Information Retrieval (CS60092)

End-Semester Examination **Maximum Marks 90**

This question paper has 4 pages and 12 questions.

Note: There are no clarifications. In case of doubt, you can take a valid assumption, state that properly and continue. Please answer PART-A in the first half of the answer sheet (initial 5 pages) and PART-B in the later half.

PART - A

Question 1:

Just as ROUGE-N based on the idea of N-gram co-occurrence statistics, one can build ROUGE-S based on *skip-gram* co-occurrence statistics. For instance skip-bigram is any *pair of words in their sentence order*, allowing for arbitrary gaps. For instance, in a sentence $w_1\ w_2\ w_3$, the skip bigrams are $w_1\ w_2$, $w_1\ w_3$ and $w_2\ w_3$. In the following we have four sentences where S1 is the reference summary and S2, S3 and S4 are candidate summaries.

S1. police killed the gunman
S2. police kill the gunman
S3. the gunman kill police
S4. the gunman police killed

(a) If there are *n* words in a sentence, find an expression for the number of possible skipgrams that could be formed from the sentence.
(b) For each of the sentences S1-S4, list the skipgrams and hence the number of co-occurrences.
(c) Find the recall for each sentence. Which sentence therefore has the best ROUGE-S?
(d) How does ROUGE-S recall compare with ROUGE-L and ROUGE-(N=2) recall for the set of above sentences.

[2+4+(2+1)+3=12]

Question 2:

Show with an example that the Rocchio classification can assign a label to a document that is different from its training set label. **[3]**

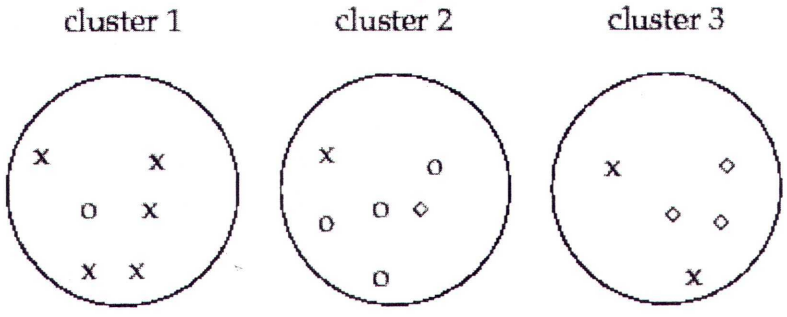
Question 3:

Consider the pairwise distance matrix of seven documents given below. Execute **single linkage hierarchical clustering** on this matrix and draw the dendogram. **[10]**

samples	A	B	C	D	E	F	G
A	0	0.5000	0.4286	1.0000	0.2500	0.6250	0.3750
B	0.5000	0	0.7143	0.8333	0.6667	0.2000	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.6667	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8000	0.8571
E	0.2500	0.6667	0.4286	1.0000	0	0.7778	0.3750
F	0.6250	0.2000	0.6667	0.8000	0.7778	0	0.7500
G	0.3750	0.7778	0.3333	0.8571	0.3750	0.7500	0

Question 4:

Consider the three clusters introduced in one of your class lectures below:



Replace every point *d* in the above figure with two identical copies of *d* in the same class. (i) Is it less difficult, equally difficult or more difficult to cluster this set of 34 points as opposed to the 17 points in the figure? (ii) Compute purity, NMI and RI for the clusters with 34 points. Show by calculation, which measures increase and which stay the same after doubling the number of points? (iii) Given your assessment in (i) and the results in (ii), which measures are best suited to compare the quality of the two clusterings?

[1+(6+3)+1=11]

Question 5:

We have data from the questionnaires survey (to ask people opinion) and objective testing with two attributes (acid durability and strength) to classify whether a special paper tissue is **good** or **bad**. Here are four training samples. **[5]**

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Y = Classification
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

Now the factory produces a new paper tissue that pass laboratory test with X1 = 3 and X2 = 7. Without another expensive survey, can we guess what the classification of this new tissue is? Use a 3-NN approach and show all the steps to classify this new paper tissue.

Question 6:

Derive an expression from w and b in terms of the centroids in Rocchio classification. Assume that there are only two classes: c_1 and c_2. **[4]**

PART - B

Question 7:

Suppose your corpus contains the following 3 sentences:

S1: drug repurposing by integrated literature mining

S2: text mining for drug drug interaction

S3: network analysis and mining for disease drug interaction

Use a mixture language model to rank these document as per relevance to the query – 'drug repurposing'. Use lambda = 0.5. Do not remove the stop words. **[6]**

Question 8:

Consider the following documents as well as the class label assigned to each of these.

S1: drug repurposing by integrated literature mining → DR

S2: text mining for drug drug interaction → DR

S3: network analysis and mining for disease drug interaction → DR

S4: mining the social network → SA

Use a Naïve Bayes text classifier with add-1 smoothing to find the appropriate category for the sentence. Do not remove the stop words.

S5: social network analysis and mining **[6]**

Question 9:

Suppose you want to index corpus from a new language, for which average word size (number of characters) per token is 10, and the average word size per term is 14. Also, it is not very uncommon to have words having 24 characters in this language, so you assign 24 bytes per term for the dictionary storage.

(a) Assume that in your corpus, you have 6 million unique words. Estimate the size of dictionary, while using the standard array of fixed width entries.

(b) How much compression can you achieve on this, if you store dictionary as a (long) string, with pointer to the next word showing end of the current word? [Report the final size of the dictionary]

(c) On top of that, suppose you use blocking with 8 strings in a block? What would be the additional saving? [Report the size after this step] **[3+3+4 = 10]**

Question 10:

Suppose you have a collection containing 5000 documents, and the user issues a query, `social text mining'. Use lnc.lnc scheme to assign relevance scores to document S1 to S4, as given in Question 7 above for this query. You should not use any stemming but remove the stop words, 'by, for, and, the'.

Suppose you use the relevance score above to obtain a ranked list of these 4 documents. Assume that the user marks the 'third' document in the ranking list returned by the system as relevant. If you apply relevance feedback [alpha=0.5, beta=0.5] and rerank the documents, provide the new rankings. **[5+4 = 9]**

Question 11:

As per the binary independence model for ranking the documents as per the odds of relevance, you finally use a retrieval status value (RSV). Write down this RSV. How do you compute it in a pseudo relevance feedback settings. **[4]**

Question 12:

Suppose your collection contains 10 documents. The table below provides the relevance judgments as per two queries q1 and q2, along with the output of the system at the top 10 rank positions for these two queries. **[2+4+4 = 10]**

Doc No.	Judgment (q1)	Judgment (q2)	Ranked order (for q1)	Ranked order (for q2)
D1	R	NR	D4	D2
D2	NR	NR	D3	D1
D3	R	R	D6	D5
D4	R	NR	D8	D7
D5	NR	R	D7	D9
D6	NR	R	D10	D10
D7	R	R	D9	D3
D8	NR	NR	D1	D4
D9	NR	R	D2	D8
D10	R	R	D5	D6

Answer the following questions.

(a) Report precision@5 and Recall@5 of the system.

(b) Report MAP of the system.

(c) Suppose that for all the documents marked as 'R', if they are odd (e.g., D3) their relevance is 1, and if they are even, their relevance is 2. Report the NDCG@5 of the system for these two queries separately.