FACULTY OF SCIENCE
ACADEMY OF COMPUTER SCIENCE AND SOFTWARE ENGINEERING

MODULE	IT8X030: BIG DATA ANALYTICS
CAMPUS	AUCKLAND PARK CAMPUS (APK)
ASSESSMENT	NOVEMBER EXAM 2020

DATE: 2020-11 **SESSION:** 8:30 - 10:30

ASSESOR(S): PROF D.T. VAN DER HAAR

MODERATOR: DR PATRICIA E.N. LUTU (UP)

DURATION: 120 MINUTES **MARKS:** 80

Please read the following instructions carefully:

1. You must complete this examination yourself within the prescribed time limits.
2. YYou are bound by all university regulations. Please **take special note of the regulations** regarding assessment, plagiarism, and ethical conduct.
3. You must complete and submit the *"Honesty Declaration : Online Assessment"* document along with your submission to EVE. No submissions without an accompanying declaration will be marked.
4. Your answers together with the declaration must be submitted as a pdf file. The file name should have the following format:
   STUDENTNUMBER_SURNAME_INITIALS_SUBJECTCODE_ASSESSMENT e.g.
   202012345_SURNAME_IAM_IT8X030_FSAO.pdf
5. Additional time for submission is allowed for as per the posted deadlines on EVE. If you are experiencing technical difficulties related to submission please contact me as soon as possible.
6. No communication concerning this examination is permissible during the examination session except with Academy staff members. The invigilator is available via email (dvanderhaar@uj.ac.za) and on the "UJ Big Data" Discord server throughout the assessment (https://discord.gg/6cCm7N4).
7. This paper consists of 4 pages excluding the cover page.

QUESTION 1

(a) Outline three key **differences** between the fields of **Big Data** and **Machine Learning**. [3]

(b) Describe how one would **deploy and monitor** a data science project. [4]

Total: 7

QUESTION 2

One way of performing data-driven decisions is to leverage retail customer buying behaviour to construct discount packages that maximise volume focused revenue. An excellent way to achieve this is through the use of basket analysis on buying behaviour captured through a rewards programme. Below is a table representing nine (9) store transactions (baskets) and six products (items). A "1" indicates membership of the item in the transaction.

	A	B	C	D	E	F
0	0	0	1	1	1	0
1	0	1	0	0	1	1
2	1	1	1	1	0	0
3	0	0	1	0	1	1
4	1	0	0	1	0	1
5	0	0	1	1	1	1
6	0	1	1	0	1	0
7	1	1	0	1	0	1
8	1	1	0	1	1	1

(a) *Derive* two (2) viable **association rules** with more than one antecedent (in the format $X, Y \rightarrow Z$), which would apply to the above baskets. [2]

(b) *Calculate* the **support** and **confidence** for the above identified association rules. [4]

(c) *Are* there any interesting association rules for the above baskets? Provide a brief explanation for your answer. [1]

Total: 7

QUESTION 3

(a) Assuming the *above* baskets for the previous question are boolean-based mappings of items in a basket, what is the **Jaccard** distance between transaction (row) **0** and **1**? [2]

(b) Derive **bigrams** for the following string: [3]
The cat sat on the mat

(c) Briefly *describe* **locality sensitive hashing**. [2]

Total: 7

QUESTION 4

(a) Briefly *describe* how divisive agglomerative clustering works. [2]

(b) *Discuss* how the **BFR** algorithm works. [5]

Total: 7

QUESTION 5

(a) Given the following matrix A, **derive** AA^T: [6]

$$\begin{bmatrix} 2 & 4 & 2 & 2 & 4 \\ 5 & 4 & 5 & 8 & 3 \\ 3 & 1 & 8 & 1 & 4 \\ 2 & 1 & 4 & 4 & 6 \\ 8 & 6 & 5 & 7 & 5 \end{bmatrix}$$

(b) What is a **disadvantage** of the CUR algorithm within the context of dimensional reduction? [1]

Total: 7

QUESTION 6

(a) Briefly *describe* the **types of queries** that are performed on a data stream. [4]

(b) *What* is the **Greedy** algorithm within the context of web advertising and what is its competitive ratio? [3]

Total: 7

QUESTION 7

(a) *Explain* the roles each five types of layers play in a typical **Convolutional Neural Network (CNN)**. [5]

(b) *Name* two (2) examples of metrics you would use to assess a salient **segmentation** algorithm. [2]

Total: 7

QUESTION 8

(a) Briefly *describe* three **limitations** of collaborative filtering-based recommender systems. [3]

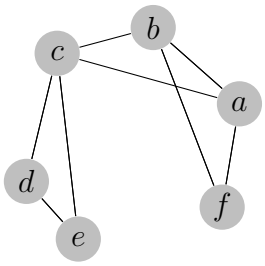
(b) Given the following utility matrix to depict review scores for five different games for five users, calculate the **cosine** distance (angle) between user 2 and 4 and say whether you would recommend the same types of games for user 2 as for user 4: [4]

	A	B	C	D	E
0	4	2	3	1	2
1	4	4	5	5	0
2	1	2	5	3	0
3	4	1	4	3	4
4	5	4	0	3	5

Total: 7

QUESTION 9

Analyse the undirected graph below and answer the questions that follow.



(a) *Where* would you cut the graph to make **good** partitions? [1]

(b) Provide the above graph's **Laplacian** matrix representation. [6]

(c) *Provide* two dependencies or tools that can be used to **visualise** data. [2]

Total: 9

QUESTION 10

The Netflix show "The Social Dilemma" has highlighted some key concerns around how social media website recommendations can manipulate users to perform revenue-conducive behaviour. It has spurred renewed talks about regulation of the companies that provision these social media platforms. However, other parties say the benefit of these social media platforms outweigh these concerns. Write a one and a half page (with pt font size) report that will discuss the implications of using Big Data to build recommender systems within social media platforms, along with the extent of social responsibility these social network providers need to meet and how it can potentially manipulate users. The report should be structured as follows:

- Discuss exactly how you would **implement** a system like this by paying attention to the following: [6]
  - The **infrastructure, architecture, methods and algorithms** that you would use.
  - The **constraints** of your implementation.
- A discussion on the **potential** of using Big Data to achieve socially responsible recommendations in social media, which includes: [9]
  - The **privacy** implications and **benefits** of using Big Data within the context.
  - How **manipulation** can take place within the context.
  - Should regulation be increased **or** decreased within the context?

Total: 15

The End!