COMP430 DATA PRIVACY AND SECURITY ASSIGNMENT1 REPORT, FALL 2022

Name-Surname: Barış Kaplan, **KU ID:** 0069054, **KU Email:** bkaplan18@ku.edu.tr

FOR SEED VALUE = 4

Name of anonymization	MD Cost	LM Cost	k parameter value	Total execution time
Clustering	24022	1480.9144619269528	4	00:11.077203
Clustering	29334	1676.106440781425	8	00:11.486594
Clustering	32270	1697.933483183475	16	0:00:12.544645
Clustering	30105	1817.4434334892	32	0:00:13.832239
Clustering	31256	1901.99298331	64	00:15.389239
Clustering	32399	1989.3484873473	128	00:17.238397

FOR SEED VALUE = 8

Name of anonymization	MD Cost	LM Cost	k parameter value	Total execution time
Clustering	24022	1480.9144619269528	4	00:10.602627
Clustering	29334	1676.106440781425	8	0:00:11.575255
Clustering	32270	1697.9334831835	16	0:00:13.545217
Clustering	30105	1817.4434334892	32	0:00:14.775987
Clustering	31256	1901.99298331	64	0:00:17.994001
Clustering	32399	1989.3484873473	128	0:00:18.833402

Some notes:

- **In this assignment, I have utilized the Tree and Node classes of the treelib module of Python. If you have installed the treelib module beforehand, you are good to go. However, if you have not installed the treelib module, please run “pip install treelib” command from your desktop’s terminal or from your ide’s terminal. After you run this command, the treelib module will be successfully installed.**
- **I have implemented the randomized anonymizer algorithm. Moreover, I have done the bottom up anonymizer partially. Even though I have read various relatedd articles, corresponding PDF descriptions, and several related technical blogs for an extreme amount of time, I could not figure out how to solve some bugs / errors in these parts. My bottomup anonymizer and randomized anonymizer codes have some bugs. Thus, I cannot get an output from both of these parts.**

Take aways & what I have learned & my expectations: While representing the domain generalization hierarchy I read, I have used the Tree and Node classes of the treelib module of Python. While coding, I have learned how and where to use some functions of treelib such as leaves(), get_node(), an subtree(). Moreover, while I am doing research and coding, I have learned the difference between the randrange() and randint() functions. Moreover, I have learned the exception types of treelib module, how to handle these treelib exceptions, and how these exceptions are handled internally in the treelib library. Moreover, while I am coding, I have examined and understood the approaches of the internal implementations of the functions (such as leaves(), subtree(), create_node(), and so on) in thre treelib module of python. For the clustering, I have observed that as the md cost increases, lm cost also increases. Moreover, I have observed that the value of the k parameter is directly proportional with the total execution time, lm cost, and md cost. In other words, as the value of the k parameter increases from 4 to 128 gradually; the md cost, lm cost, and total execution time tend to increase as well. Since the clustering-based anonymization includes more subparts (compared to randomized and bottom-up) and since it is more detailed than the others, I have expected to get higher execution times for the clustering-based anonymization algorithm. While I am coding, I have also learned how to use some python functions such as randrange(), math.prod(), randint(), and so forth. When I change the seed value from 4 to 8, I have observed some minor changes in the execution times.When I change the seed value to 8, I expected to observe more changes in the values of the execution times. However, I observed minor changes in the total execution times when I change the seed value. Moreover, for each seed value (4 and 8), even though I expected higher execution times for each k value, I obtained much lower execution times than my expectation.