Q.1. Given below is the **Table** which lists the 4 documents in the Training set with the appropriate class they belong to and the Test document.

	Doc id	**Words in document**	**Class**
Training Set	D1	India Hockey	Sports
	D2	England England Cricket	Sports
	D3	India India Trade	Politics
	D4	England Crisis	Politics
Test Document	**D5**	**Hockey India Cricket England**	

a) Classify the test document using Multinomial Naïve Bayes model. (Computational steps are to be shown clearly).
b) What is Bernoulli NB model and how it differs from Multinomial NB model, in estimating the parameters. Give suitable application where Bernoulli model can be used instead of Multinomial NB model.

[4+3 = 7 Marks]

a) Prior Probability
P(sports) = P(politics)=2/4=1/2 ---------- 0.5 marks

Conditional Probabilites --- 2 marks

P(Hockey|Sports) = 1+1/5+6 = 2/11 P(Hockey|politics) = 1/11
P(India|sports) = 2/11; P(India|politics) = 3/11
P(cricket|sports) = 2/11; P(cricket|politics) = 1/11
P(England|sports) = 3/11; P(England|politics) = 2/11

Classifier – 1 mark
P(D5|sports) = 1/2*2/11*2/11*2/11*3/11 = 0.0008
P(D5|politics)= 1/2*1/11*3/11*1/11*2/11= 0.0002

Hence D5 belongs to Sports Class – 0.5 mark

3 marks
b) Bernoulli model is Equivalent to the binary independence model which generates an indicator for each term of the vocabulary, either 1 indicating presence of the term in the document or 0 indicating absence

- The Bernoulli model estimates P(t|c) as the fraction of documents of class c that contain term t
- When classifying a test document, it uses binary occurrence information, ignoring the number of occurrences

Suitable for small text fragments like twitter or customer reviews.[Any other suitable example]

Q.2. a) While Evaluating the Classification Algorithms which measures are used for measuring the aggregate performance of the classes? Discuss the measures with suitable examples.

Correct answer is Macroaveraging and Microaveraging. Along with an example –2 marks

(b) The following list of Rs and Ns represents relevant (R) and nonrelevant (N) returned documents in a ranked list of 20 documents retrieved, in response to a query from a collection of 10,000 documents. The top of the ranked list (the document the system thinks is most likely to be relevant) is on the left of the list. This list shows 6 relevant documents. Assume that there are 8 relevant documents in total in the collection.

R R N N N N N N R N R N N N R N N N N R

(i) What is the precision of the system on the top 20?
6/20=0.3 - 1 mark

(ii) What is the F1 score on the top 20?
Recall = 6/8=0.75 F1= 2PR/P+R=2*0.75*0.3/0.75+0.3= 0.45/1.05=0.43 – 1 mark

(iii) Assume that these 20 documents are the complete result set of the system. What is the MAP score for the query?
MAP= 1/6[1/1+2/2+3/9+4/11+5/15+6/20] = 0.555 - 2 marks

[2+1+1+2=6 Marks]

Q.3. (a) An expert system is designed to assist medical professionals by recommending treatment plans based on patient symptoms. The system clusters patients with similar symptoms to suggest appropriate treatments. Given below is a set of 6 patient symptom records represented as vectors, where each patient is represented by two symptom severity scores: **Symptom A Severity (SA)** and **Symptom B Severity (SB)**.

Patients: P1(1,2), P2(2,3), P3(4,4), P4(5,5), P5(8,7), P6(9,6)

Apply K-Means algorithm and cluster these patients into **2 clusters (K = 2)** using the initial centroids: P1(1,2) and P5(8,7). Show the resultant cluster formation and centroid computation after the first iteration. [Normalization not required].

Initial Centroids P1;P2

Cluster 1 - 1 mark
Dist (P1,P2) = 1.4
Dist(P1,P3) = 3.6
Dist(P1,P4)=5.0
Dist(P1,P5)=8.6
Dist(P1,P6) = 8.9

Cluster 2 - 1 mark
Dist (P5,P1) = 8.5
Dist (P5,P2) = 7.2
Dist(P5,P3) = 5
Dist(P5,P4)=3.6
Dist(P5,P6) = 1.4

Cluster1 {P1,P2,P3,} – 0.5 mark
Cluster 2{P4,P5,P6} – 0.5 mark

Centroid of Cluster1 = [(1+2+4)/3, (2+3+4)/3]= (2.33,3.0) – 0.5 mark

Centroid of cluster2 = [(5+8+9)/3, (5+7+6)/3]= (7.33,6.0) – 0.5 mark

(b) You are assigned the task of clustering the students in a class based on their heights. State and justify whether you will choose single-linkage clustering or complete linkage clustering.

Complete link cluster would be best for avoiding chaining effect and also to form well separated clusters.

[4+2=6 Marks]

Q.4. (a) Calculate Purity of the following clustering algorithm results. Di 's are documents and Ci 's are classes.

The true labels of the documents are:

{(D1 : C1),(D2 : C2),(D3 : C1),(D4 : C1),(D5 : C2)}

Resultant Clusters are as follows:

Cluster 1: D1, D2, D3, D4

Cluster 2: D5

Purity of cluster 1 = 3/4 - 0.5 mark
Purity of cluster2 = 1/1 - 0.5 mark
Overall Purity = 3+1/5 - 1 mark

(b) What is the disadvantage of using Purity as an evaluation measure?

- **High purity is easy to achieve when the number of clusters is large – in particular, purity is 1 if each document gets its own cluster.**
- **Hence purity cannot be used to trade off the quality of the clustering against the number of clusters**

(c) In a Transaction Monitoring Application , if fraudulent patterns are to be classified , which algorithm is best suited among Naïve Bayes, kNN and Rocchio? Justify your answer.

Answer is Naïve Bayes, to be given with proper justification.

(d) What is URL Normalization with reference to Crawling? Give a suitable example.

- When a fetched document is parsed, some of the extracted links are *relative* URLs
- E.g., http://en.wikipedia.org/wiki/Main_Page has a relative link to /wiki/Wikipedia:General_disclaimer which is the same as the absolute URL http://en.wikipedia.org/wiki/Wikipedia:General_disclaimer
- During parsing, must normalize (expand) such relative URLs

[2+1+2+2=7 Marks]

Q.5. "Professor Thompson's Research Network" consists of 4 researchers: A (AI expert), B (Data Scientist), C (Computer Vision expert), D (Machine Learning expert).They collaborate on research projects, and the professor wants to rank their influence based on their collaboration network.

The collaboration data shows:

- A collaborates with B, C, and D
- B collaborates with A, D, and C
- C collaborates with A and B
- D collaborates with A and B

Using this data, calculate the PageRank of each researcher, using the power iteration method, assuming a Teleportation rate of 0.8.[Show only the first and second iteration result of Page Rank computation. Assume uniform distribution for initial PageRank].

[5 Marks]

Link Matrix

	A	B	C	D
A	0	1	1	1
B	1	0	1	1
C	1	1	0	0
D	1	1	0	0

Probability transition matrix

	A	B	C	D
A	0	0.33	0.33	0.33
B	0.33	0	0.33	0.33
C	0.5	0.5	0	0
D	0.5	0.5	0	0

Multiply by 1-alpha i.e. 0.2

	A	B	C	D
A	0	0.066	0.066	0.066
B	0.066	0	0.066	0.066
C	0.1	0.1	0	0
D	0.1	0.1	0	0

Add alpha/N i.e. 0.8/4=0.2

	A	B	C	D
A	0.2	0.266	0.266	0.266
B	0.266	0.2	0.266	0.266
C	0.3	0.3	0.2	0.2
D	0.3	0.3	0.2	0.2

First iteration

Uniform distribution. So initial vector is X= (0.25,0.25,0.25,0.25)

Now, perform the multiplication for each row of P i.e. $X.P^1 =$:

- First row:
  $0.2\times0.25+0.266\times0.25+0.266\times0.25+0.266\times0.25=0.25$

- Second row:
  $0.266\times0.25+0.2\times0.25+0.266\times0.25+0.266\times0.25=0.250.\ 0.2667\times0.25=0.25$

- Third row:
  $0.3\times0.25+0.3\times0.25+0.2\times0.25+0.2\times0.25=0.25$

- Fourth row:
  $0.3\times0.25+0.3\times0.25+0.2\times0.25+0.2\times0.25=0.25$

Therefore $X.P^1=(0.25,0.25,0.25,0.25)$

Second iteration

X. P^2 which will result with the same value i.e. (0.25,0.25,0.25,0.25)

After 2 iterations, each researcher has a PageRank of 0.25

Q.6. (a) Consider the Utility Matrix given below for 4 users (**U1 to U4**) and 5 items **(IT1 to IT5**). Compute the Jaccard Similarity and Cosine Similarity between the users (U1,U2) and (U1,U4).

User/item	IT1	IT2	IT3	IT4	IT5
U1	5	4		3	2
U2	3	2	1	1	
U3			4	3	2
U4	4	3	5		1

1. Jaccard Similarity:

Users (U1, U2):

- Rated items for U1: IT1, IT2, IT4, IT5
- Rated items for U2: IT1, IT2, IT3, IT4
- **Intersection** (common rated items): IT1, IT2, IT4 → 3 common items
- **Union** (total distinct rated items): IT1, IT2, IT3, IT4, IT5 → 5 distinct items

J(U1,U2)= Size of Intersection/Size of Union = 3 / 5 =0.6
Similarly, J(U1,U4)= 3 / 5 =0.6

2. Cosine Similarity:

For users U1 and U2

U1: [5, 4, 0, 3, 2]
U2: [3, 2, 1, 1, 0]

Cos(u1,u2)=u1.u2 / |u1|.|u2|

- **Numerator**:

  $(5\cdot3)+(4\cdot2)+(0\cdot1)+(3\cdot1)+(2\cdot0)=15+8+0+3+0=26$

- **Denominator**:

$|u1|=\text{sqrt}(5^2+4^2+3^2+2^2) = \text{sqrt}(54) = 7.348$

$|u2|= \text{sqrt}(3^2+2^2+1^2+1^2) = \text{sqrt}(15) = 3.872$

Denominator = 7.348 * 3.873 = 28.45

Cos(u1,u2)= 26 / 28.45 = 0.914

For users U1 and U4

U1: [5, 4, 0, 3, 2]
U4: [4, 3, 5, 0, 1]

Cos(u1,u4)=u1.u4 / |u1|.|u4|

- **Numerator**:

  (5·4)+(4·3)+(0·5)+(3·0)+(2·1)=20+12+0+0+2=34

- **Denominator**:

$|u1|=\text{sqrt}(5^2+4^2+3^2+2^2) = \text{sqrt}(54) = 7.348$

$|u4|= \text{sqrt}(4^2+3^2+5^2+1^2) = \text{sqrt}(51) = 7.141$

Denominator = 7.348 * 7.141 = 52.48

Cos(u1,u2)= 34 / 52.48 = 0.648

(b) What is Dictionary based Query Translation? What are the different methods in it?
This approach tries to identify and select the possible translations of each source word from a bilingual dictionary
Different methods:
1. Word by word translation - for each word in a query, select the first translation word / select all the translation words
2. Global translation for the whole query - For each query word, determine all the possible translations (through a dict.)

(c) Explain the term "Sensory Gap", in multimedia IR, with an appropriate example.
The gap between the object in the real world and information in recorded scene.

Suppose a user is searching for images of a "happy dog" using an image retrieval system. They input the query "happy dog," expecting results that match the emotional and conceptual idea of a joyful or playful dog. But the system does not inherently understand emotions or abstract concepts like "happiness."

(d) Assume there are 5000 documents in a collection and among them 30 documents contain the term **"science".** Suppose in document **d,** the frequency of term **"science "** is 25, what is the TF-IDF value of the term for that document **d**?

Term Frequency (TF):

Given that the frequency of the term **"science"** in document d is 25, we compute:

$TF(t,d)=1+\log_{10}(25)$

$\log_{10}(25) = 1.3979$

$TF(t,d)=1+1.3979=2.3979$

Inverse Document Frequency (IDF):

$IDF(t)=\log_{10}(5000 / 30) = \log_{10}(166.67) = 2.22$

TF-IDF:

$TF\text{-}IDF(t,d)=2.3979 * 2.22 = 5.32$

[4+2+1+2=9 Marks]

****