FACULTY OF SCIENCE
ACADEMY OF COMPUTER SCIENCE AND SOFTWARE ENGINEERING

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

DATE: 2021-01	**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 10 pages excluding the cover page.

QUESTION 1

(a) Name four **challenges** one would typically face when dealing with data. Provide a one-sentence description of each challenge. [4]

Solution:

1. Usage
2. Quality
3. Context
4. Streaming
5. Scalability

(b) Describe the **MapReduce** programming model in terms of: the hardware architecture that supports MapReduce computations, the primary steps that are used to run a MapReduce job, and the software components that must be implemented by application programmers. [3]

Solution:
MapReduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster. A MapReduce program is composed of a map procedure, which performs filtering and sorting, and a reduce method, which performs a summary operation

1. Sequentially read a lot of data
2. Map: Extract something you care about
3. Group by key: Sort and shuffle
4. Reduce: Aggregate, summarise, filter or transform
5. Write the result

Total: 7

QUESTION 2

As we navigate the COVID19 pandemic there have been key research collaborations that have lead to significant breakthroughs by certain authors and it is important to identify the key relationships that lead to a key author with a valuable discovery. One way to achieve this is through the use of basket analysis on journal publications and the publication authors within them. Below is a table representing nine (9) journal volumes (baskets) and six authors (items). An "1" indicates membership of the item in the transaction.

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

Solution:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	interest	antecedent_len
8	(B, F)	(A)	0.444444	0.555556	0.333333	0.750000	1.35	0.08642	1.777778	0.194444	2
7	(B, A)	(F)	0.555556	0.444444	0.333333	0.600000	1.35	0.08642	1.388889	0.155556	2
9	(A, F)	(B)	0.333333	1.000000	0.333333	1.000000	1.00	0.00000	inf	0.000000	2
3	(F)	(A)	0.444444	0.555556	0.333333	0.750000	1.35	0.08642	1.777778	0.194444	1
11	(F)	(B, A)	0.444444	0.555556	0.333333	0.750000	1.35	0.08642	1.777778	0.194444	1
2	(A)	(F)	0.555556	0.444444	0.333333	0.600000	1.35	0.08642	1.388889	0.155556	1
10	(A)	(B, F)	0.555556	0.444444	0.333333	0.600000	1.35	0.08642	1.388889	0.155556	1
0	(B)	(A)	1.000000	0.555556	0.555556	0.555556	1.00	0.00000	1.000000	0.000000	1
1	(A)	(B)	0.555556	1.000000	0.555556	1.000000	1.00	0.00000	inf	0.000000	1
4	(C)	(B)	0.444444	1.000000	0.444444	1.000000	1.00	0.00000	inf	0.000000	1
5	(D)	(B)	0.444444	1.000000	0.444444	1.000000	1.00	0.00000	inf	0.000000	1
6	(F)	(B)	0.444444	1.000000	0.444444	1.000000	1.00	0.00000	inf	0.000000	1

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

Solution:
See above table

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

Solution:
No there are not (the highest is $B, F \rightarrow A$)

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]

Solution:
$d(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$ so
1 - $\frac{2}{4}$ which equals
$\frac{2}{4} = 0.5$

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

Solution:
Where _ depicts white space

- The
- _ca
- t_s
- at_
- on_
- the
- _ma
- t

(c) Briefly describe how the **A-Priori** algorithm can be used to solve a similarity problem. [2]

Solution:
The first pass is used to find frequent singletons, but the second pass is used to count only candidate pairs to avoid searching for items that are not similar.

Total: 7

QUESTION 4

(a) *Describe* the point assignment **clustering** strategy. [2]

Solution:

1. Maintain a set of clusters.
2. Place points into their "nearest" cluster

Points are considered in some order, and each one is assigned to the cluster into which it best fits. This process is normally preceded by a short phase in which initial clusters are estimated. Variations allow occasional combining or splitting of clusters, or may allow points to be unassigned if they are outliers (points too far from any of the current clusters).

(b) *Discuss* how the **k-means** algorithm works. [5]

Solution:

Assumes Euclidean space/distance. Start by picking k, the number of clusters. Initialize clusters by picking one point per cluster. Example: Pick one point at random, then k 1 other points, each as far away as possible from the previous points

1. For each point, place it in the cluster whose current centroid it is nearest
2. After all points are assigned, update the locations of centroids of the k clusters
3. Reassign all points to their closest centroid (Sometimes moves points between clusters)

Repeat 2 and 3 until convergence. Convergence: Points don't move between clusters and centroids stabilize

Total: 7

QUESTION 5

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

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

Solution:

$$\begin{bmatrix} 1 & 5 & 8 & 7 & 8 \\ 3 & 7 & 3 & 4 & 7 \\ 0 & 2 & 4 & 0 & 5 \\ 5 & 1 & 2 & 2 & 8 \\ 6 & 2 & 4 & 7 & 2 \end{bmatrix} \begin{bmatrix} 1 & 3 & 0 & 5 & 6 \\ 5 & 7 & 2 & 1 & 2 \\ 8 & 3 & 4 & 2 & 4 \\ 7 & 4 & 0 & 2 & 7 \\ 8 & 7 & 5 & 8 & 2 \end{bmatrix} = \begin{bmatrix} 203 & 146 & 82 & 104 & 113 \\ 146 & 132 & 61 & 92 & 86 \\ 82 & 61 & 45 & 50 & 30 \\ 104 & 92 & 50 & 98 & 70 \\ 113 & 86 & 30 & 70 & 109 \end{bmatrix}$$

(b) Define **Mahalanobis** distance. [1]

Solution:
The Mahalanobis distance is a measure of the distance between a point P and a distribution D. It is a multi-dimensional generalization of the idea of measuring how many standard deviations away P is from the mean of D.

Total: 7

QUESTION 6

(a) Briefly *describe* the **BALANCE** algorithm within the context of web advertising. [3]

Solution:

1. By Mehta, Saberi, Vazirani and Vazirani
2. BALANCE solves this problem
3. For each query, pick the advertiser with the largest unspent budget
4. Break ties arbitrarily (but in a deterministic way)

(b) *Describe* the steps in the **Girvan-Newman** algorithm. [4]

Solution:

- Undirected unweighted networks
- Repeat until no edges are left:
  - Calculate betweenness of edges
  - Remove edges with highest betweenness
- Connected components are communities
- Gives a hierarchical decomposition of the network

.

Total: 7

QUESTION 7

(a) *List* three examples of **Convolutional Neural Networks (CNN)** architectures. [3]

Solution:

1. LENET-5
2. VGG16/19
3. Inception /v2/v3/etc.
4. Alexnet
5. Mobilenet
6. etc.

(b) *What* are the roles of **Dropout** layers in a deep neural network? [2]

Solution:
Regularization for minimising overfitting

(c) *Name* two (2) examples of **metrics** you would use to assess a deep learning model in the training phase. Provide a brief definition of each measure. [2]

Solution:

1. Accuracy
2. loss

Total: 7

QUESTION 8

(a) Briefly *describe* three **advantages** of content-based recommender systems. [3]

Solution:
Any two (2) of the following

1. No need for data on other users (No cold-start or sparsity problems)
2. Able to recommend to users with unique tastes
3. Able to recommend new & unpopular items (No first-rater problem)
4. Able to provide explanations (Can provide explanations of recommended items by listing content-features that caused an item to be recommended)

(b) The following utility matrix depicts review scores for five different games for five users. Calculate the **cosine** distance (angle) between user 2 and 4. State whether you would recommend the same types of games for user 2 and user 4. Give reasons for your answer. [4]

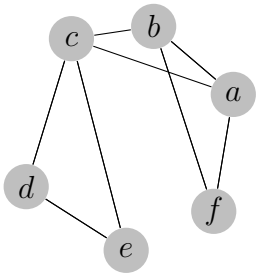
	A	B	C	D	E
0	2	5	1	4	5
1	4	1	3	3	4
2	3	0	5	5	3
3	5	1	2	1	4
4	1	2	3	2	0

Solution:
The Cosine distance between row 2 and 4 is: 0.19967326933495888, no you would not

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? Give reasons for your answer. [1]

Solution:
A c-b or a-c cut

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

Solution:
L = D - A
Adjacency matrix (assuming a,b,c,d,e,f order)[2 marks]

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

Degree Matrix [2 marks]

$$\begin{bmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$

Laplacian matrix [2 marks]

$$\begin{bmatrix} 3 & -1 & 0 & 0 & 0 & -1 \\ -1 & 3 & 0 & 0 & 0 & -1 \\ -1 & -1 & 4 & -1 & -1 & 0 \\ 0 & 0 & 1 & 2 & -1 & 0 \\ 0 & 0 & -1 & -1 & 2 & 0 \\ -1 & -1 & 0 & 0 & 0 & 2 \end{bmatrix}$$

(c) *What* does the **Flask** Python package do? [2]

Solution:
It helps withmaking lightweight REST API's

Total: 9

QUESTION 10

As we transition to the new normal of working from home we are faced with a fresh surge of cyber security attacks and budget cuts. Some analysts are saying this is the cue for putting the brakes on for big data systems because big data systems tend to make businesses softer targets and they cost a great deal of money. Write a one and half page report to discuss the implications of using Big Data to build valuable systems in business. **In your report, you should specify** exactly how Big Data impacts cyber security and cost:

- A discussion on the **potential** of using Big Data to achieve business competitiveness, contrasted against cyber security and cost:
  - The **advantages** and **limitations** of using Big Data within business.
  - The **infrastructure, architecture, methods and algorithms** that you would use.
  - The **cyber security** implications of using Big Data within the context.
  - The **cost** implications of using Big Data within the context.
  - Your **opinion** on whether this economic climate is good for the introduction of new big data systems.

Solution:

An appropriate discussion that highlights the advantages and limitations (privacy, resource requirements, etc.) of big data within the context.

Total: 15

The End!