BIG DATA –Spring 2018

ASSIGNMENT 3

Due Date: 8th March 2018 till 10:00 PM on google classroom.

Upload the Source code and the output file on Google Classroom with your roll number.

INPUT FILE

The input file is almost same as the Assignment 2 except the name of the studio that produced the movie is also added at the end. Now Input File consists of the name of the Movie, its release Year, the movie rating (on a scale of 1-5) and the studio name.

Input File

```
The Dark Dragon 1988 3.5 FOX
Evil Witch 1988 2 HBO
Dark Lord 1988 3 HBO
SuperMan 1992 4.5 FOX
Twister 1988 3.5 WARNER
Tornado 2010 4 FOX
The Rise of Aron 2010 2.5 FOX
Jumangi 1988 4 WARNER
Snow Queen 1988 2.5 HBO
Red Riding Hood 2010 4.5 HBO
Lords 2010 3 WARNER
```

QUESTION

Write MapReduce program using **Pair's Approach** for the following task: For each year, list the percentage of the movies produced by each studio.

Sample Output for above Input in the format (Year, Studio -> percentage of the movies produce by that studio in given year)

```
1988, FOX → 16%
1988, HBO → 50%
1988, WARNER → 33%
1992, FOX → 100%
2010, FOX → 50%
2010, HBO → 25%
2010, WARNER → 25%
```

Hint:

Percentage of the movies produced by each studio = $\frac{\text{No. of movies produced by the Studio in a given year Y}}{\text{Total no. of movies produced in that year Y}}$

This problem is somewhat similar to relative frequency word co-occurrence problem discussed in the class. *You will have to provide **Mapper, Reducer, Partitioner and Comparator.***

Hints:

1. *Consult book Hadoop the definitive guide for comparator details (Chapter 5, pg 136 & onwards)*
2. *Have a look at http://codingjunkie.net/order-inversion/*