The Theory of Computation
Bernard M.E. Moret
Addison-Wesley-Longman, 1998

Solution Manual
Chapter 8, Additional Problem 8.4

1. First note that the sum of the intra-cluster and the inter-cluster dissimilitudes is the constant $\sum_{x,y} d(x, y)$. This shows that an optimal solution for one is an optimal solution for the other and vice-versa (thereby solving part (ii)); it also allows us to Turing-reduce one problem to the other. Thus it is enough to show that one of the two problems is NP-equivalent. Consider then *k-max clustering*; its decision problem can be restricted (by setting $k = 2$) to the known NP-complete problem *Max cut*; thus the optimization version is NP-hard.
2. This was proved above.
3. We show that the existence of any ε-approximate algorithm for *k-min clustering* would allows us to solve in polynomial time a known NP-complete problem, *graph k-coloring*. Specifically, given a graph, $G = (V, E)$, to be colored in k colors, we transform it to a k-min clustering problem as follows. The set of vertices becomes the set of elements; the dissimilitude measure is defined as

$$d(x, y) = \begin{cases} 2\varepsilon n^2 & \text{if } \{x, y\} \in E \\ 1 & \text{otherwise} \end{cases}$$

   Now, if the graph had a k-coloring, we can obtain a clustering of cost no larger than n^2 by putting in cluster i all elements corresponding to vertices colored i; if the graph did not have such a coloring, however, at least two elements corresponding to connected vertices will be in the same cluster, thereby guaranteeing a cost of at least $2\varepsilon n^2$. Our ε-approximate algorithm will return a solution of cost no worse than εn^2 in the first case, no better than $2\varepsilon n^2$ in the second (it cannot do better than optimal), so that we can distinguish the two cases and thus decide graph k-colorability.

4. A suitable choice for the greedy measure is the minimum intra-cluster dissimilitude—which is rather surprising in view of the result of part (iii). However, the choice is perfectly reasonable, since the two measures are essentially the same as seen in part (ii). It remains to show that such an approximation algorithm has a ratio guarantee of $\frac{k-1}{k}$.
5. There are two reasons. One explains why one of the two problems can be hard to approximate while the other is relatively easy: their optimal values sum to a constant and are generally orders of magnitude apart, so that a small error in one becomes a large error in the other. The second explains why it is *k-min clustering* that is hard: it has the smaller optimal values.