COMP9313 - Project 5

Changxun Fan (z5006334)

Quesiton 1

The following code of for computing the relative frequency is problematic. Describe how you can fix it.

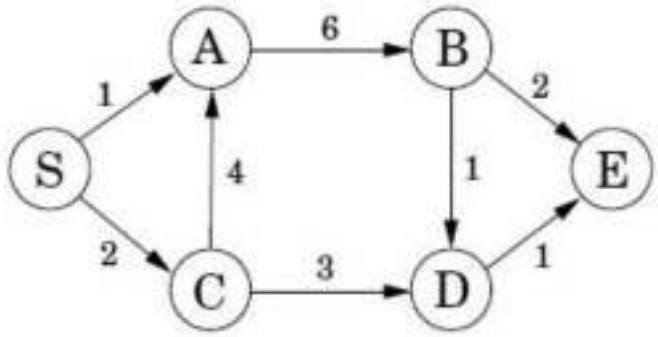
```
class Mapper
    method Map(docid a, doc d)
        for all term w in doc d do
            for all term u in Neighbor(w) do
                Emit(pair(w, u), count 1)
                Emit(pair(w, *), count 1)

class Reducer
    curMarginal := 0
    method Reduce(pair p, counts [c1, c2, ...])
        s := 0
        for all count in counts[c1, c2, ...] do
            s := s+ c
        // not is total count
        if(!p.contains(*))
            Emit(p, s/curMarginal)
        else
            curMarginal := s
```

Answer: In the mapper, for each word u in the neighburs of w, emit the word pair (w, u) and pair(w, *) one time respectively. (w, u) is used to count the word frequency and (w, *) is used to count the total word frequency. So in the reducer, we have to find the total word frequency first, and that must contain character "*".

Question 2

Given the following graph, assume that you are using the single shortest path algorithm to compute the shortest path **from node S to node E**. Show the output of the mapper (sorted results of all mappers) and the reducer (only one reducer used) in each iteration (including both the distances and the paths).



Answer:

- Input file:
  - S -> 0 | A:1, C:2
  - A -> ∞ | B:6
  - B -> ∞ | D:1, E:2
  - C -> ∞ | A:4, D:3
  - D -> ∞ | E:1
  - E -> ∞
- Iteration1
  - Map:
    - Read: S -> 0 | A:1, C:2
    - Emit: (A, 1), (C, 2), and the adjacency list (S, A:1, C2)
  - Reduce:
    - Receives: (A,1), (C, 2), (S, <0, (A:1, C:2)>)
    - Emit:
      - A -> 1 |B:6
      - C -> 2 | A:4, D:3
      - S -> 0 | A:1, C:2

- Iteration2
  - Map:
    - Read: A -> 1 |B:6
    - Emit: (B, 7), (A, <1, (B:6)>)
    - Read: C -> 2 | A:4, D:3
    - Emit: (A, 6), (D, 5), (C, <2, (A:6, D:5)>)
  - Reduce:
    - Receives: ((B, 7), (A, <1, (B:6)>)), ((A, 6), (D, 5), (C, <2, (A:6, D:5)>))
    - Emit:
      - A -> 1 | B:6
      - B -> 7 | D:1, E:2 (update)
      - C -> 2 | A:4, D:3
      - D -> 5 | E:1 (update)
      - S -> 0 | A:1, C:2
- Iteration3
  - Map:
    - Read: A -> 1 |B:6
    - Emit: (B, 7), (A, <1, (B:6)>)
    - Read: B -> 7 | D:1, E:2
    - Emit: (D, 8), (E, 9), (B, <7, (D:1, E:2)>)
    - Read: C -> 2 | A:4, D:3
    - Emit: (A, 6), (D, 5), (C, <2, (A:6, D:5)>)
    - Read: D -> 5 | E:1
    - Emit: (E, 6), (D, <5, (E:1)>)
    - Read: S -> 0 | A:1, C:2
    - Emit: (A,1), (C, 2), (S, <0, (A:1, C:2)>)
  - Reduce:
    - Receives: ((B, 7), (A, <1, (B:6)>)), ((D, 8), (E, 9), (B, <7, (D:1, E:2)>)), ((A, 6), (D, 5), (C, <2, (A:6, D:5)>)), ((E, 6), (D, <5, (E:1)>)), ((A,1), (C, 2), (S, <0, (A:1, C:2)>))
    - Emit:
      - A -> 1 | B:6
      - B -> 7 | D:1, E:2

        - C -> 2 | A:4, D:3
        - D -> 5 | E:1
        - E -> 6 (update)
        - S -> 0 | A:1, C:2
- Iteration4
    - Map:
        - Read: A -> 1 |B:6
        - Emit: (B, 7), (A, <1, (B:6)>)
        - Read: B -> 7 | D:1, E:2
        - Emit: (D, 8), (E, 9), (B, <7, (D:1, E:2)>)
        - Read: C -> 2 | A:4, D:3
        - Emit: (A, 6), (D, 5), (C, <2, (A:6, D:5)>)
        - Read: D -> 5 | E:1
        - Emit: (E, 6), (D, <5, (E:1)>)
        - Read: E -> 6
        - Emit:
        - Read: S -> 0 | A:1, C:2
        - Emit: (A,1), (C, 2), (S, <0, (A:1, C:2)>)
    - Reduce:
        - Receives: ((B, 7), (A, <1, (B:6)>)), ((D, 8), (E, 9), (B, <7, (D:1, E:2)>)), ((A, 6), (D, 5), (C, <2, (A:6, D:5)>)), ((E, 6), (D, <5, (E:1)>)), ((A,1), (C, 2), (S, <0, (A:1, C:2)>))
        - Emit:
            - A -> 1 | B:6
            - B -> 7 | D:1, E:2
            - C -> 2 | A:4, D:3
            - D -> 5 | E:1
            - E -> 6
            - S -> 0 | A:1, C:2
- No Update in Iteration4, terminate that.

Question 3

Suppose we are maintaining a count of 1s using the DGIM method. We represent a bucketby (i, t), where i is the number of 1s in the bucket and t is the bucket timestamp (time of themost 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, 201through 210, the stream has 0101010101. What will the sequence of buckets be at the endof these ten inputs?

Answer:

- the current list of bucket is:

$$(16,148)(8,162)(8,177)(4,183)(2,192)(1,197)(1,200)$$

- at 201 clock, 0 enters into the window, no update.

$$(16,148)(8,162)(8,177)(4,183)(2,192)(1,197)(1,200)$$

- at 202 clock, 1 enters into the window, update.

$$(16,148)(8,162)(8,177)(4,183)(2,192)(1,197)(1,200)(1,202)$$
$$\Downarrow$$
$$(16,148)(8,162)(8,177)(4,183)(2,192)(2,200)(1,202)$$

- at 203 clock, 0 enters into the window, no update.

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

- at 204 clock, 1 enters into the window, update .

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

- at 205 clock, 0 enters into the window, no update.

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

- at 206 clock, 1 enters into the window, update.

$$(16, 148)(8, 162)(8, 177)(4, 183)(2, 192)(2, 200)(1, 202)(1, 204)(1, 206)$$
$$\Downarrow$$
$$(16, 148)(8, 162)(8, 177)(4, 183)(2, 192)(2, 200)(2, 204)(1, 206)$$
$$\Downarrow$$
$$(16, 148)(8, 162)(8, 177)(4, 183)(4, 200)(2, 204)(1, 206)$$

- at 207 clock, 0 enters into the window, no update.

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

- at 208 clock, 1 enters into the window, update.

$$(16, 148)(8, 162)(8, 177)(4, 183)(4, 200)(2, 204)(1, 206)(1, 208)$$
$$\Downarrow (208 - 148 + 1) > 60,\ \textit{drop the oldest bucket}$$
$$(8, 162)(8, 177)(4, 183)(4, 200)(2, 204)(1, 206)(1, 208)$$

- at 209 clock, 0 enters into the window, no update.

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

- at 210 clock, 1 enters into the window, update.

$$(8, 162)(8, 177)(4, 183)(4, 200)(2, 204)(1, 206)(1, 208)(1, 210)$$
$$\Downarrow$$
$$(8, 162)(8, 177)(4, 183)(4, 200)(2, 204)(2, 208)(1, 210)$$

So the final list of buckets is:

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

Question 4

Consider three users u1, u2, and u3, and four movies m1, m2, m3, and m4. The users rated themovies using a 4-point scale: -1: bad, 1: fair, 2: good, and 3: great. A rating of 0 means thatthe user did not rate the movie.

The three users' ratings for the four movies are: u1 = (3, 0, 0, -1), u2 = (2, -1, 0, 3), u3 = (3, 0, 3,1)

(i) (3 pts) Which user has more similar taste to u1 based on cosine similarity, u2 or u3? Show detailed calculation process.

(ii) (2 pts) User u1 has not yet watched movies m2 and m3. Which movie(s) are you going torecommend to user u1, based on the user-based collaborative filtering approach? Justifyyour answer.

Answer:

(i)

	m1	m2	m3	m4
u1	3	0	0	-1
u2	2	-1	0	3
u3	3	0	3	1

$$similarity(r_x, r_y) = Cosine(r_x, r_y) = \frac{r_x \cdot r_y}{||r_x|| \cdot ||r_y||}$$

According to the above equation, we can calculate that:

$$u1 * u2 = 3 * 2 + 0 * (-1) + 0 * 3 + (-1) * 3 = 3$$
$$u1 * u3 = 3 * 3 + 0 * 0 + 0 * 3 + (-1) * 1 = 8$$
$$u2 * u3 = 2 * 3 + (-1) * 0 + 0 * 3 + 3 * 1 = 9$$
$$||u1|| = \sqrt{3^2 + 0^2 + 0^2 + (-1)^2} = \sqrt{10}$$
$$||u2|| = \sqrt{2^2 + (-1)^2 + 0^2 + 3^2} = \sqrt{14}$$
$$||u3|| = \sqrt{3^2 + 0^2 + 3^2 + 1^2} = \sqrt{19}$$

$$cos(u1, u2) = \frac{u1 \cdot u2}{||u1|| \cdot ||u2||} = 0.2535$$
$$cos(u1, u3) = \frac{u1 \cdot u3}{||u1|| \cdot ||u3||} = 0.5803$$
$$cos(u2, u3) = \frac{u2 \cdot u3}{||u2|| \cdot ||u3||} = 0.6529$$

so sim(u1, u3) > sim(u1, u2), the user u3 has more similar taste to u1 based on cosine similarity.

(2)

$$similarity(r_x, r_y) = Person(r_x, r_y) = \frac{\sum_{s \in S_{xy}}(r_{xs} - \overline{r}_x)(r_{ys} - \overline{r}_y)}{\sqrt{\sum_{s \in S_{xy}}(r_{xs} - \overline{r}_x)^2 (r_{ys} - \overline{r}_y)^2}}$$

According to the above equation, we can calculate that:

$$\overline{r_{u_1}} = \frac{3 + (-1)}{2} = 1$$
$$\overline{r_{u_2}} = \frac{2 + 3}{2} = 2.5$$
$$\overline{r_{u_3}} = \frac{3 + 1}{2} = 2$$

so we can calculate the similarity between these users using the above information

- u1 and u2

$$S_{12} = m_1, m_4$$
$$sim(u_1, u_2) = \frac{(3-1) * (2-2.5) + (-1-1) * (3-2.5)}{\sqrt{(3-1)^2 + (-1-1)^2} * \sqrt{(2-2.5)^2 + (3-2.5)^2}} = -1$$

- u1 and u3

$$S_{13} = m_1, m_4$$
$$sim(u_1, u_3) = \frac{(3-1) * (3-2) + (-1-1) * (1-2)}{\sqrt{(3-1)^2 + (-1-1)^2} * \sqrt{(3-2)^2 + (1-2)^2}} = 1$$

According to the Pearson Correlation Coefficient, user 3 has the highest correleation with user 1, based on that we can recommend movie 3 to user1.