COMP9313 22T3 Project 1 (12 marks)

Problem statement:

In this project, your task is to perform data analytics over a dataset of online social networks using **MRJob**.

Input files:

The dataset contains users' check-in history, in which each record is in format of "userID, locID, check_in_time", where userID (string type) is the ID of a user, locID (string type) is the ID of a location, and check_in_time is the timestamp of the user's check-in at this location. A sample file is like below:

```
u1,l1,t1
u1,l1,t2
u1,l2,t3
u2,l1,t4
u2,l3,t5
u3,l2,t6
u3,l2,t7
u3,l3,t8
```

This small sample file can be downloaded at:

https://webcms3.cse.unsw.edu.au/COMP9313/22T3/resources/81827

Problem desciption:

We denote the number of check-ins at location loc_i by user u_j as $n_{loc_i}^{u_j}$ and the number of check-ins from u_j as n_{u_j}. Thus, $n_{u_j} = \sum_{loc_i \in L_{u_j}} n_{loc_i}^{u_j}$, where L_{u_j} is the set of locations visited by u_j.

The probability that u_j checked-in at loc_i is computed as $prob_{loc_i}^{u_j} = \frac{n_{loc_i}^{u_j}}{n_{u_j}}$. Your task is to compute $prob_{loc_i}^{u_j}$ for each user at each location which has be visited by this user.

Output format:

Store the result in HDFS in format of: "loc_i\t$u_j$, $prob_{loc_i}^{u_j}$ ". The results are first sorted by location IDs in **ascending** order, and then by the user's check-in probabilities in **descending** order. If two users have the same probability, sort them by their IDs in **ascending** order.

For example, given the above data set, the output is (there is no need to remove the quotation marks which are generated by MRJob):

```
"l1"  "u1,0.6667"
"l1"  "u2,0.5"
"l2"  "u3,0.6667"
"l2"  "u1,0.3333"
"l3"  "u2,0.5"
"l3"  "u3,0.3333"
```

The entire output could be checked at:

https://webcms3.cse.unsw.edu.au/COMP9313/22T3/resources/81828

Code format:

Please name your python file as "project1.py" and compress it in a package named "zID_proj1.zip" (e.g. z5123456_proj1.zip).

Command of running your code:

We will use more than one reducer to test your code. Assuming we use 2 reducers, we will use the following command to run your code:

```
$ python3 project1.py -r hadoop hdfs_input -o hdfs_output --jobconf
mapreduce.job.reduces=2
```

In this command, `hdfs_input` is the input path on HDFS, and `hdfs_output` is the output folder on HDFS.

The code template can be downloaded from:

https://webcms3.cse.unsw.edu.au/COMP9313/22T3/resources/81830

Warning: Please ensure that the code you submit can be compiled. Any solution that has compilation errors will receive no more than 4 points.

Marking Criteria:

Your source code will be inspected and marked based on readability and ease of understanding. The documentation (comments of the codes) in your source code is also important. Below is an indicative marking scheme:

Result correctness: 6
Algorithm design (the use of design patterns learned to reduce memory consumption and to improve efficiency): 5
Code structure, Readability, and Documentation: 1

Submission:

Deadline: Sunday 9th Oct 11:59:59 PM

You can submit through Moodle:

If you submit your assignment more than once, the last submission will replace the previous one. To prove successful submission, please take a screenshot as assignment submission instructions show and keep it by yourself. If you have any problems in submissions, please email to siqing.li@unsw.edu.au.

Late submission penalty

5% reduction of your marks for up to 5 days

Plagiarism:

The work you submit must be your own work. Submission of work partially or completely derived from any other person or jointly written with any other person is not permitted. The penalties for such an offence may include negative marks, automatic failure of the course and possibly other academic discipline. Assignment submissions will be examined manually.

Relevant scholarship authorities will be informed if students holding scholarships are involved in an incident of plagiarism or other misconduct.

Do not provide or show your assignment work to any other person - apart from the teaching staff of this subject. If you knowingly provide or show your assignment work to another person for any reason, and work derived from it is submitted you may be penalized, even if the work was submitted without your knowledge or consent.