

COMP9313 2017s2 Assignment

Chia-Jui Chuang
z5095972

Question 1

class Mapper

```
// the priority queue has the highest price at the front
hMap = new HashMap<userID, PriorityQueue<Pair<Price, Product>>>;
method map(key k, line l)
    tokens = l.split("\\t");
    if(!hMap.containsKey(tokens.userID))
        hMap.insert(userID, new PriorityQueue<>);
    pq = hMap.get(userID);
    pq.insert(Pair(tokens.price, tokens.product));
    if(pq.size() > 5)
        pq.removeBack(); // This is to ensure there are at most 5 items

method cleanup()
    for each userID in hMap
        pq = hMap.get(userID)
        for each pair in pq // there should be at most 5 items
            emit(userID, pair)
```

class Reducer

```
method reduce(userID, pairs[p1, p2, ...])
    pq = new PriorityQueue<Pair<Price, Product>>;
    for each pair in pairs
        pq.insert(pair)
    int max = pq.size() < 5 ? pq.size() : 5;
    for(int i = 0; i < max; ++i)
        pair = pq.front();
        emit(userID, pair.product + "\\t" + pair.price);
        pq.removeFront();
```

Note:

Since only one reducer will be used as specified in the assignment specs, no partitioner has been setup here. However, if we were to scale this mapreduce job, I would add a partitioner to balance the load on each reducer.

If the key-value pairs could not fit into memory, I would use a separate combiner instead of using the in-mapper combining technique. Nonetheless, the mapper has been setup to remove the extra pairs, in this case when the size is greater than 5, as part of the mapping process, the likelihood of using a separate combiner is fairly low.

Question 2

$$h1(n) = 3n + 2 \bmod 7$$

$$h2(n) = 2n + 1 \bmod 7$$

Row	C1	C2	h1	h2
0	0	1	2	6
1	1	0	5	1
2	0	1	1	3
3	0	0	4	5
4	1	1	0	0
5	1	1	3	2
6	1	0	6	4

Initialisation

	C1	C2
h1	∞	∞
h2	∞	∞

Row 0

	C1	C2
h1	∞	2
h2	∞	6

Row 1

	C1	C2
h1	2	2
h2	1	6

Row 2

	C1	C2
h1	2	1
h2	1	3

Row 3

	C1	C2
h1	2	1
h2	1	3

Row 4

	C1	C2
h1	0	0
h2	0	0

Row 5

	C1	C2
h1	0	0

h2	0	0
----	---	---

Row 6

	C1	C2
h1	0	0
h2	0	0

Signature Matrix:

$$\text{SIM}(C1, C2) = 1$$

Jaccard Similarity:

$$\text{SIM}(C1, C2) = 1/3$$

Question 3

1) @ 202

(16, 148) (8, 162) (8, 177) (4, 183) (2, 192) (2, 200) (1, 202)

2) @ 204

(16, 148) (8, 162) (8, 177) (4, 183) (2, 192) (2, 200) (1, 202) (1, 204)

3) @ 206

(16, 148) (8, 162) (8, 177) (4, 183) (4, 200) (2, 204) (1, 206)

4) @ 208

(16, 148) is dropped in this step since it is outside the window.

(8, 162) (8, 177) (4, 183) (4, 200) (2, 204) (1, 206) (1, 208)

5) @ 210

(8, 162) (8, 177) (4, 183) (4, 200) (2, 204) (2, 208) (1, 210)

Question 4

	m1	m2	m3
u1	2		3
u2	5	2	
u3	3	3	1
u4		2	2

(a) user-user

$$\text{sim}(u1, u2) = \frac{(2 \times 5 + 3 \times 0)}{\sqrt{(2^2 + 3^3)} \times \sqrt{(5^2)}} = 0.5547$$

$$\text{sim}(u1, u3) = \frac{(2 \times 3 + 3 \times 1)}{\sqrt{(2^2 + 3^3)} \times \sqrt{(3^2 + 1^1)}} = 0.7894$$

$$\text{sim}(u1, u4) = \frac{(3 \times 2)}{\sqrt{(2^2 + 3^3)} \times \sqrt{(2^2)}} = 0.8321$$

$$r_{u1m2} = \frac{(0.5547 \times 2 + 0.7894 \times 3 + 0.8321 \times 2)}{(0.5547 + 0.7894 + 0.8321)} = 2.3627$$

(b) item-item

$$\text{sim}(m2, m1) = \frac{(2 \times 5 + 3 \times 3)}{\sqrt{(2^2 + 3^3 + 2^2)} \times \sqrt{(5^2 + 3^2)}} = 0.7903$$

$$\text{sim}(m2, m3) = \frac{(3 \times 1 + 2 \times 2)}{\sqrt{(2^2 + 3^3 + 2^2)} \times \sqrt{(1^2 + 2^2)}} = 0.7593$$

$$r_{u1m2} = \frac{(0.7903 \times 2 + 0.7593 \times 3)}{(0.7903 + 0.7593)} = 2.4900$$