# First TRUE Homework Assignment (5% of grade)

Due on Wednesday, April 15 at 5 pm.

For each assignment, deliver the following in folder /home/public/assignment (you should have write permissions)

1. Your java source code file: name the file lastname_exercisex.java where x is either 1 or 2
2. One output file from a reducer: name the file lastname_exercisex.txt where x is either 1 or 2.

## There should be two files per student per exercise.

1. Maximum temperature per year: You have two files with temperature readings from various weather stations (if you put them into the same folder in hdfs you call your mapreduce job with that folder, it's going to automatically read both files and pass all rows from both files to the mappers).

   You need to write a mapreduce job in java that will calculate the maximum temperate for each year. The output should be pairs (year,temperature).

   The data is from the NOAA web site.
   Example record: 0029029070999991901010106004+64333+023450FM-12+000599999V0202701N015919999999N0000001N9-00781+99999102001ADDGF108991999999999999999999

   Year = positions at 15-19
   In this example: 1901
   Temperature = positions at 87-92
   In this example: -0078
   If Temperature = 9999, it should be interpreted as missing value.
   The temperature quality is in position 92-93. If it is in the range {0,1,4,5,9}, then the temperature reading is accurate and satisfactory.
   The data set is available in /home/public/course
2. The second data set comes from IBM and it is a machine learning classification problem. It's in the csv format. Your task is to get the average value of the fourth column per every different combination of columns 30,31,32,32 among all records with the last column equal to 'false'
   Your result should look like:
   1,0,1,1, average value of column 4
   0,0,1,0, average value of column 4
   Etc: for all possible combinations of fields 30,31,32,33
   The data is available in /home/public/course/ibm.csv
   The result is equivalent to the sql query: select ave(u[4]), u[30],u[31],u[32],u[33] from u group by u[30],u[31],u[32],u[33] where u[last column] = 'false' (u[i] is the i'th column)

IMPORTANT: copy the input files from /home/public/course directly into hdfs (and not into your home directory on the local filesystem)

To create the jar file, copy /home/public/course/build.xml into your local home directory. Edit build.xml by replacing “zzzzzzz” with the name of your main class. Then issue ‘ant’ which will create the jar file. (Alternatively you can use maven if you prefer with the pom file available in the same folder.)