**11676 – Big Data Analytics**

**lneves**

**HW 5 – MapReduce**

For this assignment, I have created a HadoopDriver class to start the mapreduce jobs and to store the test results of my code into Cassandra. I have decided, since I was afraid of the bottleneck of reading all instances serially, to have my mapreduce to be divided in two steps. The first will use the DecisionTreeMapper and DecisionTreeReducer. The mapper here reads an instance from Cassandra and generates an instance object, what allow me to preprocess data from Cassandra in parallel. From this, I will assign a random number from 0 to number of trees -1. Assuming the random function results are uniformly distributed, this way I was able to create N decision trees for my random forest. My reducer will receive the instances for each tree, train each tree and persist them to Cassandra. This way, if I want to select only the “best trees”, I have all trees on my database. The second pass will have an identity mapper that will just send the tree read from Cassandra to a reducer that will aggregate my trees and persist a random forest model to Cassandra. Those would be my RandomForestMapper and RandomForestReducer. Finally, my driver reads the forest from Cassandra and persists the error metrics running on a test set.