Question 1: Relevance Feedback
Points: ___

In the relevance feedback approach, we are given a set of relevant documents, D_r, and a set of non-relevant documents, D_{nr}. By using the Rocchio method we can modify the initial query q_0 and obtain the modified query q_m as follows

$$q_m = \alpha q_0 + \frac{\beta}{|D_r|} \sum_{d_j \in D_r} d_j - \frac{\gamma}{|D_{nr}|} \sum_{d_j \in D_{nr}} d_j \qquad \text{(a)}$$

where α, β, γ are tuning parameters.

Questions:

1) Suppose that a user's initial query is ``*cheap CDs cheap DVDs cheap CDs*''. The user examines two documents, $d_1$ and $d_2$. She judges $d_1$, with content ``*CDs cheap software cheap CDs*'' relevant and d_2 with content ``*cheap thrills DVDs*'' non relevant. Assume that in document and query vectors we are using only the pure term frequency, i.e. $w_{ij} = freq(i,j)$. Using Rocchio relevance feedback as given in Equation (a), what would the modified query vector be after relevance feedback? Assume, α=1, β=0.75, γ=0.25. Please list the vector elements in alphabetical order.

2) For the example given in 1) re-compute the modified query vector using term weights computed by the standard *tf-idf* method, i.e. $w_{ij} = tf(i,j)\ idf(i)$. We assume that the document collection is $D=\{d_1,d_2\}$. Briefly discuss and compare the answers obtained in 1) and 2). Please list the vector elements in alphabetical order.

3) In Rocchio's method, if we are given a document, say d_j, then what setting of the tuning parameters should be used for finding other documents like d_j?

Question 2: Schema Fragmentation
Points: ___

Given the following relational table *T.*

PK	Player	goals	fouls	viewer satisfaction
1	Beckham	1	10	80
2	Beckham	2	15	95
3	Zidane	1	8	70
4	Zidane	2	12	65
5	Zidane	8	20	75
6	Rooney	1	8	60

Let applications at three different sites S1, S2, and S3 access the table making the following queries.

Q1: SELECT * FROM T
Q2: SELECT* FROM T WHERE fouls ≥ 10 AND viewer satisfaction ≥ 75
Q3: SELECT * FROM T WHERE player = Rooney
Q4: SELECT * FROM T WHERE player = Zidane AND fouls < 15

The following table lists the daily access frequencies of the queries from each of the sites.

	Q1	Q2	Q3	Q4
S1	10	0	f_{13}	10
S2	5	f_{22}	f_{23}	0
S3	0	5	10	10

Assume that horizontal fragmentation has resulted in the following two fragments:

F1: {1, 2, 5}

F2: {3, 4, 6}

Questions:

1. Find the frequencies f_{13} and $f_{23.}$

2. Assume that the cost of query evaluation is measured in terms of cost of transferring the result tuples to the requesting site and that the following communication costs for transferring one tuple between two sites are given.

	S1	S2	S3
S1	0	10	5
S2	10	0	5
S3	5	5	0

   What is the total query evaluation cost incurred by the application running at site S3 assuming that F1 is hosted at S1 and F2 is hosted at S2?

3. Determine f_{22} assuming that F1 is hosted at S1 and F2 is hosted at S2 and total communication cost for S2 is 450.

Question 3: Mobile Broadcasting
Points: ___

Assume a mobile data broadcast is given that consists of broadcast disks with the following numbers of different pages on each disk and their corresponding access probabilities.

Disk	Number of pages	Access Probability of each page
D1	6	100/3696
D2	8	225/3696
D3	4	324/3696

Questions:

1) Calculate the optimal broadcast schedule for $f_{min} = 1$ and $f_{min} = 2$. Denote the i^{th} page on the j^{th} Disk by P_{ji}.

2) Compute the expected average delay for each of the two broadcast schedules.

3) In each case what is the difference between the achieved delay and the theoretically lowest achievable delay? Can this difference be decreased? If yes how?

Question 4: Hierarchical Peer-to-peer networks
Points: ___

In a two-level hierarchical P2P network the set of peers *P* of size *n* is partitioned into *p* disjoint clusters $C_1,..,C_p$ of equal size. The communication among peers, e.g. for locating a resource, can occur at two levels:

1. Intra-cluster communication: messages exchanged among peers within the same cluster
2. Inter-cluster communication: messages exchanged among peers from different clusters

Two different approaches have been studied for designing such hierarchical P2P networks:

1. *Superpeer overlay networks*: each cluster elects one peer, called the superpeer. The superpeers are connected in an unstructured overlay network. All inter-cluster communication is performed among the superpeers from the different clusters. Peers within a cluster communicate directly with their superpeer in a client-server style. Peer's resources are registered at their superpeer. Search is performed by first asking the local superpeer. If the resource is not registered there, the search is flooded in the superpeer network. The TTL and connectivity are assumed to be set such that flooding search reaches all superpeer nodes.

2. *Homogeneous hierarchical overlay networks*: peers of the same cluster form local overlay networks for intra-cluster communication. In addition all peers participate in a joint global overlay network for inter-cluster communication. We assume that both the local and global overlay network follow the Chord design. Peers have assigned the same identifiers in the local and global overlay network and maintain two separate routing tables for the local and global overlay network. Resource keys are inserted in the local overlay network of the cluster to which the peer providing the resource belongs. In addition resource keys are also inserted into the global overlay network. Search is performed by first routing in the local overlay network. If the resource is not found there it is searched by routing in the global overlay network.

Questions:

1) Assume that peers are more likely to request resources from their own cluster. Let $1/p < \delta \leq 1$ denote the probability that a resource from the own cluster is requested. What is the expected search cost depending on the value of δ in terms of the size of the network n and the number of clusters p for the two types of hierarchical overlay networks? Assume that the search cost is measured as the expected average number of peers that receive at least one search message during a search.

2) Assume that the total number of peers is $n=2^{20}$ and the number of clusters is $p=2^6$. For which value of δ the homogeneous hierarchical overlay networks are more search efficient than the superpeer overlay networks? For which value of δ the homogeneous hierarchical overlay networks are more search efficient than a standard flat Chord network?

3) Now assume that on average for each search one new resource key is inserted into the network. The values of n and p are as in 2). Considering the combined search and insertion cost, for which value of δ the homogeneous hierarchical overlay networks are more efficient than the superpeer overlay networks? For which value of δ the homogeneous hierarchical overlay networks are more efficient than a standard flat Chord network?

4) Difficult: Assume that in the homogeneous hierarchical overlay network resources are only stored at the peer responsible for the resource identifier according to the local overlay network of the corresponding cluster. Discuss how still resources could be searched using the global overlay network. What cases need to be considered and what would be the search cost as compared to the basic approach described before?