Programming for Data Science

Content and skills covered by the assignment

- Understand advanced concepts of programming in Python.
- Have a critical appreciation of the main strengths and weaknesses of a range of Python packages and understand how to use them.
- Have a critical appreciation of how to acquire and clean datasets for analysis.
- Understand how to manipulate potentially large datasets efficiently.
- Be able to write computer programs in Python using industry-standard packages.
- Be able to select appropriate data structures for modelling various data science scenarios.
- Be able to select the appropriate algorithm and programming package for a given problem.
- Be able to write a computer program in Python to collect or read data from available sources, and clean these datasets using the appropriate packages.
- Effective written communication.
- Planning, organising, and time-management.
- Problem solving and analysis.

For Question 1, you are asked to perform the following tasks based on the following target website, which contains artificial content designed for this assignment: https://sitescrape.awh.durham.ac.uk/comp42315/

a) Please design and implement a solution to crawl the publication title, year and author list of every unique publication record on the target website.

b) This information should then be stored and displayed in an appropriate format. Displayed to the user should be: publication title, year, author list, number of authors and impact factor. [Consider what would be an appropriate way of displaying this information to a user. HINT: You will need to develop an algorithm to work this out.]

c) The records should be manipulatable by the user, at minimum you should be able to displayed sorted information according to: descending year values, descending number of author values, the titles from A to Z, and finally by the impact of the papers.

d) Explain your design and highlight any features in this question's report part of your Jupyter Notebook in no more than 300 words. Reflect upon the importance of well structured HTML.

2. For this question, you will analyse an existing corpus of privacy policies, following the paper https://arxiv.org/abs/2201.08739. You need to download policy-metadata.zip (25.4 MB) and policy-texts.zip (375.1 MB) from https://zenodo.org/records/7426577. You do not need to download any of the labelled datasets or policies, only the raw text and the metadata. This question consists of 3 parts.

a) For each policy text file, you need to compute the Flesch Reading Ease score, the SMOG index, the Coleman-Liau index, the Flesch-Kincaid Grade and the Dale-Chall readability score. After manually identifying which of those metrics are given in policy metadata.csv, you need to compare the metrics you’ve obtained against those provided, identifying any difference. If you have found any difference, reflect on the possible reasons.

b) For each policy text file, you need to compute a sentiment analysis, including polarity and subjectivity, and present an overall analysis of the corpus, including a box plot for each metric, as well as as analysis of the median of each metric over the years.

c) Finally, you need to do a correlation analysis, assessing if any of the readability metrics in (a) are correlated to the sentiment analysis metrics in (b).

3. For this question, you are asked to perform the task based on the target dataset (drone.csv). The target drone dataset contains different features related to the drone physical attributes. The target variable is 'warning', i.e. warning level. If the warning level is high then a drone must be grounded immediately to avoid any damage to the drone itself as well as the surroundings (in case of a crash). Design and implement the solution to use data analysis and visualisation for the following tasks:

a) You are required to extract a subset that includes the 'defined_features' and the 'target_variable' (in the subset, there will be 11 features in total including target variable). You are required to extract the 'defined_features' that are as indicated below: defined_features = ['voltage_v', 'voltage_filtered_v', 'current_a', 'current_filtered_a', 'discharged_mah', 'remaining', 'scale', 'load', 'ram_usage'] Perform exploratory data analysis on the drone dataset ('defined_features'), highlight the features that are statistically important and highly correlated to the 'target variable', and visualise them legibly using an appropriate visual method. Save the statistically important features as a subset 'selected_features' and compare them with the 'defined_features'. Highlight any differences and report your findings.

b) Perform analysis to identify the complex relationship between drone physical attributes ('selected_features') and the warning level ('target_variable') using a probabilistic method. Highlight and visualise the attributes with the highest probabilistic relationship with the target variable. Justify the design choice and showcase the findings using an appropriate visualisation tool.

c) Perform predictive analysis using at least one machine learning algorithms and validate its performance based on suitable performance metrics. Justify the design choice and showcase the findings using an appropriate visualisation tool.