CS422 Final Exam

Part I - Short Answers

Q1.

Core point is a point that has at least a specified number of points (MinPts) within Eps. These are points that are at the interior of a cluster. Counts the point itself.

A border point is not a core point, but is in the neighborhood of a core point.

A noise point is any point that is not a core point or a border point.

Thus by definition, a noise point cannot exist inside of the neighborhood of a core point.

Q3.

The metric/cost function which is used to measure how close the estimated product matrix P is to the actual values in M is root-mean-square error (RMSE)

When M contains blank elements, the blank elements are ignored i.e. we don't take these elements in our RMSE calculation.

Q4.

$N = 1000000$

1% documents contain shingle 'soccer'

$\therefore N = 1000000 \times 0.01 = 10000$

Out of '1000' document length, 'soccer' appears '17' times.

$$\text{Inverse Document Frequency} = \log_{10}\left(\frac{1000000}{10000}\right) = \underline{2}$$

$$\text{Term Frequency} = 17/1000 = \underline{0.017}$$

$$\text{Score of TF-IDF} = 0.017 \times 2$$

$$= \underline{0.034}$$

Q2.

If an anomaly 'xi' data point is removed from the sample, the log-likelihood of sample distribution increases. The variance parameters remain unaffected by each other incase the features are orthogonal.

Q5.

The dimension of matrix would be n * n

the dimension of matrix would be $n*n$.
The sum of each column would be 1.
Pagerank for the k nodes would converge to $1/k$ and 0 for the $n-k$ nodes, without teleport/skip.

Q6.

Given $N = 10{,}000{,}000$

Minhash characteristic matrix M has a dimension of $27^k * N = 27^5 * 10{,}000{,}000$

Signature matrix S created using permutation
$= 50 * 10{,}000{,}000$

$P(C_1 \cong C_2$, in one of the bands$) = 0.8^5 = \underline{0.328}$

$P(C_1 \ncong C_2, \forall$ bands$) = (1 - 0.328)^{10} = \underline{0.01878}$

The probability that document are similar
$= 1 - 0.01878 = \underline{0.98122}$.

Q7.

Minimum distance between clusters:

$$A, C = |3 - (-2)| = \boxed{5}$$
$$B, C = |17 - (-2)| = 19$$
$$A, B = |9 - 17| = 8$$

Single Linkage : $\underline{A \,\&\, C.}$

Maximum distance b/w clusters:

$$A, C = |9 - (-8)| = 17$$
$$B, C = |19 - (-8)| = 27$$
$$A, B = |3 - 19| = \boxed{16}$$

Complete linkage : $\underline{A \,\&\, B}$

Q8.

SSE can be decreased in 'loose' clusters by following methods:

1) splitting clusters
2) Increasing K (introduce new centroids)

SSE can be increased in 'close' clusters by following methods:

1) combining clusters
2) decreasing K

Part II - Long Answers

Q1.

Let average distance from x_i to points in C be 'a'.
Let average distance from x_i to points in D be 'b'.

then silhouette coef. is given as,

$$S = (b-a) / \max(a,b)$$

Then if S is negative then $a > b$ i.e. the sample is closer to cluster D.

Q2.

Given : mean $(\bar{x}) = 100$
variance $(\sigma^2) = 10000$

standard deviation $(\sigma) = \sqrt{\sigma^2}$
$= \sqrt{10000} = 100$

4 - standard deviation above $\bar{x} = 100 + 4(100)$
$= 500$

4 - standard deviation below $\bar{x} = 100 - 4(100)$
$= -300$

Q3.

Given: User 1 = [4, 2, 3, 2, 4]
User 2 = [5, 3, 4, 3, 5]

Average for User 1 = 3
Average for User 2 = 4

Centered value for user 1 : [1, -1, 0, -1, 1]
Centered value for user 2 : [1, -1, 0, -1, 1]

$$\text{Cosine similarity} = \frac{a \cdot b}{|a| \cdot |b|} = \frac{1+1+0+1+1}{\sqrt{4} \cdot \sqrt{4}}$$

$$= \frac{4}{4} = 1$$

Collaborative filtering algorithm will include the user 2 since cosine similarity for User 1 and User 2 is 100% which is more than threshold value of 75%

of 15%.

Q4.

Total movies, M = 100 ; where 20 movies belong to each of the 5 given genres.

weight per genre:

family = 20/100 = 0.2

animation = 20/100 = 0.2

adventure = 20/100 = 0.2

drama = 20/100 = 0.2

documentary = 20/100 = 0.2

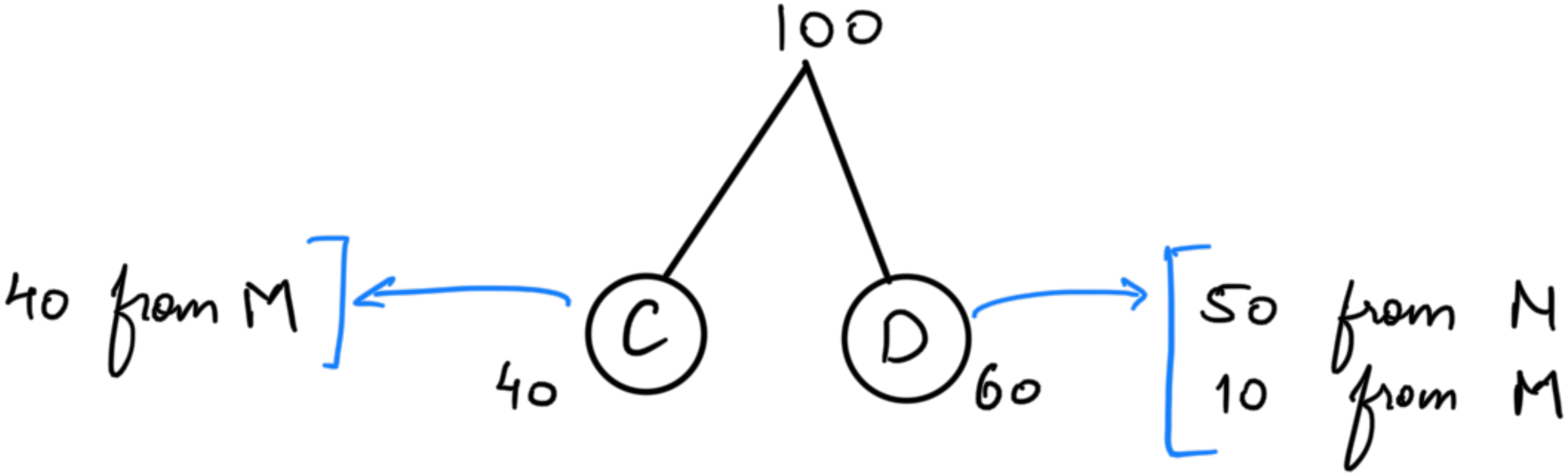
$$\text{mean of ratings} = \frac{3+3+3+3+5}{5} = 3.4$$

User watches all movies on the basis of genres.

User likes movies based on documentaries on the user rating.

Part III - Essay Question

Q1.



Entropy formula $= -\sum_{i=1}^{n} P_{ij} \log_2 P_{ij}$

Cluster C:

$P_{ij} = 40/40 = 1$

$$\text{entropy}(C) = -1 \log_2(1) = 0$$

Cluster D:

$$\text{entropy} = -(10/60)\log_2(10/60) - (50/60)\log_2(50/60)$$

$$\text{entropy}(D) = 0.65$$

LUCKY 7 - BONUS QUESTIONS

Q1. The Turing Award was rewarded.

Q2. Researchers at MIT used recipies from food blogs and other sites where people post recipies.

Q3. An AI-generated portrait sold in recent art auction at $ 432,500

Q4. The name of the organization is OpenAI.

Q5. It was the pong.

Q6. It is Hanabi.

Q7. Finland recently started a program to train 1% of its population.