FACULTY OF SCIENCE
ACADEMY OF COMPUTER SCIENCE AND SOFTWARE ENGINEERING

MODULE	IT8X030: BIG DATA ANALYTICS
CAMPUS	AUCKLAND PARK CAMPUS (APK)
ASSESSMENT	NOVEMBER EXAM 2021

DATE: 2021-11-08	**SESSION:** 8:30 - 10:30
ASSESOR(S):	PROF D.T. VAN DER HAAR
MODERATOR:	DR S. EYBERS (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. You 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/EyYcw5fTSW).
7. This paper consists of 4 pages excluding the cover page.

QUESTION 1

(a) *What* is **k-fold cross-validation** and *why* would you use it?. [3]

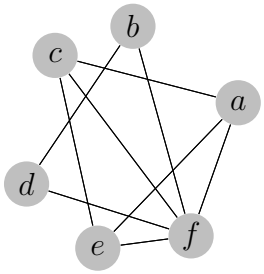
(b) *What* are two (2) **types of decomposition** that can occur in parallel computing. [2]

(c) *Name* two (2) **technologies** that be used to virtualise computing hardware. [2]

Total: 7

QUESTION 2

(a) Provide the below graph's **Laplacian** matrix representation. [4]



(b) *Describe* **online learning** and the role it plays in stream model based algorithms. [3]

Total: 7

QUESTION 3

One aspect of fighting the COVID19 pandemic is having a better understanding of the symptoms experienced when a patient contracts the virus. Some argue that a pattern exists in how certain symptoms present themselves in certain patients. One way of determining if this relationship exists to determine if an association exists using basket analysis. Below is a table representing nine (9) patients (baskets) and seven symptoms (items). A "1" indicates membership of the item in the patient. For example patient 0 exhibited symptoms W, T and U.

	Y	Z	V	X	W	T	U
0	0	0	0	0	1	1	1
1	1	0	1	1	1	1	0
2	1	0	1	0	1	0	0
3	0	1	1	0	1	1	1
4	0	1	1	1	1	0	1
5	1	0	1	0	1	1	0
6	0	1	0	1	1	1	1
7	0	1	0	1	0	1	0
8	1	1	0	1	0	1	1

(a) *Derive* two (2) viable **association rules** with more than one antecedent (in the format $A, B \rightarrow C$) with a support greater than 0.25, which would apply to the above baskets. [2]

(b) *Calculate* the **interest** for the above identified association rules. [4]

(c) *Are* there any interesting association rules (certain symptoms that present themselves uniformly together) for the above baskets? [1]

Total: 7

QUESTION 4

(a) Assuming the *above* baskets for the previous question are boolean-based mappings of items in a basket, what is the **Jaccard** distance between basket (row) **3** and **4**? [2]

(b) *Derive* **word trigrams** for the following string: [3]
"You take the blue pill the story ends you wake up in your bed and believe whatever you want to believe"

(c) Briefly *describe* **hierarchical agglomerative clustering**. [2]

Total: 7

QUESTION 5

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

$$\begin{bmatrix} 5 & 4 & 3 & 4 & 5 \\ 5 & 7 & 6 & 3 & 3 \\ 5 & 7 & 2 & 3 & 7 \\ 8 & 4 & 8 & 0 & 5 \\ 5 & 4 & 1 & 2 & 6 \end{bmatrix}$$

(b) *Name* one **non-linear** dimensional reduction method. [1]

Total: 7

QUESTION 6

(a) *List* Pedro Domingo's five (5) **paradigms of machine learning**. [5]

(b) *How* is **multi-class classification** achieved with SVM's? [2]

Total: 7

QUESTION 7

(a) *What* is a **loss function** and how does it *relate* to deep learning? [3]

(b) Apply the following structuring element (kernel): $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ [4]

to the following image matrix (assuming a stride of 1 is used with no padding):

$$\begin{bmatrix} 6 & 8 & 10 \\ 12 & 12 & 0 \\ 4 & 15 & 10 \end{bmatrix}$$

Total: 7

QUESTION 8

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

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

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

(c) *Name* and *describe* one (1) **metric** that can be used to assess recommender systems. [2]

Total: 9

QUESTION 9

(a) *Discuss* the **online graph matching** problem. [5]

(b) *Provide* two **considerations** one should keep in mind when **visualising** data. [2]

Total: 7

QUESTION 10

Sasria just announced that claims linked to damage caused by the local July unrest could amount to between R20 billion and R25 billion, thereby surpassing similar events of unrest internationally. It also has a significant impact on the economy and subsequently unemployment, which makes it an unfavourable outcome for government. One idea that came out from all of this is to establish a command center that monitors coordination efforts on social media and attempts to prevent instigators from acting through the use of big data analytics. Write a one and a half page report (with a minimum of 500 words with 12 pt font size and single spacing) that will discuss the feasibility, methods and implications of using Big Data to monitor for potentially dangerous coordination efforts within social media platforms. In the report you must discuss exactly how you would **implement** a system like this by paying attention to the following:

1. The **infrastructure** that you would use.
2. What **data modalities** you would leverage.
3. Outline the **data preparation** approaches that would be applicable
4. Relevant **methods or algorithms** that you would consider.
5. The **benefits and constraints** of your implementation.

Total: 15

The End!