Additional Problems:

Problem 1. K-means clustering.
Given are the points A = (1,2), B = (2,2), C = (2, 1), D = (-1, 4), E = (-2, -1), F = (-1,-1)
a) Starting from initial clusters Cluster1 = {A} which contains only the point A and
Cluster2 = {D} which contains only the point D, run the K-means clustering algorithm and report the final clusters. Use L1 distance as the distance between points which is given by
d((x1, y1), (x2, y2)) = | x1 – x2 | + | y1 – y2 |
b) Draw the points on a 2-D grid and check if the clusters make sense.

Problem 2. Bottom-up hierarchical clustering
Given are the 1-dimensional points A = 1, B = 2, C = 3, D = 8, E = 9, F = 10.
Compute single-link bottom-up hierarchical clustering using d(x, y) = |x – y| as the distance between points.

Problem 3. Gradient Descent
Describe the basic idea of gradient descent
(See here: http://en.wikipedia.org/wiki/Gradient_descent)

Problem 4.
Why will Nearest Neighbor classifier fail if the dimension of the data is large

Problem 5. MapReduce Application

Show how counting the frequency of all words in a document can be implemented with MapReduce. Use pseudo-code. Specify both the code in the mapper function and the reducer function.

Problem 6.
If I run K-means on a data set with n points, where each points has d dimensions for a total of m integrations in order to compute k clusters how much time will it take? (answer is a function of n, m, k, d).

Problem 7. Cluster Hypothesis for IR
a) State the cluster hypothesis for Information Retrieval
b) Describe how it can be empirically verified.

(see book)

Problem 8. Relevance Feedback for the Vector Space model.

Assume that there are only 3 different words in the collection: A, B and C
The query is “A B”
The user is presented the documents D1 = “A A B C A”
D2 = “B C A A C”
D3 = “C C A A B”
ranked in this order as an answer to this query.
The user marks document D3 as relevant. Using the Rochio algorithm, show the ranking of D1 and D2 after the user feedback.

Reference to formula:

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j$$

For this problem set alpha = 0.5, beta = 0.5, gamma = 0.0