CMSC 5333 – Advanced Database
Final Exam (Project)
2000 pt
Due: **Thursday, April 28, 11:59 PM**

Data anonymization is a process that takes data pertaining to certain users and removes identifying information. One technique for data anonymization is called "k-anonymity." In this technique, unique identifiers are completely removed, and some data that could lead to identification (called **quasi-identifiers**) are generalized (replaced with a less-specific value).

A data set is said to be k-anonymous if every query involving quasi-identifiers returns either zero or at least k rows. For example:

Zip	Age	GPA
73013	32	4.0
73017	25	3.4
73024	34	2.7
73025	22	3.2
73013	26	3.1
73017	28	3.7
73024	34	3.4
73013	22	2.2
73013	23	1.6
73025	27	2.4
73024	43	3.7
73013	26	3.9

----->

Zip	Age	GPA
7301*	25-35	4.0
7301*	25-35	3.1
7301*	25-35	3.9
7301*	20-25	2.2
7301*	20-25	1.6
7301*	20-25	3.4
730**	25-35	3.7
730**	25-35	2.7
730**	25-35	3.4
7302*	20-45	3.7
7302*	20-45	2.4
7302*	20-45	3.2

If we let Zip code and Age be quasi-identifiers, then the data set on the left is not anonymized at all (although some queries on quasi-identifiers would return 2 entries, others would return 1). However, if we transform it into the data set on the right (having generalized some of the values), then queries on zip code and age would either return no rows, or at least 3 rows. Thus, the set on the right is 3-anonymous.

Your task for this is as follows:

1. Download the "adult" data set from the UCI Machine Learning Repository here:

   http://archive.ics.uci.edu/ml/machine-learning-databases/adult/

   In this folder is "adult.data" which is the data set, and "adult.names" which describes the fields in the data set.

2. Create code to run a MapReduce job (or chain of jobs) that will produce a 3-anonymous data set from the adult data, where the quasi-identifiers are:

   Age, Education, Race, Sex.

3. Submit your code, and the first page of your results (as gathered from Hue) to this assignment.

A high-level summary of how to do the k-anonymity algorithm:

Sort the rows by the quasi-identifiers (doing this minimizes the amount of generalization necessary)
For each set of k rows
    If all the quasi-identifiers are the same, leave them alone
    Otherwise, generalize the identifiers which are different

Note that your code can be in Java, Pig, Hive, or any other platform on which you can do MapReduce jobs.

Good Luck!

Note on Academic Honesty: This assignment is to be done **individually**. Some of you may have to use other students' hardware, and that is fine. However, you may not use other students' code, nor should your code resemble any other students' code. If there is any evidence that you worked with another student to create your code, you will not receive credit for the assignment.