FACULTY OF SCIENCE
ACADEMY OF COMPUTER SCIENCE AND SOFTWARE ENGINEERING

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

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 11 pages excluding the cover page.

QUESTION 1

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

Solution:
The data is split into k representative folds to train and test a machine learning model. It is used to asses a machine learning model (especially when the dataset is small).

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

Solution:

1. It can be based on data and the task
2. or based on static or dynamic processes

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

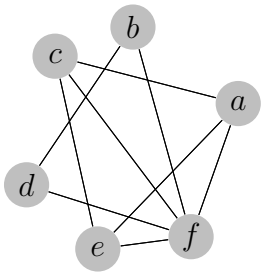
Solution:

1. Docker
2. VirtualBox
3. VmWare
4. any technology really...

Total: 7

QUESTION 2

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



Solution:
L = D - A
Degree Matrix [1 marks]

$$\begin{bmatrix} 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 5 \end{bmatrix}$$

Adjacency matrix (assuming a,b,c,d,e,f order)[1 marks]

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

Laplacian matrix [2 marks]

$$\begin{bmatrix} 3 & 0 & -1 & 0 & -1 & -1 \\ 0 & 2 & 0 & -1 & 0 & -1 \\ -1 & 0 & 3 & 0 & -1 & -1 \\ 0 & -1 & 0 & 2 & 0 & -1 \\ -1 & 0 & -1 & 0 & 3 & -1 \\ -1 & -1 & -1 & -1 & -1 & 5 \end{bmatrix}$$

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

Solution:

Online learning trains a model incrementally as new data comes in to slowly adapt to changes that occur in the data (the very nature of data streams). It allows for modeling problems where we have a continuous stream of data

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]

Solution:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	interest	antecedent_len
0	(Y, W)	(V)	0.333333	0.555556	0.333333	1.0	1.8	0.148148	inf	0.444444	2
2	(U, X)	(Z)	0.333333	0.555556	0.333333	1.0	1.8	0.148148	inf	0.444444	2
4	(Z, W)	(U)	0.333333	0.555556	0.333333	1.0	1.8	0.148148	inf	0.444444	2
1	(V)	(Y, W)	0.555556	0.333333	0.333333	0.6	1.8	0.148148	1.666667	0.266667	1
3	(Z)	(U, X)	0.555556	0.333333	0.333333	0.6	1.8	0.148148	1.666667	0.266667	1
5	(U)	(Z, W)	0.555556	0.333333	0.333333	0.6	1.8	0.148148	1.666667	0.266667	1

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

Solution:

See above table, but also include:

```
U,T->W (Support: 0.33, Interest: -0.0278) and
Z,X->T (Support: 0.333, Interest: 0.0833) are fine too
```

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

Solution:

No there are not any

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]

Solution:
$d(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$ so
1 - $\frac{4}{6}$ (0.67) which equals
$\frac{2}{6} = 0.33$

(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"

Solution:
Where _ depicts white space

- 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]

Solution:

- Bottom up clustering
- Initially, each point is a cluster
- Repeatedly combine the two "nearest" clusters into one

Total: 7

QUESTION 5

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

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

Solution:

$$\begin{bmatrix} 5 & 4 & 3 & 4 & 5 \\ 5 & 7 & 6 & 3 & 3 \\ 5 & 7 & 2 & 3 & 7 \\ 8 & 4 & 8 & 0 & 5 \\ 5 & 4 & 1 & 2 & 6 \end{bmatrix} \begin{bmatrix} 5 & 5 & 5 & 8 & 5 \\ 4 & 7 & 7 & 4 & 4 \\ 3 & 6 & 2 & 8 & 1 \\ 4 & 3 & 3 & 0 & 2 \\ 5 & 3 & 7 & 5 & 6 \end{bmatrix} = \begin{bmatrix} 91 & 98 & 106 & 105 & 82 \\ 98 & 128 & 116 & 131 & 83 \\ 106 & 116 & 136 & 119 & 103 \\ 105 & 131 & 119 & 169 & 94 \\ 82 & 83 & 103 & 94 & 82 \end{bmatrix}$$

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

Solution:
Any methods, but some options include:

- Isomap
- Auoencoder
- T-SNE
- Deep learning variants
- etc.

Total: 7

QUESTION 6

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

Solution:

1. Rule based learning (Decision trees, Random Forests, etc)
2. Connectivism (neural networks, etc)
3. Bayesian (Naive Bayes, Bayesian Networks, Probabilistic Graphical Models)
4. Analogy (KNN & SVMs)
5. Unsupervised Learning (Clustering, dimensionality reduction, etc)

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

Solution:
Using:

- One versus one
- One versus many

,

Total: 7

QUESTION 7

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

Solution:
A loss function is a measure to determine the effectiveness of the network and is the objective of training. During training we try to minimise the loss using optimisation techniques.

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

Solution:
$\begin{bmatrix} 20. & 22. \\ 16. & 15. \end{bmatrix}$

Total: 7

QUESTION 8

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

Solution:
Any two (2) of the following

1. Cold Start: Need enough users in the system to find a match
2. Sparsity: The user ratings matrix is sparse and it is hard to find users that have rated the same items
3. First rater: Cannot recommend an item that has not been previously rated New items, Esoteric items
4. Popularity bias: Cannot recommend items to someone with unique taste. Tends to recommend popular items

(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

Solution:
The Cosine distance between row 2 and 4 is: 1 - 0.813115628 = 0.18688437181825834, no you would not

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

Solution:

- Root-mean-square error (RMSE) $\sqrt{\Sigma_{xi}(r_{xi} - r^*_{xi})^2}$
  where r_{xi} is predicted, r^*_{xi} is the true rating of x on i
- Precision at top 10: % of those in top 10 item Rank Correlation: Spearman's correlation between system's and user's complete rankings

Total: 9

QUESTION 9

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

Solution:
I like to use an analogy:

- Initially, we are given the set boys
- In each round, one girl's choices are revealed (That is, the girl's edges are revealed)
- At that time, we have to decide to either:
  - Pair the girl with a boy
  - Do not pair the girl with any boy
- Example of application: Assigning tasks to server

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

Solution:

- Know your audience
- Think about the content (relationships, time frame, compositions, comparisons)
- Colours matter
- Use interactive maps
- Use the tools out there
- Build a flexible, unconstrained query engine

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.

Solution:

An appropriate discussion that highlights the automation efforts related to identifying potential instigators (using sentiment analysis, natural language processing, idealogical triggers), computing infrastructure (cloud offering), data modalities (text, image and video), data prep (missing values, normalisation), methods (encoding, dimensional reduction, clustering, etc), advantages (decision making ability, cost saving, risk mitigation, etc.) and limitations (privacy, resource requirements, etc.) of big data within the context.

Total: 15

The End!