ICT707

Data Science Practice

Task 3

Semester 2

Assessment and Submission Details

Marks: 40% of the Total Assessment for the Course

Due Date: 11:59pm Sunday, Week 12

Submit your assignment to Blackboard Task 3. Please follow the submission instructions in Blackboard.

The assignment will be marked out of a total of 100 marks and forms 40% of the total assessment for the course. **ALL** assignments will be checked for plagiarism by SafeAssign system provided by Blackboard automatically.

Refer to your Course Outline or the Course Web Site for a copy of the "Student Misconduct, Plagiarism and Collusion" guidelines.

Late submission will be penalised according to the policy in the course outline. Please note Saturday and Sunday are included in the count of days late.

Requests for an extension to an assignment **MUST** be made to the course coordinator prior to the date of submission and requests made on the day of submission or after the submission date will only be considered in exceptional circumstances. Assignment submission extensions will only be made using the official University guidelines.

Assignment Task

This assignment consists of two deliverables, being:

- One code implementation (40%). This requires a zip file which should include:
  - The code file in Jupyter Notebook format.
  - Relevant data set files.
  - A pdf or HTML file which is printed/converted from your Notebook **after having all cells executed**.
- A report (60%). The report must be uploaded as a separate file.

Part I - PySpark source code (40%)

Important Note:

- For code reproduction, your code must be self-contained. That is, it should not require other libraries besides PySpark environment we have used in the semester. The data files are packaged properly with your code file.
- The data sets used in the lecture slides should not be used as the data set of the assignment. This will result in 0 mark for the coding component.

In this component, we need to utilise Python 3 and PySpark to complete the following data analysis tasks:

1. Exploratory data analysis
2. Recommendation engine
3. Classification

You need to choose a dataset from Kaggle (https://www.kaggle.com/datasets) to complete these tasks. Remember to include the data set file in you source code submission.

Note: In your notebook, please use Heading 1 Markdown cell to separate each sub task.

Task I.1: Exploratory data analysis

This subtask requires you to explore your dataset by

- telling its number of rows and columns,
- doing the data cleaning (missing values or duplicated records) if necessary
- selecting 3 columns, and drawing 1 plot (e.g. bar chart, histogram, boxplot, etc.) for each to summarise it

Task I.2: Recommendation engine

This subtask requires you to implement a recommender system on Collaborative filtering with Alternative Least Squares Algorithm. You need to include

- Model training and predictions
- Model evaluation using MSE

Task I.3: Classification

This subtask requires you to implement a classification system with Logistic regression. You need to include

- Logistic Regression model training
- Model evaluation

Part II –Report (60%)

You are required to write a report with the following content:

- Provide a high-level survey on the advances of data science in the past 2 years.
- Compare the features of Spark version 2.4 that we used this semester and the new version 3.0.
- Explain your design and implementation of the machine learning parts in your code, including the following topics:
  - Background of your selected data set
  - For each task, which learning algorithm is used and what are its key parameters and how you set them up
  - For each task, provide comments/evaluation for the model learnt

Your report should use the following template:

Table of Contents

1.0 Advancement of Data Science (500 words)

2.0 Comparison of Spark 2.4 and 3.0 (250 words)

3.0 Machine Learning Implementation (250 words)
- 3.1 Data set
- 3.2 Collaborative filtering
  - Features of the model, key parameters and configuration
  - Evaluation
- 3.3 Logistic regression
  - Features of the model, key parameters and configuration
  - Evaluation

References

The marking rubrics are viewable on the blackboard.

Report Format

Your report should be about 1000 words.

The report MUST be formatted using the following guidelines:

- Title Page – Must not contain headers, footers, or page numbering. Include your name as the report's author.
- Header – Report title

- Footer – your name and the page number
- Paragraph text – 12 point Calibri single line spacing
- Headings – Arial in an appropriate type size
- Margins – 2.5cm on all margins
- Page numbering
    - Introduction and onwards to use conventional numerals (1, 2, 3, 4) starting at page 1 from the introduction.
- The report is to be created as a single Microsoft Word document (version 2007 or later). No other format is acceptable and doing so will result in the deduction of marks.

Please follow the conventions detailed in:

Summers, J. & Smith, B., 2014, *Communication Skills Handbook*, 4th Ed, Wiley, Australia.

Referencing

The report is to include (at least 5) appropriate references and these references should follow the Harvard method of referencing. Note that ALL references should be from journal articles, conference papers, technical papers or a recognized expert in the field. DO NOT use Wikipedia as a reference. The use of unqualified references will result in the deduction of marks.

Assignment Return and Release of Grades

Assignment grades will be available on the blackboard in two weeks after the submission. Details of marking will also be accessible via online rubrics on the blackboard.

Where an assignment is undergoing investigation for alleged plagiarism or collusion the grade for the assignment and the assignment will be withheld until the investigation has concluded.

Assignment Advice

This assignment will take several weeks to complete and will require a good understanding of machine learning and PySpark for successful completion. It is imperative that students take heed of the following points in relation to doing this assignment:

1. Ensure that you clearly understand the requirements for the assignment – what must be done and what are the deliverables.
2. If you do not understand any of the assignment requirements – Please ASK your tutor.
3. Each time you work on any aspect of the assignment reread the assignment requirements to ensure that what is required is clearly understood.
4. We have practiced nearly all coding tasks in DataCamp before. If you have any difficulty, redoing the practices in DataCamp is recommended.

5. Prior to submitting your code, you should ensure not only that it executes as required, but also looks professional. It is expected that you adhere to python standards for naming and indenting. All methods should be adequately documented such that another programmer examining your code will readily know what the code is doing.

Plagiarism and Collusion Advice

1. All work must be submitted through SafeAssign.
2. SafeAssign will pick up any similarities between work online as well as work from other students (in this semester and previous)
3. Please make sure you reference your work properly. If you are using any material from the internet or any books from the library, you need to cite the work correctly. Failure to do so will result in possible cases of Academic Misconduct.
4. Please do not share your work with other students. Do not give anyone your files to have a look. SafeAssign will pick up collusion, but keep in mind the percentages for Collusion may not report accurately until all student assignments have been submitted. Both the person copying and the person providing will potentially be held accountable.
5. You can submit a draft assignment through SafeAssign before making the actual submission.

End of Assignment