CT102 Information Systems
Assignment 2
Date: Friday 10th December 2021
Due: on or before Monday 10th January 2022 via Blackboard
Total Marks: 20 (5 per question)

Q. 1. *(modified from summer 2019)*
Given the following four symbols, and their probability of occurrence, use the compression technique of arithmetic coding to find the real-valued interval that corresponds to the coding of the message **set**.
Please ensure to show all your workings. Note that the symbol order given (s, t, a, e) should be used and the interval should begin at 0.0.

s: P(s) = 0.15 **t:** P(t) = 0.2 **a**: P(a) = 0.3 **e**: P(e) = 0.35

Clearly explain your approach and the steps taken, showing all your workings.

Q. 2. *(summer 2021)*
Given the following five symbols, and their probabilities of occurrence, use the compression technique of Huffman encoding to find the encoding of the word **water**

P(**w**) = 0.05 P(**r**) = 0.1 P(**a**) = 0.22 P(**t**) = 0.26 P(**e**) = 0.37

Clearly explain your approach and the steps taken, showing all your workings.

***** For Q1 and Q2 please ensure to follow algorithm **exactly** as given in lecture notes and as shown in the worked examples, paying particular attention to the ordering of the symbols.

Q. 3. *(Summer 2021)*
The following table summarises the ratings that people (with names given) have given to music artists (with artist's name given) where ratings are in the range [1-5] and 1 indicates "dislike" and 5 indicates "like".

	Lewis Capaldi	Drake	Dermot Kennedy	Ariana Grande	Lisa O'Neill	Dua Lipa
Brian	4		5	2	3	
Luke	2	5	3	4		
Anna		1	4		5	4
Sonya	5		5	3		
Hugo		5		4		1
Linda		4		5	2	

(a) Show how the Pearson Correlation formula can be used to find the correlation between the two people **Hugo** and **Luke**. Clearly show the formula used and show, and explain your workings, in your answer.
(b) Explain how the result you obtain could be used in a recommendation system.

Q. 4. *(Summer 2021)*
Given the following edge list that represents the edges that exist between six nodes, where (A, B) represents the fact that there is a directed edge from node A to node B.

(A, B), (A, C), (B, A), (B, D), (B, F), (C, A), (C, B), (C, E), (D, E), (D, F),
(E, A), (E, B), (E, C), (E, F), (F, B), (F, D).

Assuming the given network represents a social network of six people:

(a) Distinguish between the *average outdegree* and the *edge density* of the network by calculating the *average outdegree* and the *edge density* of the given social network, explaining the formulae and showing your workings.
(b) Calculate the *local clustering coefficient* value of the node **C** explaining your formula and showing your workings.
(c) Find and list the shortest path between the two nodes **A** and **F**, explaining your approach.

5. ***** Please include the ticked plagiarism declaration in your solution or the following:

Plagiarism Declaration:
"I am aware of what plagiarism is and include this here to confirm that this work is my own"

Please note that any suspected cases of plagiarism, or absence of a plagiarism declaration, will not receive a mark until assurances can be given in person as to the origins of the solution.