**Homework 4B**

Due: 4/17/2016

In this homework you will learn how to use Pig Latin.

To connect to server outside from college network first connect to cs1.utdallas.edu server and then ssh the cs6360.utdallas.edu using **ssh <Netid>@cs6360** command. e.g ssh gga110020@cs6360

**Instructions to work:**

1) Create and use only one directory with net id per user on the cluster. (e.g /yournetid)

2) Do not modify/add files in the **/Spring-2016-input/** directory or other users directories (all logs are tracked)

**Connecting to Pig on the cluster.**

Login to cs6360 as stated in previous homework.

Run the command below.

pig

If everything is OK, you should be able to see “grunt>” in your shell.

**Dataset**

We will use the datasets located under **/Spring-2016-input/** in the HDFS in the Programming/Master Node CS6360.utdallas.edu. Please use this folder and don’t copy/modify any other folder on the server. All dataset files are **single-colon (:)** separated.

**movies.dat**: **MovieID:Title:Genres**

**ratings.dat**: **UserID:MovieID:Rating:Timestamp**

**users.dat: UserID:Gender:Age:Occupation:Zip-code**

**Pig Latin**

**Q1:**

Using Pig Latin script, list the unique userid of male users whose age between 20 - 40 and who has rated the lowest rated Comedy AND Drama movies. (You should consider all movies that has Comedy **AND** Drama both in its genre list.)

Print only users whose zip starts with **7**. Consider average rating to calculate the lowest rated movies. While finding the Comedy and Drama movies, you should count all users not only the male users.

**Q2:**

Using Pig Latin script, Implement co-group command on MovieID for the datasets **ratings** and **movies.** Print first 5rows.

**Q3:**

Write a UDF(User Define Function) FORMAT\_GENRE in Pig which basically formats the genre in movie in the following:

Before formatting: Children's

After formatting: 1) Children's **<NetId>**

Before formatting: Animation|Children's

After formatting: 1) Children's & 2) Animation **<NetId>**

Before formatting: Children's|Adventure|Animation

After formatting: 1) Children's, 2) Adventure & 3) Animation **<NetId>**

Using Pig Latin script, use the FORMAT\_GENRE function on movies dataset and print the movie name with its genre(s).