SCHOOL OF ARCHITECTURE, COMPUTING & ENGINEERING

Submission instructions

- Cover sheet to be attached to the front of the assignment when submitted
- Question paper to be attached to assignment when submitted
- All pages to be numbered sequentially
- All work has to be presented in a ready to submit state upon arrival at the ACE Helpdesk. Assignment cover sheets or stationery will **NOT** be provided by Helpdesk staff

Module code	CN7031		
Module title	Big Data Analytics		
Module leader	Amin Karami		
Assignment tutor	A Karami, F Jafari, MA Ghazanfar, N Qazi		
Assignment title	Big Data Analytics: Coursework		
Assignment number	1		
Weighting	100%		
Handout date	Week 5 (30th October 2020)		
Submission date	Presentation: Week 12 (14th-18th December 2020) Turnitin Submission: 25th December 2020 (midnight)		
Learning outcomes assessed by this assignment	1-8		
Turnitin submission requirement	Yes	**Turnitin GradeMark feedback used?**	No
UEL Plus Grade Book submission used?	No	**UEL Plus Grade Book feedback used?**	No
Other electronic system used?	Yes	**Are submissions / feedback totally electronic?**	Yes
Additional information			

Form of assessment:

☐ Individual work ☒ Group work

For **group work** assessment which requires members to submit both individual and group work aspects for the assignment, the work should be submitted as:

☒ Consolidated single document ☐ Separately by each member

Number of assignment copies required:

☒ 1 ☐ 2 ☐ Other

Assignment to be presented in the following format:

☒ On-line submission
☐ Stapled once in the top left-hand corner
☐ Glue bound
☐ Spiral bound
☐ Placed in a A4 ring bound folder (not lever arch)

Note: To students submitting work on A3/A2 boards, work has to be contained in suitable protective case to ensure any damage to work is avoided.

Soft copy:

☐ CD (to be attached to the work in an envelope or purpose made wallet adhered to the rear)
☐ USB (to be attached to the work in an envelope or purpose made wallet adhered to the rear)
☒ Soft copy not required

CN7031 - Big Data Analytics
Group assignment 2020-21 Academic Year

This coursework (CRWK) must be attempted in **the groups of 4 or 5 students**. This coursework is divided into two sections: (1) Big Data analytics on a real case study and (2) group presentation. All the group members **must attend the presentation**. Presentation would be online through Microsoft Teams. **If you do not turn up in the presentation date with the video call, you will fail the module**.

Overall mark for CRWK comes from two main activities as follows:

1- Big Data Analytics report (around 3,000 words, with a tolerance of ± 10%) in HTML format (60%)
2- Presentation (40%)

Marking Scheme

Topic	Total mark	Remarks *(breakdown of marks for each sub-task)*	
Big Data Analytics using Spark SQL	30	*(6)*	Providing 2 queries using Spark SQL.
		(14)	Developing advanced SQL statements. Refer to: https://spark.apache.org/docs/3.0.0/sql-ref.html
		(10)	Visualizing the outcomes of queries into the graphical and textual format, and be able to interpret them.
Big Data Analytics using PySpark	60	*(45)*	Analyzing the dataset through 3 statistical analytics methods including advanced descriptive statistics, correlation, hypothesis testing, density estimation, etc.
		(15)	Designing one classifier, then evaluate and visualize the accuracy/performance. Applying a multi-class classifier is considered for full mark.
Documentation	10	*(10)*	Write down a well-organized report for a programming and analytics project.
Total:	**100**		

IMPORTANT: you must use CRWK template in the HTML format, otherwise it will be counted as plagiarism and your group mark would be zero. Please refer to the "THE FORMAT OF FINAL SUBMISSION" section.

Good Luck!

Big Data Analytics using Spark

CN7031 – Big Data Analytics

(1) Understanding Dataset: CSE-CIC-IDS2018[1]

This dataset was originally created by the University of New Brunswick for analyzing DDoS data. You can find the full dataset and its description here. The dataset itself was based on logs of the university's servers, which found various DoS attacks throughout the publicly available period to generate totally 80 attributes with 6.40GB size. We will use about 2.6GB of the data to process it with the restricted PCs to 4GB RAM. Download it from here. When writing machine learning or statistical analysis for this data, note that the Label column is arguably the most important portion of data, as it determines if the packets sent are malicious or not.

a) The features are described in the "IDS2018_Features.xlsx" file in Moodle page.

b) The labels are as follows:

```
'DoS attacks-Hulk',
'DDOS attack-LOIC-UDP',
'Label',
'DDOS attack-HOIC',
'DoS attacks-Slowloris',
'SQL Injection',
'SSH-Bruteforce',
'Brute Force -Web',
'Infilteration',
'Benign',
'DoS attacks-GoldenEye',
'DoS attacks-SlowHTTPTest',
'Bot',
'Brute Force -XSS',
'FTP-BruteForce'
```

- "Label": normal traffic
- "Benign": susceptible to DoS attack

c) In this coursework, we use more than 8.2-million records with the size of 2.6GB. As a big data specialist, firstly, we should read and understand the features, then apply modeling techniques. If you want to see a few records of this dataset, you can either use [1] Hadoop HDFS and Hive, [2] Spark SQL or [3] RDD for printing a few records for your understanding.

[1] Source: https://registry.opendata.aws/cse-cic-ids2018/ & https://www.unb.ca/cic/datasets/ids-2018.html

(2) Big Data Query & Analysis using Spark SQL [30 marks]

This task is using Spark SQL for converting big sized raw data into useful information. Each member of a group should **implement 2 complex SQL queries (refer to the marking scheme)**. Apply appropriate visualization tools to present your findings numerically and graphically. Interpret shortly your findings.

You can use https://spark.apache.org/docs/3.0.0/sql-ref.html for more information.

- **What do you need to put in the HTML report per student?**
    1. At least two Spark SQL queries.
    2. A short explanation of the queries.
    3. The working solution, i.e., plot or table.

- **Tip:** The mark for this section depends on the level of your queries complexity, for instance using the simple *select* query is not supposed for a full mark.

(3) Advanced Analytics using PySpark [60 marks]

In this section, you will conduct advanced analytics using PySpark.

3.1. Analyze and Interpret Big Data using PySpark (45 marks)

Every member of a group should analyze data through **3 analytical methods** (e.g., advanced descriptive statistics, correlation, hypothesis testing, density estimation, etc.). You need to present your work numerically and graphically. Apply tooltip text, legend, title, X-Y labels etc. accordingly.

Note: we need a working solution without system or logical error for the good/full mark.

3.2. Design and Build a Machine Learning (ML) technique (15 marks)

Every member of a group should go over https://spark.apache.org/docs/3.0.0/ml-guide.html and apply one ML technique. You can apply one the following approaches: Classification, Regression, Clustering, Dimensionality Reduction, Feature Extraction, Frequent Pattern mining or Optimization. Explain and evaluate your model and its results into the numerical and/or graphical representations.

Note: If you are 4 students in a group, you should develop 4 different models. If you have a similar model, the mark would be zero.

(4) Documentation [10 marks]

Your final report must follow the **"The format of final submission"** section. Your work must demonstrate appropriate understanding of building a user friendly, efficient and comprehensive analytics report for a big data project to help move users (readers) around to find the relevant contents.

THE FORMAT OF FINAL SUBMISSION

1- You can use either Google Colab (https://colab.research.google.com/) or Ubuntu VMWare for this CRWK.

2- You have to convert the source code (*.ipynb) to **HTML**. Watch the video in the Moodle about "how to submit the report in HTML format".

3- Upload **ONLY one single HTML file per group** into Turnitin in Moodle. One member of each group must submit the work, **NOT** all members. The name of the file must be in the format of "`Your-Group-ID_CN7031`", such as ***Group200_CN7031.html*** if you are belonging to the group 200.

4- The submission link will be available from week 10, and you are free to amend your submitted file several times before submission deadline. Your last submission will be saved in the Moodle database for marking.

PLAGIARISM

If there are copied PySpark codes from somewhere or someone else, all the group members will get zero, and should attend the "breach of regulation" committee for further explanations and the probable additional penalties.

FEEDBACK TO STUDENTS

Feedback is central to learning and is provided to students to develop their knowledge, understanding, skills and to help promote learning and facilitate improvement.

- Feedback will be provided as soon as possible after the student has completed the assessment task.
- Feedback will be in relation to the learning outcomes and assessment criteria.
- It will be offered via Turnitin GradeMark or Moodle post.

As the feedback (including marks) is provided before Award & Field Board, marks are:

- Provisional
- available for External Examiner scrutiny
- subject to change and approval by the Assessment Board

ASSESSMENT FORM FOR PRESENTATION

CN7031 – Big Data Analytics (40%)

Students have to fill this section correctly. Assessors will not be liable for any mistakes.
Group No: .....
1st Student (full name and ID):
2nd Student (full name and ID):
3rd Student (full name and ID):
4th Student (full name and ID):
5th Student (full name and ID):

Assessment Criteria:

Criteria	1st	2nd	3rd	4th	5th	Mark
Demonstrate/interpret Spark SQL queries						10
Understand Spark and its mechanism						5
Demonstrate/interpret PySpark codes						15
Ability to answer questions						10
Overall mark						**40**

Date & Time: …….

Assessors' signature and comments: