

Minimal Generalization Algorithm: I choose the Minimal Generalization Algorithm (MinGen) to implement k-anonymization [1] to the Adults.txt dataset.

Summary: The below are the algorithm steps

1. Firstly, read the adults.txt file data and then prompt the user to input K-value.
2. Check whether the given file is K-anonymous or not. If yes, then return this file as output. Otherwise, go to the 3rd step.
3. Select the age, work type and gender as quasi identifiers and create all the possible generations of the table as
 Gender -> 0, 1 {male and female at 0th level; * at 1st level}
 Age -> 0,1,2,3,4,5,6 {take the ages range with multiples of 5 as 5, 10, 15, 25, 30}
 Work type -> 0, 1, and 2 {as mentioned in the assignment2 description}.
4. Apply each generation to the data table and check whether it is k-anonymous or not
 If not, then please filter out that generation.
 If yes, then calculate the precision for that generation using the below formula and store these values against the generations.

$$Precision = 1 - \frac{\sum_{i=1}^{Na} h}{Na}$$
 Na = No of quasi identifier attributes. (age, gender, work type in the assignment)
 h = Attribute height to which it is generalized.
 D = Total height of the attribute generation. (Age-6, gender -1, work type -2).
5. Repeat the step4 for all the generations
6. Finally, compare and find the best precision value-generation (from the 4th step calculated values) among all the K-anonymous generations. Then, apply that generation to the given input adults.txt file and write the same on the output file.

Advantages:

- We can properly interpret the data that obtained as results from the generalization and suppression process. Thus, we can say what has been done to the data to get these results.
- Data presented in the results is truthful and so, we can use it for fraud-detection, health care assessments and to identify specific patterns.
- We can guarantee the output result with k-anonymity using MinGen algorithm.

Drawbacks: Mingen algorithm is an expensive and inefficient algorithm compared to other algorithms. We pose the generalization at each cell level and it is impractical to search all the possible generalizations on the mid-sized tables.

Reference [1]: [Latanya Sweeney: Achieving k-Anonymity Privacy Protection Using Generalization and Suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10\(5\): 571-588 \(2002\).](#)