

Lab 5 – Pig-I

Submit your *own work* on time. No credit will be given if the lab is submitted after the due date. Follow the instructions completely.

For all the questions below, Submit the pig script file along with your input files and output file. Paste screenshots wherever applicable.

1) [1] Producing Word Count in Pig

Create a “.pig” script file and write all the commands needed to run word count example in Pig. Create your own input file.

2) [2] Join in Pig – Top 5 most visited sites

Create some sample data for “users.csv” and “pages.csv” files as discussed in today’s lecture.

Find the top 5 most visited sites by users aged between 18 - 25.

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

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

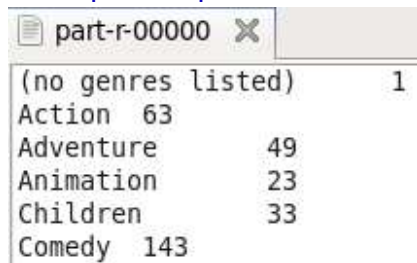
Note that in the movies.csv file, the column *genres* have multiple values in it for one movie which are separated by a pipe symbol (|)¹.

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.

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.

- 3) [1] How many male lawyers are listed in the users.txt file? (write the complete pig script and the final count of male lawyers)
- 4) [2] What is the userId of the oldest male lawyer? (write the complete pig script and the userId of the oldest male lawyer)
- 5) [2] How many movies are there whose title start with letter "A" or "a"? Show the count of these movies by genre.

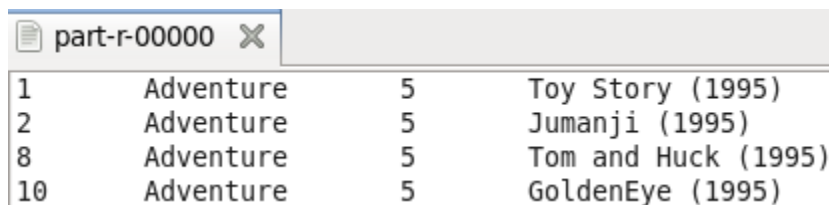
Example Output:



(no genres listed)	1
Action	63
Adventure	49
Animation	23
Children	33
Comedy	143

- 6) [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
---------	--------	--------	-------



1	Adventure	5	Toy Story (1995)
2	Adventure	5	Jumanji (1995)
8	Adventure	5	Tom and Huck (1995)
10	Adventure	5	GoldenEye (1995)

- 7) [2] Out of these highest rated top 20 movies found in Q6, how many times male programmers have watched these movies?

Hint: Answer is more than 100!

- 8) [2] Modify the above pig script 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!](#)