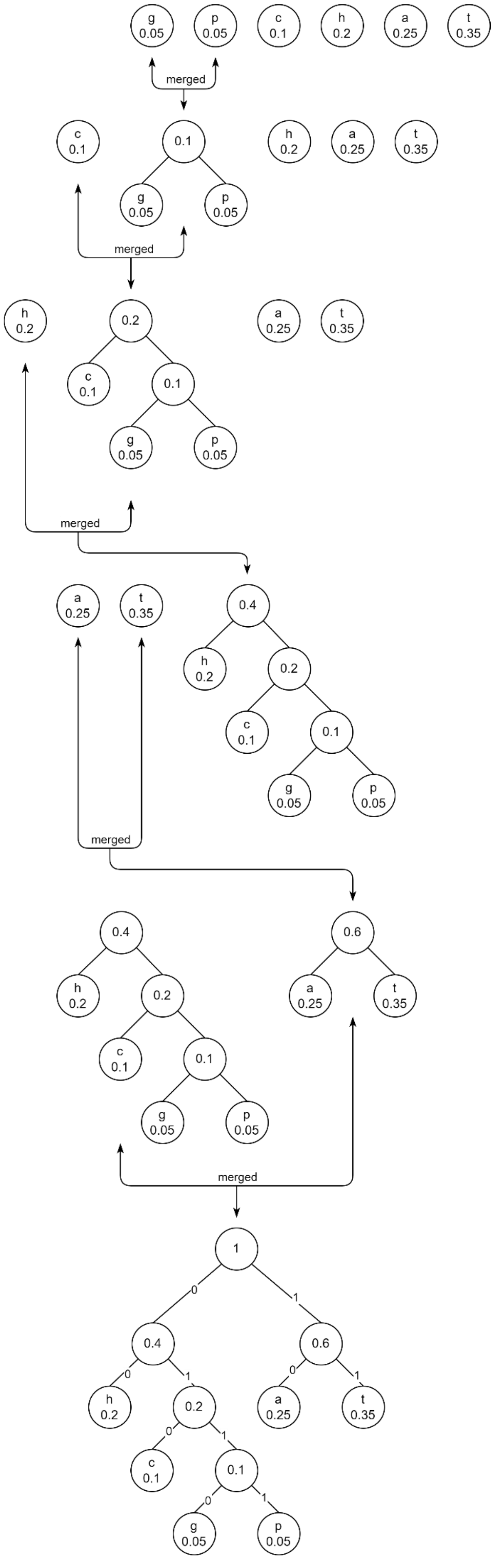


Q. 1. (modified from Summer 2023) Given the following six symbols, and their probabilities of occurrence: P(t) = 0.35 P(a) = 0.25 P(h) = 0.2 P(c) = 0.1 P(p) = 0.05 P(g) = 0.05

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

$$c: 010$$

$$h: 00$$

$$a: 10$$

$$t: 11$$

$$g: 0111$$

$$p: 0110$$

$$t: 11$$

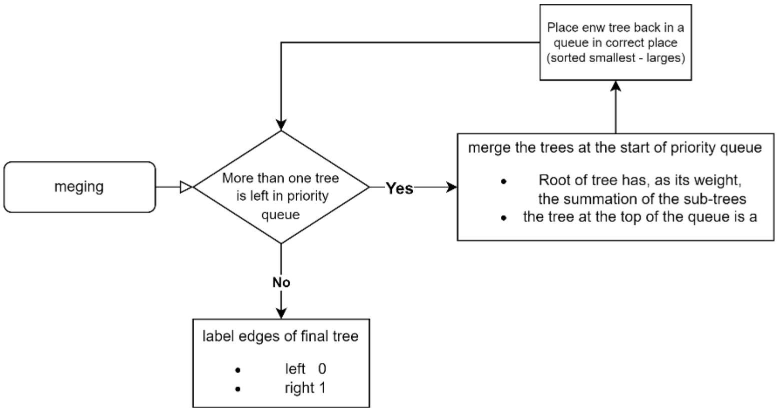
$$chatgpt : 0100010110111011011$$

Explanation:

To encode word ***chatgpt*** I used Huffman compression algorithm given in the lectures.

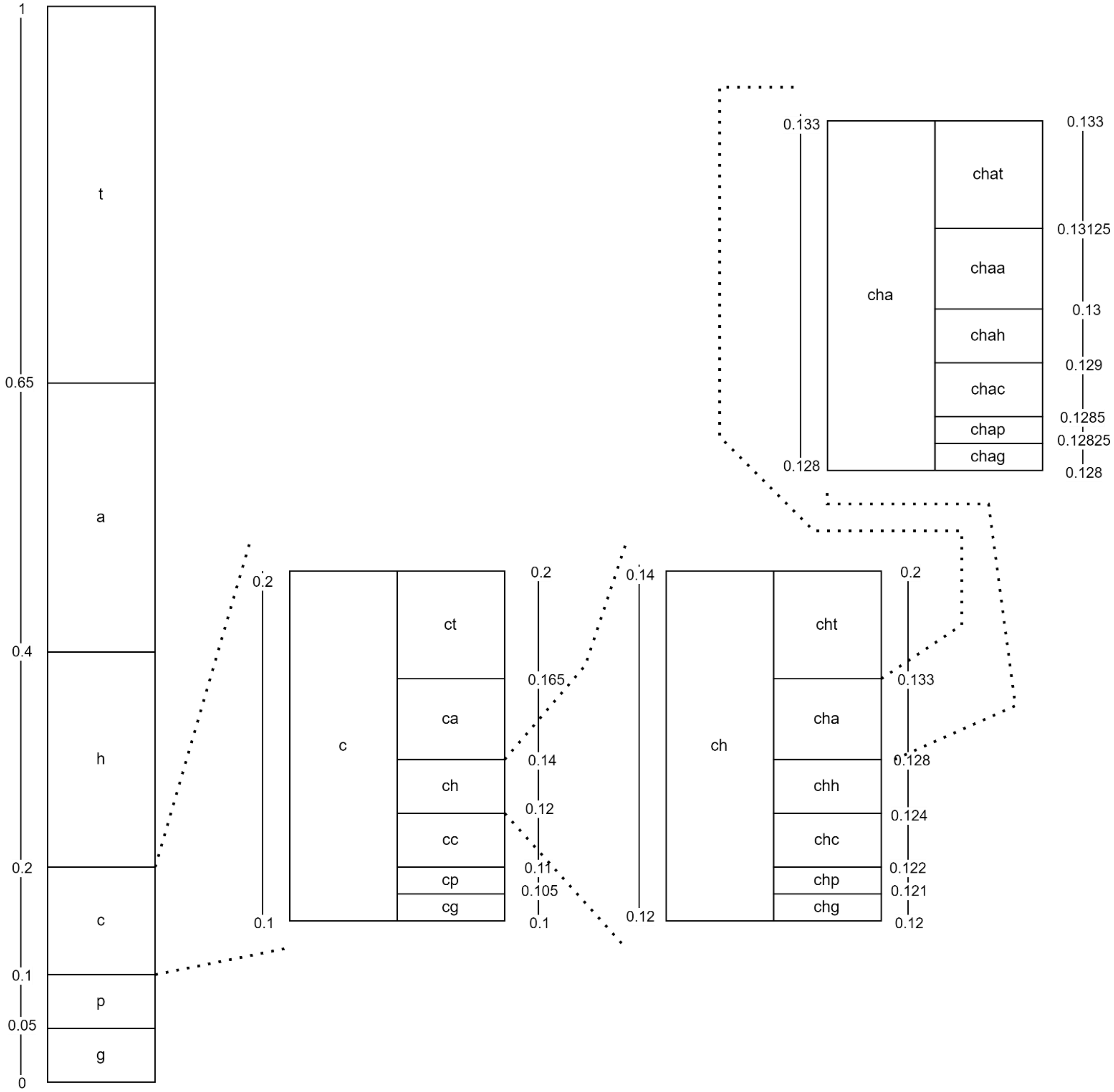
First step was to create a trivial tree(node) for each letter with an associated weight and sort them in non-decreasing order (with exception when the weights are equal, then sort them alphabetically or by the tree size (smallest to largest)

Then I started merging while following this rules:


Place enw tree back in a queue in correct place (sorted smallest - larges)
meging
More than one tree is left in priority queue
Yes
merge the trees at the start of priority queue
• Root of tree has, as its weight, the summation of the sub-trees
• the tree at the top of the queue is a
No
label edges of final tree
• left 0
• right 1

Text in schema above is taken from the presentation to accurately explain steps of the merging part of the algorithm

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”.



chat: [0.13125; 0.133)

Q. 2 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	Baptise
Keenan	0	**4**	**3**	0	**1**
Ruth	3	**5**	**3**	0	**2**
Tim	2	0	4	5	0
Siobhan	2	5	3	4	0

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.

ii)

Pearson Correlation formula is a to find a correlation between "neighbours". The result is a weighted average od deviation from the neighbours'' mean

Pearson Correlation formula

$$\omega_{a,u} = \frac{\sum_{i=0}^{m}\left(r_{a,i} - \overline{r_a}\right) \times \left(r_{u,i} - \overline{r_u}\right)}{\sqrt{\sum_{i=1}^{m}\left(r_{u,i} - \overline{r_{au}}\right)^2} \times \sqrt{\sum_{i=1}^{m}\left(r_{a,i} - \overline{r_a}\right)^2}}$$

$r_{a,i}$ *is rating of user a for item i*
$\overline{r_a}$ *is the average rating given by user a*
$r_{u,i}$ *is rating of user u for item i*
$\overline{r_u}$ *is the average rating given by user u*

Calculate average rating for Keenan and Ruth

$$\overline{r_K} = \frac{4+3+1}{3} \approx 2.67$$

$$\overline{r_R} = \frac{3+5+3+2}{4} = 3.25$$

$$\omega_{K,R} = \frac{(4-2.67)\times(5-3.25)+(3-2.67)\times(3-3.25)+(1-2.67)\times(2-3.25)}{\sqrt{(4-2.67)^2+(3-2.67)^2+(1-2.67)^2} \times \sqrt{(5-3.25)^2+(3-3.25)^2+(2-3.25)^2}} =$$

$$= \frac{4.3325}{2.16 \times 2.17} \approx 0.92$$

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

Results from part I can be used for recommendation, a high positive correlation could suggest that both user in our case Keenan and Ruth have similar preferences. Similarity high negative correlation would suggest that both users have different preferences.

Q.3. 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.

Suitable Primary Key would **ID column**, because it is contains unique values that can represent each row in a table effectively. Individual and non-repeating values are essential for a primary key for effectively identifying each row in a database

ii) Identify the redundant data stored in the timetable table.

Redundant data are Lecturer entries. Each **Lecturer** is assigned for specific **Code**

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

```
SELECT Venue, Building FROM Timetable WHERE Cap > 60;
```

Q.4. 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),

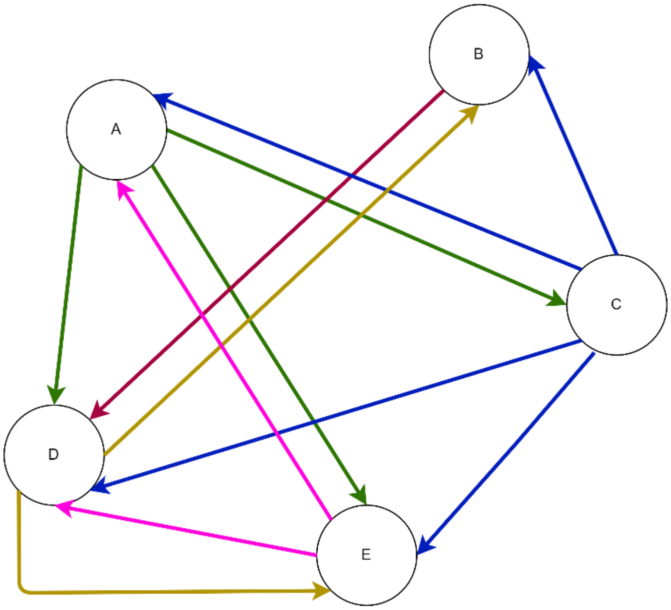
(B, D),

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

(D, B), (D, E),

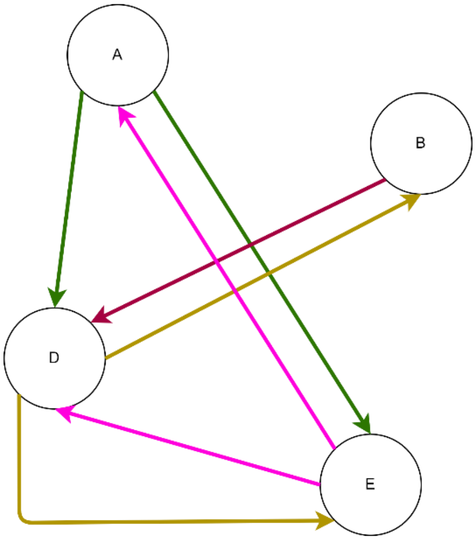
(E, A), (E, D).

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.



	A	B	C	D	E	out degree
A			1	1	1	**3**
B				1		**1**
C	1	1		1	1	**4**
D		1			1	**2**
E	1			1		**2**
In degree	**2**	**2**	**1**	**4**	**3**	$\frac{12}{5} = 2.4$

Clustering coefficient of C



C has 4 neighbours: A, B, D, E

$$CC(C) = \frac{Number\ of\ edges\ between\ neighbors\ of\ C}{Maximum\ possible\ edges\ between\ neighbors\ of\ C} = \frac{7}{4 \times 3} =$$

$$= \frac{7}{12} = 0.58$$

Outdegree of node C is 4 and clustering coefficient of C is equal to 0.58