LAB 5
APACHE PIG LAB – PART - 2

Submit the pig script files for each question along with only the output files. Paste screenshots wherever applicable.

**You've been given a sample Movies Data set.
The details of the files and schema are as follows:**

movies.csv	List of 9000+ movies and their details	{movieId, title, genres}
users.txt	List of 900+ users and their details	{userId, age, gender, occupation, zipCode}
ratings.txt	~2M file with movie rating details	{userId, movieId, rating, timestamp}

Note that in the movies.csv file, you might find a "comma (,)" in some of the titles.
Also, the column *genres* have multiple values in it for one movie which are separated by a pipe symbol (|)[1]. You'll need to take care of them properly!

Now let's do some analysis on this real-world data set using Pig in Hadoop mode (not local mode). But for testing purposes, you can try first in local mode as it's faster.

---

[1] Those with database experience will notice that this is a violation of the first normal form as defined by E.F. Codd. This intentional denormalization of data is very common in OLAP systems in general, and in large data-processing systems such as Hadoop in particular. RDBMS systems tend to make joins common and then work to optimize them. In systems such as Hadoop, where storage is cheap, and joins are expensive, it is generally better to use nested data structures to avoid the joins.

Homework Questions

ℹ

1) [2] **How many male lawyers are listed in the users.txt file?**
Write the complete pig script and the final count of male lawyers.

2) [2] **What is the userId of the oldest female lawyer?**
Write the complete pig script and the userId of the oldest female lawyer.

3) [3] **How many movies are there whose title start with letter "A" or "a"? Show the count of these movies by genre.**
If a movie is both Action and Comedy then you'll be counting it twice in both the Action and Comedy genres.

Example Output ->

```
part-r-00000  X
(no genres listed)      1
Action   78
Adventure        68
Animation        31
Children         50
Comedy   179
Crime    45
Documentary      26
```

4) [3] **Display a list of top 20 highest rated (rating=5) "Adventure" movies sorted by movieId.**
The sample output file should look something like this:

movieId genres   rating   title

```
part-r-00000  X
1     Adventure    5    Toy Story (1995)
2     Adventure    5    Jumanji (1995)
8     Adventure    5    Tom and Huck (1995)
10    Adventure    5    GoldenEye (1995)
13    Adventure    5    Balto (1995)
15    Adventure    5    Cutthroat Island (1995)
29    Adventure    5    City of Lost Children, The (Cité des enfants perdus, La) (1995)
44    Adventure    5    Mortal Kombat (1995)
53    Adventure    5    Lamerica (1994)
```

5) [Bonus - 1] **Modify the pig script in Q4 above so that the output file will now show the "header" for each tab separated field.**

Example o/p =>

```
part-r-00000  X
MovieId Genre    Rating  Title
1       Adventure    5       Toy Story (1995)
2       Adventure    5       Jumanji (1995)
8       Adventure    5       Tom and Huck (1995)
10      Adventure    5       GoldenEye (1995)
13      Adventure    5       Balto (1995)
```

You might want to take a look at CSVExcelStorage for how to add header line to output file.

More Help here!