workflow diagram, and show the architecture design of your program.
>
> **Section 2. Implementation details.** Please describe all important implementation details (e.g., modules, classes/functions, 3-rd party packages and tools).
>
> **Section 3. Results.** Please include all required outputs (figures, tables, and/or numbers) in this section.
>
> **Section 4. User manual/guide.** You are the developers of your tool, please write a user manual so others can read and understand how to use your tool.

Notes:

- Grading will be based on the correctness of your answers and clarity of your report.
- The report should be in WORD or PDF format. We are NOT able to accept other formats like LaTex or Markdown.
- Please submit all your code files (.ipynb and .py files) , and the Jupyter Notebook file should contain all the outputs that show the success of each module (i.e., the demo should run without any error messages and give the right answer).
- You CAN use any 3rd party libraries or codes found online once their licenses permit or appropriate credits are given to the original developers.
- You do NOT need to implement any Graphical User Interface for Q3.
- It is possible that some articles are associated with multiple dates. Only one of these dates is the official publication date, and you may tell that from the text associated with that date (e.g., "accepted", "online preview", "published"). For simplicity, you may just use the last date.