Q.1. Given below is the **Table** which lists the 4 documents along with the tf-idf score for the vocabulary items and the class to which they belong to. Classify the test document **D5** using Rocchio Classification Algorithm. Show the computation steps clearly.[Documents are to be Length Normalized before applying classification].

Doc/ Terms	**Feather**	**Wings**	**Blood**	**Milk**	**Teeth**	**birth**	**Class**
D1	1.0	0.85	0.5	0	0	0	BIRD
D2	0.72	0.65	0.75	0	0.15	0	BIRD
D3	0.65	0.9	0.55	0	0	0	MAMMAL
D4	0.21	0.75	0.8	0.6	0.9	1.0	MAMMAL
D5	0.43	0.1	0.61	0.1	0.2	0.08	?

[7 Marks]

Length of D1 = sqrt[1+0.85^2+0.5^2] = 1.4
Length of D2 = sqrt[0.72^2+0.65^2+0.75^2+0.15^2]= 1.24
Length of D3 = sqrt[0.65^2+0.9^2+0.55^2]=1.24
Length of D4 = sqrt[0.21^2+0.75^2+0.8^2+0.6^2+0.9^2+1]=1.85
Length of D5 = sqrt[0.43^2+0.1^2+0.61^2+0.1^2+0.2^2+0.08^2]=0.79

D1 = <1.0,0.85,0.5,0,0,0>/1.4= <0.71, 0.6, 0.36,0,0,0>
D2 = <0.72,0.65,0.75,0,0.15,0>/1.24 = <0.58,0.53,0.61,0,0.12,0>
D3 = < 0.65,0.9,0.55,0,0,0>/1.24= <0.53,0.73,0.44,0,0,0>
D4=< 0.21,0.75,0.8,0.6,0.9,1.0>/1.85 = <0.11,0.41,0.43,0.33,0.49,0.54>
D5= <0.43,0.1,0.61,0.1,0.2,0.08>/0.79 = <0.55,0.13,0.77,0.13,0.25,0.10>

Centroid of Bird class = <0.65,0.57,0.49,0,0.06,0>
Centroid of Mammal class = <0.32,0.57,0.44,0.17,0.25,0.27>
Distance between D5 and Bird = 0.59
Distance between D5 and Mammal = 0.62

It belongs to Bird class.

Q.2. a) Which algorithm among Naïve Bayes, kNN and Rocchio is best suited for real time Sentiment Analysis (Refining sentiment classification models based on user feedback on social media posts or comments) ? Justify your answer.

Naïve Bayes is the most suitable answer.

(b) In a search system, the following is the list of relevant documents in a ranked list of 8 documents retrieved in response to two different queries, from a collection of 1000 documents. Assume there are 7 relevant documents in the collection.

- Results for Query 1 are: Results at position 1, 2 ,4 and 6 are relevant
- Results for Query 2 are: Results at position 2,3 and 7 are relevant

i) Compute the F1 score for Query1 and Query2

Calculate MAP score for the search system.

Query 1 : Precision = 4/8= 0.5; Recall = 4/7= 0.57; F1 = 0.5 - 1 mark

Query2 : Precision = 3/8=0.38; recall = 3/7 = 0.43; 0.40 – 1mark

2 marks

Query1	Precision	Query2	Precision
1 -R	1/1=1	1	
2-R	2/2=1	2-R	1/2=0.5
3		3-R	2/3= 0.66
4-R	3/4=0.75	4	
5		5	
6-R	4/6=0.67	6	
7		7-R	3/7=0.43
8		8	

MAP for Query1 = [1+1+0.75+0.67] /4 = 0.86

MAP for Query 2 = [0.5+0.66+0.43]/3= 0.53

MAP Score = (0.86+0.53)/2= 1.39/2 = 0.695=0.7

[2+4=6 Marks]

Q.3. (a) The "Research Institute" has 4 research papers on various topics, and they want to cluster them based on their similarity. The similarity measure between the papers is given below.

(P1,P2) = 0.8; (P1,P3) = 0.9 ; (P1,P4) = 0.6;
(P2,P3) = 0.92; (P2,P4) = 0.75;(P3,P4) = 0.55;

i) Cluster the Papers using Single Link HAC algorithm and give the dendrogram.

ii) Cluster the Papers using Complete Link HAC algorithm and give the dendrogram.

iii) Which algorithm(Single or Complete Link) gives the best clustering for the above problem? Justify.

[2+3+1=6 Marks]

i) Cluster the Papers using Single Link HAC algorithm and give the dendrogram.

Max similarity: (P2, P3) = 0.92

Clusters: {P1}, {P2, P3}, {P4}

Sim({P1}, {P2, P3}) = max(Sim(P1, P2), Sim(P1, P3), Sim(P1,P4)) = max(0.8, 0.9,0.6) = 0.9

Sim({P4}, {P2, P3}) = max(Sim(P4, P2), Sim(P4, P3)) = max(0.75, 0.55) = 0.75
Max similarity: (P1, P2, P3) = 0.9
Clusters: {P1, P2, P3}, {P4}

Sim({P1, P2, P3}, {P4}) = max(Sim(P1, P4), Sim(P2, P4), Sim(P3, P4)) = max(0.6, 0.75, 0.55) = 0.75
Final cluster: {P1, P2, P3, P4}
(Dendrogram diagram)

ii) Cluster the Papers using Complete Link HAC algorithm and give the dendrogram.
P1, P2, P3, P4 (Initially four clusters)
Max similarity: (P2, P3) = 0.92
Clusters: {P1}, {P2, P3}, {P4}

Sim({P1}, {P2, P3}) = min(Sim(P1, P2), Sim(P1, P3)) = min(0.8, 0.9) = 0.8
Sim({P4}, {P2, P3}) = min(Sim(P4, P2), Sim(P4, P3)) = min(0.75, 0.55) = 0.55
Max similarity: (P1, P2, P3) = 0.8
Clusters: {P1, P2, P3}, {P4}

Sim({P1, P2, P3}, {P4}) = min(Sim(P1, P4), Sim(P2, P4), Sim(P3, P4)) = min(0.6, 0.75, 0.55) = 0.55
Final cluster: {P1, P2, P3, P4}
(Dendrogram diagram)

iii) Which algorithm(Single or Complete Link) gives the best clustering for the above problem? Justify.

Best Choice: Single-Link HAC (with justification)

[2+3+1=6 Marks]

Q.4. (a) Assume there are 50 documents in a system. After applying a clustering algorithm, the resultant TP and TN are 25 and 14 respectively. Compute the Rand Index of the clustering algorithm.

N= 50; TP=25; TN=14

Total No. of Pairs = 50 C2
= 50 (50-1) / 2 = 1225
RI = (TP + TN) / (TP+TN+FP+FN)
= (25+14) / 1225
= 0.0318

(or)

Total No. of Pairs = 20 C2
= 20 (20-1) / 2 = 190
RI = (TP + TN) / (TP+TN+FP+FN)
= (25+14) / 190
= 0.205

(b) Assume you are building a content-based recommendation system for Social Media Network. What are all characteristics will you consider in building the item profile? **Justify your answer.**

Content type, Keywords and tags, Topic or category, Sentiment analysis, Popularity, Geographical relevance, Content length, Publication date and Author information

(c) What is the role of back queue router and biased front queue selector in a URL Frontier?

Back Queue Router is responsible for distributing incoming URLs into different back queues and it ensures that URLs from the same host are grouped together.

Biased Front Queue Selector is responsible for selecting URLs from the front queues and it prioritizes which URLs are fetched next, ensuring efficient and effective crawling based on predefined priorities.

[3+2+2=7]

Q.5. "Given below is the Table which has 3 documents and 4 shingles. The Permutation hash functions are given as h1(x) = (x+2) mod 4, h2(x) = 2x mod 3. Compute the Similarity between the documents and identify the near duplicate document pairs.

	D1	D2	D3
S1	1	0	1
S2	0	1	0
S3	1	1	1
S4	1	0	1

	D1 slot	D2 slot	D3 slot
H1(x)=(x+2) mod 4 H2(x)=2x mod 3			
H1(1)=3 H2(1)=2	3,2	-,-	3,2
H1(2)=0 H2(2)=1	3,2	0,1	3,2
H1(3)=1 H2(3)=0	1,0	0,0	1,0
H1(4)=2 H2(4)=2	1,0	0,0	1,0

Jaccard similarity between D1 and D2 = 1 / 2 = 0.5
Similarity between D1 and D3 = 2/2 = 1
Similarity between D2 and D3 = 1 / 2 = 0.5
A near duplicate pair typically has a similarity score close to 1. From the above, **D1 and D3** have a Jaccard similarity of **1**, meaning they are near duplicates.

[5 Marks]

Q6. (a) Consider the Utility Matrix given below for 5 users (**U1 to U5**) and 5 items **(IT1 to IT5**). Compute the Jaccard Similarity and Cosine Similarity between the Users (U2,U3) and (U1,U3).

ITEM/USER	User 1	User 2	User 3	User 4	User 5
Item1	4		2	3	
Item 2	3	2	5		

Item 3			4		2
Item 4		3		5	
Item 5	2		3		3

· **Jaccard Similarity**:

- Jaccard(U2,U3)=1 / 5 =0.2
- Jaccard(U1,U3)= 3 / 4 = 0.75

· **Cosine Similarity**:

- Cosine(U2,U3)= 10 / 26.49 = 0.38
- Cosine(U1,U3)= 29 / 39.56 = 0.73

(b) What do you mean by the term cohesion? What is the role of cohesion in Query Translation?

Frequency of two translation words together

In query translation, cohesion is important for maintaining the meaning and context of the original query when it is translated into another language or when search queries are reformulated.

(c) How does the Gray-Level Co-occurrence Matrix (GLCM) help in texture feature extraction?

GLCM texture considers the relation between two pixels at a time, called the reference and the neighbour pixel.

(d) Give the corresponding queries generated using 2 gram and 3 gram index, in solving the wildcard query ***amo**** .

2-gram query: $a AND am AND mo

3-gram query: $am AND amo

[4+2+1+2=9]

****