Comp 430 - HW1 Report

Mehmet Üstek

Important Notice: Please install the python module ‘treelib’ before running the code.

Analysis:

For simplicity and faster convergence I will use a smaller part of Adult data rather than the whole Adult data itself, since whole Adult data contains more than 45.000 records.

For partial Adult data of length 1400:

Anonymizer / k tables for Run Time Analysis, MD and LM Costs:

Time (sec)	**3**	**5**	**10**	**20**	**40**	**80**	**160**	**320**
Random	**0.08**	**0.13**	**0.30**	**1.42**	**8.04**	**58.31**	**TBD**	**TBD**
Clustering	**394**	**465**	**601**	**677**	**833.93**	**1180**	**TBD**	**TBD**
Topdown	**37.11**	**30.32**	**22.34**	**18.00**	**10.75**	**8.57**	**8.32**	**5.07**

MD Cost	**3**	**5**	**10**	**20**	**40**	**80**	**160**	**320**
Random	**11111**	**12275**	**12224**	**12980**	**13472**	**13380**	**TBD**	**TBD**
Clustering	**11985**	**10559**	**12521**	**12620**	**13603**	**13715**	**TBD**	**TBD**
Topdown	**6597**	**9714**	**12061**	**13728**	**21236**	**22953**	**22974**	**25197**

LM Cost	**3**	**5**	**10**	**20**	**40**	**80**	**160**	**320**
Random	**710.03**	**708.70**	**735.46**	**699.68**	**714.41**	**776.54**	**TBD**	**TBD**
Clustering	**712.76**	**673.24**	**718.12**	**731.09**	**762.93**	**755.25**	**TBD**	**TBD**
Topdown	**343.96**	**437.27**	**639.36**	**716.70**	**1068.34**	**1151.74**	**1144.76**	**1282.241**

TBD: Values are not calculated since the running time is long. It is trivial to estimate the corresponding values for these cells using the given data.

Discussion:

From the above data, it is trivial that for run time efficiency and stability, the best option is Top Down Approach. Random anonymizer, as expected, gives the average results based on MD and LM costs, and it is significantly faster than other algorithms when k is small. However, the run time increases monotonically when k increases. The clustering algorithm is random, thus obviously it is not the best for anonymizing data regarding the smaller k values until 40. Since I used a smaller data length of 1400, k = 40 becomes the breaking point where randomness is even better than any heuristic approach. For smaller k values, Top Down Approach proves itself to be the best choice regarding the MD and LM costs. However, the TD approach loses its charm when the k increases, since the MD and LM costs increase monotonically with respect to k. Likewise, the run time for TD is monotonically decreasing wrt k. Overall, for smaller values of k, TD is the best option regarding the lowest utility loss, faster runtime. Clustering takes an enormous amount of runtime, however with higher values of k, it gives the minimum utility loss. As expected, the random approach has a significantly better runtime for smaller k values and it gives average anonymization regarding the utility loss. Although I did expect TD to be best, I did not expect it to be this fast compared to clustering. I expected clustering to take enormous time based on its complexity analysis, however I thought its utility loss would be lower. For larger k values, it makes sense that clustering is the best approach.

In the process of coding, I learned the significance of having a competent heuristic function to anonymize the data. I learned the relationship of k values and utility loss for each heuristic approach.