Loughborough University

Coursework1 Assignment Specification
(40 % of the module assessment)
Submission Deadline: at 3pm on 13th January 2022

Data Preparation for Online Test Results

1 Problem Description

1.1 Overview

Your task is to write a Python program to create an SQLite database for a student monitoring system. You are given several CSV files and each file holds a raw dataset of students' online test results for the same undergraduate module. Your program needs to clean the datasets, format them into the desired form and store them into a database file.

1.2 Clean Data (25 marks)

The program should clean the raw datasets from the given files.

- It should read data from each CSV file and write them into one separate DataFrame (e.g., dfTest_1) and rename the columns so that the column names don't have any space, the "/" character and the mark information of the questions (e.g., "Q 1 /100" becomes "Q1").
- It must also fill the null values with 0.
- If any student has more than one result for the same test, then keep only the highest test result and remove the rest of them.
- The state and time-taken column are not needed for the monitoring system. They **must not** be included in the DataFrames.
- Make sure that the program does not change original DataFrames and produces cleaned DataFrames (e.g., dfTest_1 will be dfCleanTest_1).

1.3 Format Data (25 marks)

Student grades across the datasets are not standardised. Your program needs to make all grades out of 100 to normalise the grades in the DataFrames created in Section 1.2. For example, if a student's grade is 1 out of 4. The normalised grade will be 25 which is (1 / 4) *100. Make sure that the program renames the reformatted DataFrames appropriately (e.g., dfFormattedCleanTest_1).

1.4 Store Data (25 marks)

The program should create SQL tables and store cleaned and reformatted DataFrames from Section 1.3 into these tables. You could use a basic data model, which creates a database table for each DataFrame. Make sure that the tables have the correct data types.

1.5 Modify Data (25 marks)

One of the given files has the student results of a summative online test (namely, SumTest.csv). All other files are for the results of formative tests.
Use NumPy to generate 20 random integers between 0 and 150. Use those numbers as indices and extract the respective rows from dfFormatted CleanSumTest which is produced in Section 1.3. Randomly change the results of three questions in those rows and update the database tables accordingly. Make sure that the total of each row is also correctly updated in the tables.

2 What to Submit

1) "CW1.ipynb" file is a single Jupyter notebook:
   a) Ensure that
      i) all your results are presented in the Jupyter notebook,
      ii) your codes are executable, and
      iii) detailed comments are embedded inside the code. Additionally, you can provide any special instructions or warnings to the user (or assessor!), or draw attention to any aspects of the program that you are particularly proud of (please don't waste time by writing an excessive amount).
   b) See a copy of a Conda environment (yml file) on the module Learn page. Submitted code must work in the environment which includes
      i) Matplotlib.
      ii) Numpy.
      iii) Grapviz
      iv) Pandas

2) ResultDatabase.db is an SQLite database file which stores specified datasets which will be used for CW 2.
3) TestResult Folder contains all the given CSV files.

All the files above should be compressed into a zip file and submitted electronically as directed on the module Learn page.

3 Notes on Expectations:

Technical mastery of Python Your programs should show mastery of what you have been taught.
Design Your programs should be well structured for the task in hand so that it is as easy as possible for a programmer/user to understand the code structure and be able to develop it further.