COMP9313 2017s2 Assignment

Question 1. MapReduce (5 pts)

Assume that in an online shopping system, a huge log file stores the information of each transaction. Each line of the log is in format of "userID\tproduct\t price\t time". Your task is to use MapReduce to find out the top-5 most expensive products purchased by each user in 2016.

You only need to write down the pseudo code (Mapper, Reducer, and optionally Combiner and Partitioner) to describe the algorithm (assume only one reducer). Note that the efficiency and scalability of your solution will be evaluated.

Question 2. MinHash (5 pts)

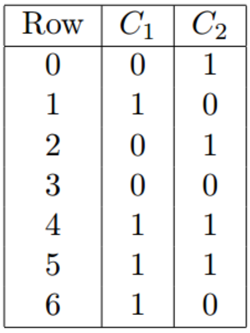
We want to compute min-hash signature for two columns, C_1 and C_2 using two pseudo-random permutations of columns using the following function:

$$h_1(n) = 3n + 2 \bmod 7$$

$$h_2(n) = 2n - 1 \bmod 7$$

Here, n is the row number in original ordering. Instead of explicitly reordering the columns for each hash function, we use the implementation discussed in class, in which we read each data in a column once in a sequential order, and update the min hash signatures as we pass through them.

Complete the steps of the algorithm and give the resulting signatures for C_1 and C_2.



Row	C_1	C_2
0	0	1
1	1	0
2	0	1
3	0	0
4	1	1
5	1	1
6	1	0

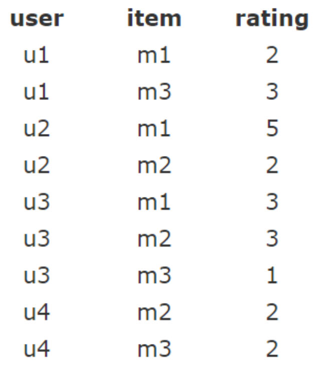
Question 3. Streaming Data (5 pts)

Suppose we are maintaining a count of 1s using the DGIM method. We represent a bucket by (i, t), where i is the number of 1s in the bucket and t is the bucket timestamp (time of the most recent 1).

Consider that the current time is 200, window size is 60, and the current list of buckets is: (16, 148) (8, 162) (8, 177) (4, 183) (2, 192) (1, 197) (1, 200). At the next ten clocks, 201 through 210, the stream has 0101010101. What will the sequence of buckets be at the end of these ten inputs?

Question 4. Collaborative Filtering (5 pts)

Consider four users u1, u2, u3 and u4, and three movies m1, m2, and m3. The ratings of movies from the users are as below:



user	item	rating
u1	m1	2
u1	m3	3
u2	m1	5
u2	m2	2
u3	m1	3
u3	m2	3
u3	m3	1
u4	m2	2
u4	m3	2

(a) Estimate the rating of u1 to m2 using the user-user collaborative filtering method (adopt the cosine similarity measure to compute the user similarities).

(b) Estimate the rating of u1 to m2 using the item-item collaborative filtering method (adopt the cosine similarity measure to compute the item similarities).

Submission:

Deadline: Sunday 5th Nov 09:59:59 PM

Please provide your solutions to these questions in a pdf file named as “answers.pdf”. Log in any CSE server (williams or wagner), and use the give command below to submit your solutions:

```
$ give cs9313 assignment5 answers.pdf
```

Or you can submit through:
https://cgi.cse.unsw.edu.au/~give/Student/give.php

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.

Late submission penalty

You will receive zero marks for this assignment.

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.