Assignment 1: NRA (0 P.)

Consider a top-k query with $m = 3$ terms, the user is interested in $k = 2$ results, and (non-weighted) summation as score aggregation. The underlying three index lists have the following (document identifier, score) entries:

L_1
d_1 0.9
d_7 0.7
d_3 0.3
d_2 0.3
d_4 0.3
d_5 0.2
d_6 0.1

L_2
d_3 0.8
d_4 0.8
d_7 0.5
d_1 0.3
d_6 0.2
d_5 0.2
d_2 0.2

L_3
d_1 0.7
d_6 0.6
d_7 0.5
d_4 0.4
d_2 0.3
d_3 0.2
d_5 0.1

(a) Apply the NRA method (no random accesses) to this setting. Document all index accessing steps, the top-k, and the set of candidates after each of them. How many sorted accesses does the method need?

Assignment 2: Frequent Itemset mining (0 P.)

(a) The *Apriori* algorithm uses a generate-and-count strategy for deriving frequent itemsets. Candidate itemsets of size $k + 1$ are created by joining a pair of frequent itemsets of size k. A candidate is discarded if any one of its subsets is found to be infrequent during the candidate pruning step. Suppose the *Apriori* algorithm is applied to the data set shown in following Table with minimum support = 3.

TID	**Itemset**
1	$\{a, c\}$
2	$\{b, c\}$
3	$\{a, b\}$
4	$\{a, b, d\}$
5	$\{c, d\}$

Draw an itemset lattice representing the data set given in above table. Label each node in the lattice with the following letter(s):

- **N**: If the itemset is not considered to be a candidate itemset by the *Apriori* algorithm. There are two reasons for an itemset not to be considered as a candidate itemset: (1) it is not generated at all during the candidate generation step, or (2) it is generated during the candidate generation step but is subsequently removed during the candidate pruning step because one of its subsets is found to be infrequent.
- **F**: If the candidate itemset is found to be frequent by the *Apriori* algorithm.
- **I**: If the candidate itemset is found to be infrequent after support counting.

(b) The area of the itemset X in data D is $area(X, D) = |X| \times supp(X, D)$, i.e. the product of the number of items in the itemset and the support. Determine which of the following claims is true and prove it (either prove that the property holds or show via counter example that it does not hold).

(i) $area(X)$ is monotonically decreasing (downwards closed), i.e. if X and Y are itemsets such that $X \subset Y$, then $area(X) \geq area(Y)$.

(ii) $area(X)$ is monotonically increasing (upwards closed), i.e. if X and Y are as above, then $area(X) \leq area(Y)$.

(iii) $area(X)$ is neither monotonically increasing nor monotonically decreasing.

Assignment 3: Density-Based Clustering (0 P.)

Consider the data in Figure 1. Answer to the following questions assuming that we are using the Euclidean distance and that $\varepsilon = 2$ and **minpts** $= 3$.

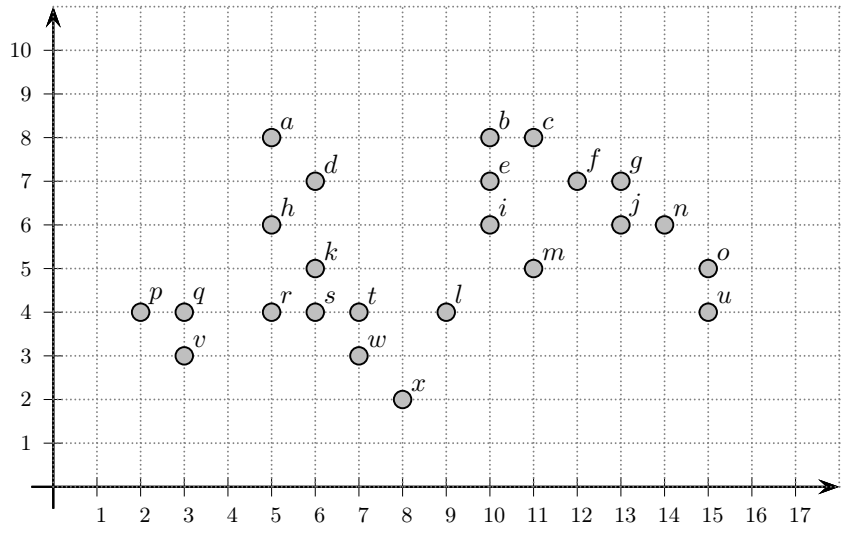

Abbildung 1: Points in a space

- List all the core points.
- Is a directly density-reachable from d?
- Is o density-reachable from i? Show the complete chain or where it breaks.
- Is l density-connected to x? Show the intermediate points that make them density-connected or that break the condition.