ECE 454 Assignment 3 Design Doc

# Part 1

Part 1 consists of one mapper. It finds the genes with highest expression rate, and store it in an array list. Then the array list is written to the context.

# Part 2

In the mapper, the samples are divided into gene values. The reducer takes them in and counts the genes with values greater than 0.5 and divides by the total number of occurrences for that gene. The result is written to context.

# Part 3

The mapper creates x number of keys for the reducer, where x is the value from 1 to sample number-1, along with the gene values. The reducer will take in the values and do the cross product calculation, and then output the result.

# Part 4

## Part 1

First, input is read and put into the bags. Then it is passed to custom UDF, where it passes a string of genes with highest expression. Then the query is written to the output file.

## Part 2

Similar to part 4 part 1, the input is read and put into the bags. Then it is passed to custom UDF. Then the query is flattened into genes. Then, using UDF, the values are calculated and written to the output file.

## Part 3

The pig portion does a cross operation on the input and filters the values for duplicates. The UDF then takes care of the cross product calculations with a foreach generate statement from pig. Result is written to output.