University of Lincoln Assessment Framework
Assessment Briefing Template 2022-2023

Module Code & Title: CMP3749M – Big Data

Contribution to Final Module Mark: 90% (However, this assignment will be marked out of 100%)

Description of Assessment Task and Purpose:

As a data scientist, your main objective is to organize and analyse data regardless how big or small the data is, often employing typical data science software. The analysis made by a data scientist must be easy enough to understand for all the stakeholders including those who have no knowledge of data science. The objective of this task is to show that you can perform an analysis over a data set to guide the stakeholders to understand the data. The data can be downloaded from Blackboard in the assessment documents area. The data needs to be analysed using the data science tools and techniques you were taught in class and contain the information detailed in the Report Guidance (see below).

You are required to write and submit a report where you need to provide answers to all questions, discuss how you completed the tasks outlined in the report guidance. In addition, you are to provide the Python (Pyspark) code you've developed for these tasks. You are expected to go into sufficient depth to demonstrate knowledge and critical understanding of the relevant processes involved. Note that most of the marks will stem from the clarity of your report, with the source code used as evidence.

Report Guidance

You are to submit a single report and associated python source code zip file containing the following two tasks:

- **Task 1 – Analysis of Nuclear Plants dataset** (strict max 2500 words) (70%)
- **Task 2 – MapReduce for Margie Travel dataset** (strict max 1000 words) (30%)

You must split the discussion into two distinct sections that provide a full and reflective account of the processes undertaken. You are expected to answer all questions in each task in detail, perform all analysis on your own (i.e., as individual work), and provide all Python (PySpark) scripts in one ZIP file as supporting evidence.

Task 1 – Analysis of Nuclear Plants dataset (strict max 2500 words) (70%)

You must clearly identify this part of the report as “Task 1 – Analysis of Nuclear Plants dataset”.

The dataset is of pressurised water reactors (a type of nuclear reactors) with various measurements in different parts of the reactor, including vibration, pressure and power levels. The first column in the spreadsheet indicates the status of the reactor, i.e., ‘normal’ or ‘abnormal’. All the other columns are features which could help us to gain insights into the status of each reactor. You are asked to provide an analysis over this data to discuss if these features could be potentially used to predict whether a reactor is normal or abnormal.

Download the data set named `nuclear_plants_small_dataset.csv’ from Blackboard on Blackboard under **Task_1_dataset**, then write Python (PySpark) code to accomplish the following:

1- As a first step, you need to load the data from the file `nuclear_plants_small_dataset.csv’ into a Pyspark DataFrame. Before making any analysis, it is required to know if there are missing values in the data. Are there any missing values? Discuss how you will deal with missing values, even if there are no missing values in this data set.

2- It is beneficial to understand the data by looking at the summary statistics. There are two groups of subjects (i.e., the normal group and the abnormal group) in this dataset. For each group, show the following summary statistics for each feature in a table: minimum, maximum, mean, and median values. For each group, plot the box plot for each feature.

3- To understand the relationship between features. If two features have high correlations, using only one of them could be enough for our analysis. Show in a table the correlation matrix of the features, where each element in the matrix shows the correlation coefficient of two features. Discuss your observations on the correlation matrix. Are there any features which are highly correlated? In any case, we will use all the features in the following tasks.

4- Shuffle the data samples and split it into a 70% training set and a 30% test set. How many examples in each group for the training dataset? How many examples in each group for the testing dataset?

5- Train a decision tree, a support vector machine model and an artificial neural network using the training set, and then apply the trained classifiers to the test set. You will obtain the predicted labels for the test set. Now evaluate and compare the classifiers, respectively, by computing the error rate (`Incorrectly Classified Samples divided by `Classified Sample’). Calculate the sensitivity and specificity. Briefly discuss the error rate, sensitivity and specificity.

Task 2 – MapReduce for Margie Travel dataset (strict max 1000 words) (30%)

You must clearly identify this part of the report as “Task 2 – MapReduce for Margie Travel dataset”.

The dataset is for Margie's Travel (MT) provides concierge services for business travelers. In an increasingly crowded market, they are always looking for ways to differentiate themselves and provide added value to their corporate customers. There are two files containing lists of data. These are located on Blackboard under **Task_2_dataset**. The coursework data folder includes the files:

- AComp_Passenger_data_no_error.csv: this data file contains details of passengers that have flown between airports over a certain period. The data is in a comma delimited text file, one line per record, using the following format:

Col. #	Field	Format
1	Passenger id	$XXXnnnnXXn$
2	Flight id	$XXXnnnnX$
3	From airport IATA/FAA code	XXX
4	Destination airport IATA/FAA code	XXX
5	Departure time (GMT)	n [10] (Unix ‘epoch’ time)
6	Total flight time (mins)	n [1. .4]

X is Uppercase ASCII

n is digit 0...9

$[n..m]$ is the min/max range of the number of digits/characters in a string.

- top30_airports_LatLong.csv: The second data file is a list of airport data comprising the name, IATA/FAA code, and location of the airport. The data is in a comma delimited text file, one line per record using the following format:

Col. #	Field	Format
1	Airport Name	X [3. .20]
2	Airport IATA/FAA code	XXX
3	Latitude	$n.n$ [3. .13]
4	Longitude	$n.n$ [3. .13]

There are two additional data input files which can be used for analysis and validation however should not be used for the final execution of the implementation:

- AComp_Passenger_data.csv: there are various errors in this data file, which illustrate a range of potential errors that could occur when handling large scale data from multiple, sometimes unreliable, sources. It is not necessary to handle this file and address these errors in your application. This file is instead provided to highlight the requirement of error handling in MapReduce applications.

- AComp_Passenger_data_no_error_DateTime.csv: it may use to convert date/time data from Unix epoch time to a human readable format for use when debugging and for validation purposes.

Write a Python code, must use MapReduce in Pyspark, to accomplish the following:

1- Determine the number of flights from each airport; include a list of any airports not used.

2- Create a list of flights based on the Flight id, this output should include number of passengers, relevant IATA/FAA codes, and departure and arrival times (times converted to HH:MM format).

3- Calculate the line-of-sight (nautical) miles for each flight and the total travelled by each passenger and thus output the passenger having earned the highest air miles.

Learning Outcomes Assessed:

On successful completion of this component a student will have demonstrated competence in apply data science toolkits in a range of applications and solve real-world problem.

- LO2: Apply data science toolkits in a range of applications and solve real-world problem

Knowledge & Skills Assessed:

Subject Specific Knowledge, Skills and Understanding: Literature searching, Referencing, Numeracy, Project Planning, Techniques and Skills in Data Science, Subject-specific knowledge.

Professional Graduate Skills: Independence and personal responsibility, adaptability, written communication, creativity, critical thinking, IT skills, self-reflection and life-long learning, problem solving, effective time management, working under pressure to meet deadlines.

Emotional Intelligence: Self-awareness, self-management, motivation, resilience, self-confidence.

Career-focused Skills: Big Data tools, techniques, skills and attributes required by employers, a range of problem strategies to present skills and attributes to employers.

Assessment Submission Instructions:

1. **Report** (Tasks 1 and 2)

The submission deadline of this assignment is included in the School Submission dates on Blackboard. You must make an electronic submission of your report to the Turnitin upload area for Assessment 2.

The report must:

- Contain your name, student number, student email address, and module name.
- Be in single PDF with
    - no more than 2500 for Task 1 – Analysis of Nuclear Plants dataset
    - no mode than 1000 for Task 2 - MapReduce for Margie Travel dataset
- Be formatted single-spaced with 11pt font size; Do not include this briefing document.

2. Source Code

Your python (Pyspark) code, should be submitted as a single zip archive, to the assessment item 2 supporting documents area on blackboard. This zip archive should contain your python code for all tasks and include code comments where appropriate to aide understanding.

All elements of both tasks are individually assessed. Your work must be presented according to the School of Computer Science guidelines for the presentation of assessed written work. Please make sure you have a clear understanding of the grading principles for this component as detailed in the accompanying Criterion Reference Grid. Your citations and referencing should be in accordance with University guidelines.

If you are unsure about any aspect of this assessment, please seek the advice of the module tutors (contact details on blackboard).

The submission deadline of this assignment is included in the School Submission dates on Blackboard.

If you are unsure about any aspect of this assessment component, please seek the advice of the module lecturers contact details are available on blackboard.

Date for Return of Feedback:

Feedback will be provided on blackboard within three weeks of submission.

Please note that all work is assessed according to the University of Lincoln Management of Assessment Policy and that marks awarded are provisional on Examination Board decisions (which take place at the end of the Academic Year.

Feedback Format:

Summative feedback will be provided on Blackboard according to CRG criteria (see CRG file).

Additional Information for Completion of Assessment:

Students are encouraged to use any lecture and their own personal notes to assist them with the completion of the assessment. Also, students are allowed to use any library and/or relevant online resource as a guide on how to solve the assessment problems.

Assessment Support Information:

Students are encouraged to seek assistance from any member of the delivery team and particularly from the module coordinator as means to complete the assessment.

Important Information on Dishonesty & Plagiarism:

University of Lincoln Regulations define plagiarism as 'the passing off of another person's thoughts, ideas, writings or images as one's own...Examples of plagiarism include the unacknowledged use of another person's material whether in original or summary form. Plagiarism also includes the copying of another student's work'.

Plagiarism is a serious offence and is treated by the University as a form of academic dishonesty. Students are directed to the University Regulations for details of the procedures and penalties involved.

For further information, see plagiarism.org