MIDTERM CS6720
(Time: 1.5 hours)
Total Points: 60
[Keep your answers short and precise.]

1. On any dataset where k-means clustering can be performed, k-medoid clustering can also be employed. In addition, k-medoid clustering can also be applied on non-Euclidean spaces. Then why use k-means at all? [2 points]

2. You have a huge database of 100 million chemical compounds. You represent each compound as a graph, and you are also given a metric distance function to compare graphs. You now want to cluster this database, but the entire database cannot be loaded into main memory. Which clustering technique (among those you learned in class) would you use and why? [3 points]

3. It is clear that parameters are easier to set in OPTICS than in DBSCAN. But if parameter selection is not a problem (let's say some oracle tells us the best parameters), would you still say OPTICS is better? Explain. [4 points]

4. In DBSCAN, a cluster is defined as follows:
   A set of points C is a cluster if
   - For any two points $p, q \in C$, p and q are density-connected
   - There does not exist any pair of points $p \in C$ and $s \notin C$ such that p and s are density-connected

   Prove or disprove that these two properties always hold in DBSCAN. [6 points]

5. Perform single-linkage hierarchical clustering in O(n^2) time? You should explain the complexity analysis. [7 points]

6. FP-tree
   a. Draw the FP-tree for the following transactions. Assume that the **items are sorted lexicographically**.[2 points]:

1	{W,X,Z}
2	{W,X,Y}
3	{V,X,Y,Z}
4	{V,Y,Z}
5	{V,W,X}
6	{V,W,X,Y}
7	{W,X}
8	{V,W,X}
9	{V,W,Y}
10	{V,W}

V: 7
W: 8
X: 7

Same as in slides. Only the letters have been changed from A-E to V-Z.

   b. For the following FP-tree, mine all frequent patterns containing item Z. Support threshold is 2. Show the recursion-tree that will be generated using FP-growth corresponding to each frequent pattern containing Z, and their corresponding conditional FP-trees. [8 points].

c. Are we guaranteed to always have the most compact FP-tree if the items are sorted based on frequency in descending order? Prove or disprove. [5 points]

7. In the multiple-support frequent itemset mining, all subsets of a frequent itemset may not be frequent. Despite this, what property makes the MSApriori algorithm work? [3 points]

8. In frequent sub-sequence mining, we looked at the problem of mining all sub-sequences that occur in at least some θ sequences from the database. We studied two different algorithms: prefixSpan and GSP. Now, consider the problem of mining frequent sub-sequences in a single long sequence.
To define the support of a subsequence in this setting, recall the definition of sub-sequence: *"A sequence $S = \langle a_1 a_2 \ldots a_r \rangle$ is a* ***subsequence*** *of another sequence $T= \langle b_1 b_2 \ldots b_v \rangle$, or S is a* ***supersequence*** *of T, if there exist integers $1 \leq j_1 < j_2 < \ldots < j_{r-1} < j_r \leq v$ such that $a_1 \subseteq b_{j1}$, $a_2 \subseteq b_{j2}$, ..., $a_r \subseteq b_{jr}$."*

The support of S in T is the number of integer sets $\{j_1, \cdots, j_r\}$ for which S is a subsequence of T. Can we use PrefixSpan or GSP to solve this problem? If yes, what modifications need to be made (if any)? If not, why not? [7 points]

9. In the implementation of PageRank, the new rank vector, R_{new}, is kept in memory, whereas the link matrix M, and the old stationary vector, R_{old}, are stored on disk. If we store R_{old} in memory instead of R_{new}, would there be any advantage or disadvantage? [2 points]

10. Consider the following approach to perform frequent itemset mining with the frequency threshold x% in a distributed framework. Given a database of transactions D, the dataset is divided in an arbitrary manner into m nodes such that $\bigcup_{i=1}^{m} D_i = D$ and $D_i \cap D_j = \emptyset$. Next, the locally frequent itemsets in each local database D_i are mined. Finally, all of the locally frequent itemsets are aggregated using x% as the frequency threshold and checked if they are globally frequent as well. These filtered globally frequent itemsets are returned as the final answer set. Is this answer set optimal? In other words, can there be any false positive or false negatives? [5 points]

11. In frequent itemset mining using the breadth-first approach, after candidate generation, we try to prune a candidate using the apriori rule. If it passes this check, we scan the candidate across the entire database to compute its true frequency. Can we avoid scanning the entire database for all candidates through some additional checks? [6 points]