Variables and Datasets

train- original training dataset

test- original testing dataset

train$a- total memory(initial used memory + initial free memory) before running a particular query)

train$b- total memory(final used memory + final free memory) after running a particular query

df$c- a new variable that defines the amount of memory each query takes for those queries which gives TRUE for garbage collector

df- dataset that is formed after segregating that part of the dataset that has gcRun==TRUE

df1- a dataset formed after taking a subset of those queries which give FALSE for garbage collector

df1$c- a new variable that defines the amount of memory each query takes for those queries which gives FALSE for garbage collector

train$c- a new variable that defines the amount of memory each query takes for all the training cases

thres- Threshold for classification of garbage collector (thres is the minimum of the sum of initially used memory and the memory of each query)

tok- It is subset of query token(for each iteration tok contains only 1 query value and is used to determine best optimal value of query memory for a particular query), only the initial values of each token is taken as best optimal query memory (since the tokens are initially used sequentially in the beggining, there would be less errors as we move down the dataset our unexplained error increases)

model1 - random forest model applied for regression and final used memory as dependent variable with the suitable parameters and independent variables taken

test1- a testing dataset which contains all the fill values of 4 variables i.e. Initial used memory, Initial free memory, memory value of each query and gcRun.

a- assumed total memory of the container is initial Used value + intial Free Value of the testing dataset before any query was run

Models Applied :

Linear Regression – Applied this model initially, but accuracy was limited so opted for other complex models.

Random Forest – The model we applied for regression analysis is random forest. We applied this model and using suitable parameters and choice of independent variables for predicting our dependent variable (finalUsedMemory).

Future Work:

Xgboost/Neural Networks – Due to time constraints we were unable to apply these models. It is an important model, which can surely give better results.

Assumptions :

The total memory of the container was used as free memory + used memory in test data-frame before any query was run; Both before and after query were always same, and whatever error were negligible.

In finding memory of each query, we took the initial values of query as they were sequentially entered and later down the datasets, our unexplained error increases.

For Classification of gcRun to TRUE and FALSE, we used a threshold value defined by the minimum of the sum of initialUsedmemory and query memory. In test datasets after predicting the initialUsedMemory through regression models and adding respective query memory , if it exceeded threshold then gcRun was taken TRUE and vice-versa.