Introduction to Programming for Data Science Assignment 2

Dr Kieran Hughes

January 2024

Instructions

This assignment is due by 11pm on Friday 22nd March. The answer to all your questions must be in a Jupyter workbook. Name your workbook as IntroProgAssignment2 - StudentID StudentNAME.ipynb. Label the start of each question and put any written answers in markdown cells. All answers should appear in order, so your answer to question 2 B should be below the answer to question 2 A and above the answer to question 2 C.

Include any libraries when you first require them. Ensure when you are uploading your work you only upload your own work and not any dataset. Just upload the ipynb file and the db file requested. Do not zip anything. This is NOT a group assignment. You must complete the assignment individually. If you share your work, your marks will be divided. Some students may be asked to demonstrate their knowledge of the submitted work.

A Note on modules and libraries: For completing this assignment you are only to use basic Python modules and libraries, mentioned either in the Assignment description, or ones used in class. These include BeautifulSoup, pandas, sqlite3, re, matplotlib, requests and numpy. Other libraries are not permitted even if they will make the assignment easier. (you may not use nltk, Counter or others that have not been referred to in class or in the assignment).

Please note that you will be docked 5% for each day you are late with submission, in accordance with IT Sligo Marks and Standards.

Questions

1. A list of research projects at the university is available on the ATU website (see link below).

   `https://www.itsligo.ie/research/research-projects/`.

   (a) Import Beautiful Soup, fetch the HTML content from the web page and create a Beautiful Soup object.

   (b) Read in all tables, find the right one, and assign it to `wanted_table`

   (c) Load the contents of the table `wanted_table` into a data frame.

   (d) Tidy up the data frame including column headings and removing all columns with no information.

   (e) Use the data frame to print a list of the students whose project title specifically refers to the environment.

2. In this question, you are tasked with exploring the "Northwind" database, a fictional database representing a company's sales data, using Python's built-in sqlite3 package. The database file "northwind.db" is on the Moodle page and documentation for sqlite3 can be found here `https://docs.python.org/3/library/sqlite3.html`. All code should be readable, commented, and handle errors where appropriate.

   (a) Import sqlite3 and write a Python function to establish a connection to the Northwind database.

   (b) Write and execute a query to retrieve the table names from the `sqlite_master` table and display the table names.

   (c) Write and execute a query to retrieve column information for the `Orders` table and display the results.

   (d) Write a Python function called `Histogram_Number_Of_Orders_By_Month`. This function should take 2 arguments, namely a connection object called `connection` and an integer representing the year called `year`. The function should retrieve the number of orders per month for the year `year`. When writing the query use the function `strftime` to extract the month part from the `OrderDate` and alias it as `Month`. Fetch all the rows and convert the result set to a data frame using pandas. Use mathplotlib to plot a histogram of the result set.

   (e) Write a function called `update_contact_name` that has 3 arguments, namely a connection object, the name of the supplier and the new contact name. The function should update the contact name for the supplier in the database via the connection. Use this function to update the contact name for "Bigfoot Breweries" to "Paul Downs".

   (f) By writing and executing a query, print a list containing the product name, the product category name and the unit price of the product for each product. The list should be ordered first by product category name in alphabetical order and then by unit price in ascending order. Note that the `Products` table and the `Categories` table both have a column `CategoryID` and so a left Join might be useful.

3. This question again uses the Northwind data base. All code should be readable, commented, and handle errors where appropriate.

   (a) Write Python code to create a new table called `Tagetproducts` if it does not already exist. The table has 3 columns, namely `TagetproductsID` (Integer), `TagetproductsName` (text) and `Price` (REAL). Apply the constraint `PRIMARY KEY` to `TagetproductsID`.

   (b) Insert the following information into `Tagetproducts` an commit the changes.

TagetproductsID	TagetproductsName	Price
1	lough Gill Cutback IPA	3.10
2	Guinness Stout	2.50
3	White Hag little Fawn	3.49

   (c) Update the price of the targeted product lough Gill Cutback IPA to 2.90.

   (d) Write a function `print_table_contents` that has 2 arguments, namely `table_name` and `connection`. The function should fetch all rows from the table via `connection`, print the column names and print each row.

   (e) Write a function `remove_expensive_target_products` that has 2 arguments, namely `max_price` and `connection`. The function should print the `Tagetproducts` table. Then it should delete, from the `Tagetproducts` table, rows with price larger than the specified value `max_price` and commit the changes. Finally it should print the `Tagetproducts` table again. Call the function `remove_expensive_target_products` where `max_price` is 3.

   (f) Upload the db file with your submission.

4. In this question you will be using the Python re module and the file Reviews.csv which is on the Moodle page.

   (a) Import pandas and load the dataset from the Reviews.csv file into a pandas data frame and print the first few rows of the data frame.

   (b) Write a regular expression pattern that is looking for occurrences of the 4 ordered characters in "good" (not including the quotation marks). The re pattern should not include any word boundaries (\b), and so should find an occurrence in "goodie". The reviews are stored in the "Text" column of the data frame. Use this re pattern to count the number of reviews that contain at least one occurrence of "good".

   (c) Write a regular expression pattern that is looking for occurrences of the word "good" but excluding cases where it is preceded by the word "no". Use this re pattern to count the number of reviews that contain at least one occurrence of the word "good" that it is not preceded by the word "no".

   (d) The User ID's are stored in the "UserId" column of the data frame. Using regular expressions find all user ID's that either start with "A9" or contain 7 consecutive digits.

   (e) Write a regular expression pattern that is looking for occurrences of at least one even digit between "9" characters. Of the User ID's found in part (d), how many match with this pattern. Note the ID "A90I4J49NU3XN" is a match,