Spring 2017, MIS 573 – Practical Big Data Analytics

Group Exercise 6

R tibble, data.table, and in-database computing

1. Please load the given dataset “adult.csv” and convert it into *tibble* and *data.table*. Assign the following column names to the variables.

```
colnames(adult_df)=
c("age","workclass","fnlwgt","education","education_num","marital_status",
"occupation","relationship","race",  "sex","capital_gain",
"capital_loss","hours_per_week","native_country","salary")
```

2. Replace the following SQL code in *sqldf*()

```
sqldf("select,  education,race,sex,salary, count(salary) from adult_df
  where age between 20 and 30
  group by education, race, sex,salary
  having count(salary) > 200
  order by count(salary)")
```

with

   a. data manipulation functions of data.table
   b. pipe operators ,%>%, on tibbles

3. Connect to the given SQLite file/database “movies.sqlite” and load the table “data” as remote tibble. Use any data piping functions to clean the data if you’d like. Fit a simple regression tree model with the target variable *rating*. Note that all data manipulations must be done in the database (no data in local R environment).