Information Retrieval (CS60092)
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur

End Semester Examination

Time: 3 hours
Full Marks: 90

State clearly any assumptions that you feel are necessary.
Solution steps / answers should be supported by proper arguments.

---

1) Consider the following matrix representing **distance** between six documents:

Document	A	B	C	D	E	F
A	0	662	877	255	412	996
B	662	0	295	468	268	400
C	877	295	0	754	564	138
D	255	468	754	0	219	869
E	412	268	564	219	0	669
F	996	400	138	869	669	0

Compute hierarchical single-linkage clustering of these six documents. Clearly show the matrices at each step of building the dendrogram.
(No marks will be given for showing only the Final Dendrogram)

[10]

2) Consider the problem of learning to classify a name as being Food or Beverage. Assume the following training set:

Document	Class
Cherry Pie Chocolate	Food
Chicken Wings Crispy	Food
Cream Soda Water	Beverage
Orange Soda	Beverage

Train a Multinomial Naive Bayes Classifier on the above data. Calculate the multinomial parameters (Priors and Conditional Probabilities). Use *Laplace Smoothing* for calculation of conditional probabilities.

What does this classifier predict about the class of the following test document: **"Chocolate Cream Soda "**? Assume *positional independence* of terms.

[7 + 3 = 10]

3)

a) Write and explain the primal formulation of the optimization problem for building a soft margin SVM.

b) Derive the equation of the hard margin SVM classifier for the following set of labeled points.

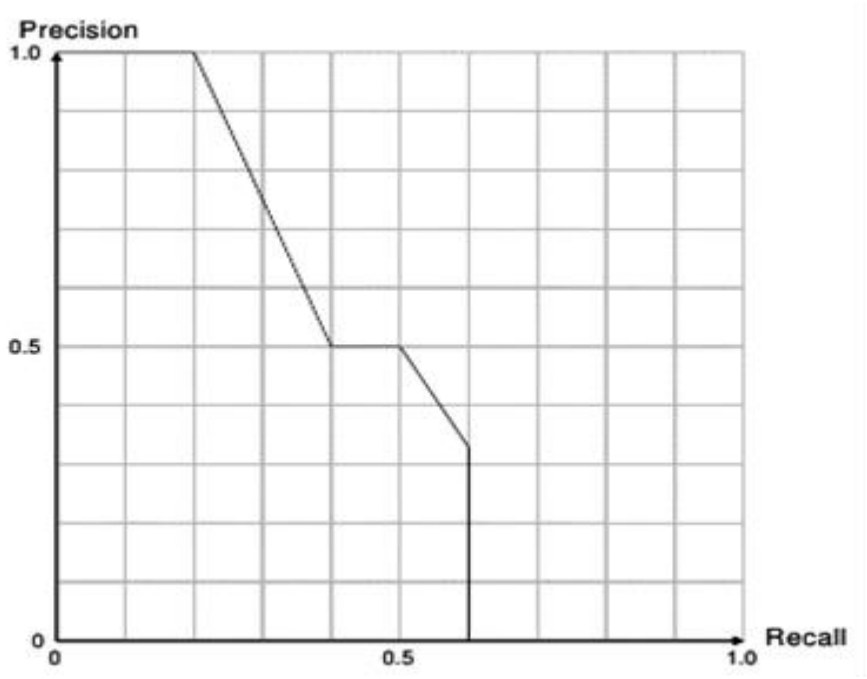
Point	x1	x2	Class
P1	10	3	+1
P1	8	2	+1
P2	4	0	+1
P3	4	2	-1
P4	2	1	-1

c) Assume that few more points are added in the following order. What should be the equation of the SVM classifier after addition of each of these points?

Point	x1	x2	Class
P5	6	0	+1
P6	0	0	-1
P7	4	1	-1

[5 + 4 + 6 = 15]

4) A document retrieval system produced the following interpolated precision-recall curve) on a particular query (based on 20 results):



You know that there are ***ten*** relevant documents.

a) What is the precision after the system has retrieved three relevant documents?

b) Going down the hit list, you discovered that you have retrieved *n* documents, and all of them are relevant. What is the maximum possible value of *n*?

c) What are the positions in the ranked list of 20 results that represent relevant documents?

d) Suppose the relevance label for the relevant documents is 1, and relevance label for the non-relevant documents is 0. Find the NDCG@20 of the result set.

[2 + 2 + 6 + 5 = 15]

5) Consider the following documents:

D1: English Channel Atlantic

D2: National Geography Channel English

D3: Doordarshan National English News

Using unigram language model, rank the above documents for the query: “National News Channel English”. To compute the model probabilities, combine MLE estimates from documents and the collection giving equal importance to both.

[10]

6)

a) Assuming Zipf's law with a corpus independent constant A = 0.1, what is the fraction of words that appear more than 5 times in any fixed corpus of W words?

b) For a search result set, value of reciprocal rank (RR) is 0.125. What are the maximum and minimum possible values of average precision at position 10 (AP@10) for the result set?

c) Suppose that *C* is a binary term-document incidence matrix. What do the entries of C^TC represent? Explain your answer properly.

[3 + 3 + 4 = 10]

7) Consider the following term document matrix C.

Terms	D1	D2	D3	D4	D5	D6
Ship	1	0	1	0	0	0
Boat	0	1	0	0	0	0
Ocean	1	1	0	0	0	0
Voyage	1	0	0	1	1	0
Trip	0	0	0	1	0	1

a) Suppose vector space model is used to represent the documents. Vector dimensions are filled with raw frequency counts of the corresponding terms. According to this representation, what is the similarity between the documents D2 and D3?

b) C is decomposed as $C = U\sum V^T$. The matrices U, ∑ and V are given below.

U =

	1	2	3	4	5
ship	−0.44	−0.30	0.57	0.58	0.25
boat	−0.13	−0.33	−0.59	0	0.73
ocean	−0.48	−0.51	−0.37	0	−0.61
voyage	−0.70	0.35	0.15	−0.58	0.16
trip	−0.26	0.65	−0.41	0.58	−0.09

$\sum$ =

2.16	0	0	0	0
0	1.59	0	0	0
0	0	1.28	0	0
0	0	0	1	0
0	0	0	0	0.39

V^T =

	d1	d2	d3	d4	d5	d6
1	−0.75	−0.28	−0.20	−0.45	−0.33	−0.12
2	−0.29	−0.53	−0.19	0.63	0.22	0.41
3	0.28	−0.75	0.45	−0.20	0.12	−0.33
4	0	0	0.58	0	−0.58	0.58
5	−0.53	0.29	0.63	0.19	0.41	−0.22

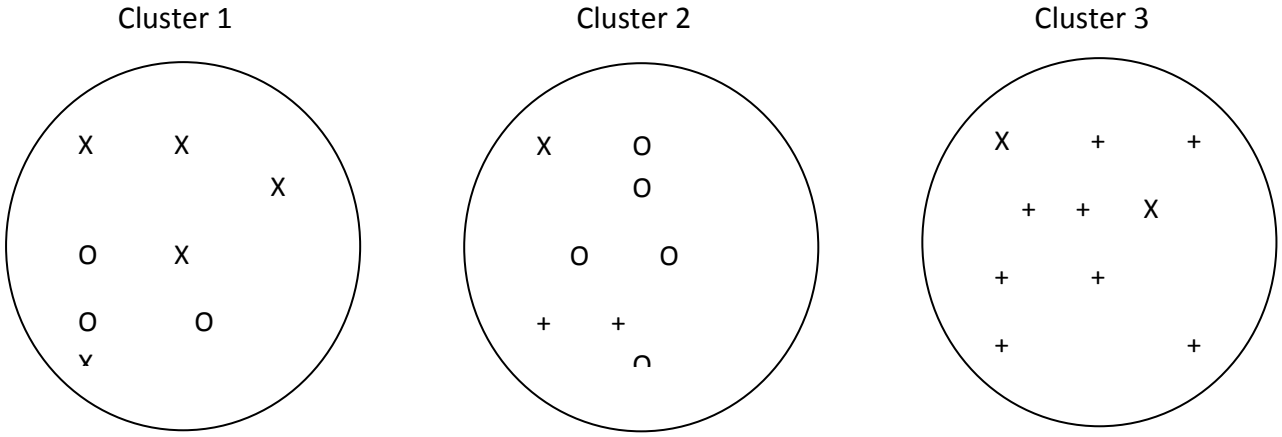
i) Suppose a low rank approximation of C is obtained as C_2 by keeping the *most* important two terms. According to C_2, what is the similarity between documents D2 and D3?

ii) Suppose another low rank approximation of C is obtained as C'_2 by keeping the *least* important two terms. According to C'_2, what is the similarity between documents D2 and D3?

c) Find out the Eigen Values of the matrix CC^T.

[2 + 2 + 2 + 4 = 10]

8) Consider the following figure for clusters found after performing flat clustering (K-Means) on a set of documents. The gold standard for each document is produced by human judges. Each document belongs to one of the three gold standard classes (x, o and +)



Calculate the following quality measures for the above clustering

a) Purity
b) NMI
c) Rand Index
d) F Measure

[2 + 4 + 2 + 2 = 10]