**CT102 Information Systems**
**Assignment 2**
**Date:** Monday 20th November 2023
**Due: on or before** Monday 15th January 2024 via Canvas
**Total Marks:** 15

**Q. 1.** *(modified from Summer 2023)*

Given the following six symbols, and their probabilities of occurrence:

P(**t**) = 0.35 P(**a**) = 0.25 P(**h**) = 0.2 P(**c**) = 0.1 P(**p**) = 0.05 P(**g**) = 0.05

(i) Use the compression technique of **Huffman encoding** to find the encoding of the word "**chatgpt**". Clearly explain your approach and the steps taken.
*(3 marks)*

(ii) Using the same symbols, and probabilities, use the compression technique of **Arithmetic encoding** to find the interval corresponding to the encoding of the word **"chat".**
*(3 marks)*

***** For parts (i) and (ii) 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. 2.** *(modified from Summer 2023)*

The table summarises the preference data associated with four people (with their names given) and TV series (with the TV series name given) they have watched (where values are in the range [1-5]).

| | The Empress | Cable Girls | Dead Wind | High Water | Baptiste |
|---|---|---|---|---|---|
| Keenan | | 4 | 3 | | 1 |
| Ruth | 3 | 5 | 3 | | 2 |
| Tim | 2 | | 4 | 5 | |
| Siobhán | 2 | 5 | 3 | 4 | |

(i) Using the preference data given by users for TV series, use the Pearson correlation formula to find the correlation between **Keenan** and **Ruth**. Show and explain the formula used and your workings. *(2 marks)*

(ii) Briefly, in your own words, explain how the result from part (i) could be used for recommendation. *(1 mark)*

**Q. 3.** *(modified from Summer 2023)*

The following table *timetable* holds details of the lecture times, locations and lecturers for modules. Each one hour timetable slot has a unique ID (ID), and associated module code (Code), module name (Name), lecturer name (Lecturer), and the semester the module is taught (Sem). Each one hour timetable slot has also an associated day (Day), time (Time) and venue (Venue). Each venue has a maximum capacity (Cap) and a building (Building).

| ID | Code | Name | Lecturer | Sem | Day | Time | Venue | Cap | Building |
|---|---|---|---|---|---|---|---|---|---|
| 1 | CT1114 | Web Dev | S Redfern | 2 | Mon | 10:00 | IT106 | 70 | IT Building |
| 2 | CT1114 | Web Dev | S Redfern | 2 | Mon | 11:00 | IT106 | 70 | IT Building |
| 3 | MA160 | Maths | M Hayes | 2 | Mon | 13:00 | AC202 | 60 | Arts-Sci |
| 4 | MA190 | Maths | G Pfeiffer | 2 | Mon | 13:00 | IT250 | 250 | IT Building |
| 5 | CT101 | Comp Sys | B Chakravarthi | 2 | Mon | 15:00 | IT125G | 125 | IT Building |
| 6 | CT101 | Comp Sys | B Chakravarthi | 2 | Mon | 16:00 | IT125G | 125 | IT Building |
| 7 | MA160 | Maths | M Hayes | 2 | Tues | 10:00 | AC202 | 60 | Arts-Sci |
| 8 | MA190 | Maths | G Pfeiffer | 2 | Tues | 10:00 | IT250 | 250 | IT Building |
| 9 | CT102 | Algorithms | J Griffith | 2 | Tues | 11:00 | IT125G | 125 | IT Building |
| 10 | CT102 | Algorithms | J Griffith | 2 | Tues | 12:00 | IT125G | 125 | IT Building |

(i) Choose a suitable Primary Key for the *timetable* table, explaining your choice. *(1 mark)*

(ii) Identify the redundant data stored in the *timetable* table. *(1 mark)*

(iii) Write an SQL query to find the venues (Venue and Building) that have capacity greater than 60. *(1 mark)*

**Q. 4.** *(modified from Summer 2022)*

Given the following edge list that represents the edges that exist in a network between five nodes, where (A, C) represents the fact that there is a directed edge from node A to node C.

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

Assuming the given network represents a social network of five people: distinguish between the outdegree and the clustering coefficient of a node by calculating the outdegree and the clustering coefficient of the node **C**. Clearly show the formulae used and show, and explain, your workings in your answer. *(3 marks)*

**PTO**

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

**Declaration:**

**"I am aware of what plagiarism is and include this to confirm that this work is my own and, further, I confirm that this work was not, wholly or in part, produced by generative AI tools"**

Please note that any suspected cases of plagiarism, or, of the use of generative AI tools will **not** receive a mark until assurances can be given in person as to the origins of the solution. Submissions will not be corrected if this declaration is absent.