CS4225/CS5425 BIG DATA SYSTEMS FOR DATA SCIENCE

Tutorial 4: Streaming and Graphs

1. In Spark Structured Streaming, why we need to specify a checkpoint location?

Answer: to save the progress information of a stream query, i.e. what data has been successfully processed. Upon failure, this info is used to restart the failed query exactly where it left off.

2. In Spark Structured Streaming, we are using below codes to collect the streaming data from sensor readings.

```
(sensorReadings
.withWatermark("eventTime", "5 minutes")
.groupBy("sensorId", window("eventTime", "10 minutes", "5 minutes"))
.count())
```

From 12:05 to 12:10, we have received below events:

Event Table

sensorID	Event Time
id1	12:06
id1	12:08

And at 12:10 the below result table is triggered:

Result Table		
Event Window	**sensorID**	**Count**
11:50-12:00	id1	1
11:55-12:05	id1	1
12:00-12:10	id1	2
12:05-12:15	id1	2

From 12:10 to 12:15, we received three more events per below:

Event Table

sensorID	Event Time
id1	11:54
id1	12:02
id1	12:13

Please provide the result table at 12:15.

Result Table		
Event Window	**sensorID**	**Count**
11:50-12:00	id1	
11:55-12:05	id1	
12:00-12:10	id1	
12:05-12:15	id1	
12:10-12:20	id1	
11:50-12:00	id1	

Answer: watermark = max event time (before 12:10) – watermark delay = 12:08 – 5 min = 12:03

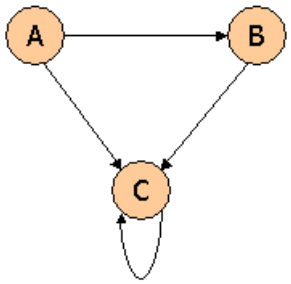
Therefore, intermediate state for 11:50-12:00 is dropped as watermark (12:03) > 12:00. The 11:50-12:00 entry will not be updated, and the rest entries are updated accordingly.

Note: the intermediate state 11:55-12:05 is NOT dropped as watermark (12:03) < 12:05, there may still be events from 12:03 to 12:05, need to be recorded. Though later, an event 12: 02 (< watermark 12:03) show up, since this entry (11:55 – 12:05) is still active we will record the 12:02 event under this entry (11:55 – 12:05) accordingly.

Result Table		
Event Window	**sensorID**	**Count**
11:50-12:00	id1	1
11:55-12:05	id1	2
12:00-12:10	id1	3
12:05-12:15	id1	3
12:10-12:20	id1	1

not update

3. Consider three Web pages with the following links:



Suppose we compute PageRank with a β of 0.7 (note: we assume that the sum of the PageRanks of the three pages must be 1, to handle the problem that otherwise any multiple of a solution will also be a solution). Compute the PageRanks a, b, and c of the three pages A, B, and C, respectively.

The PageRank equation from lecture is:

$$r_j = \sum_{i \to j} \beta \frac{r_i}{d_i} + (1 - \beta) \frac{1}{N}$$

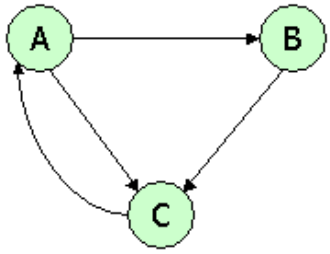
Which translates to:

a = 0.3/3
b = 0.7(a/2) + 0.3/3
c = 0.7(a/2+b+c) + 0.3/3

You can understand this term by term: the 0.3/3 comes from the teleport probability of 0.3, which is divided among 3 nodes since it randomly chooses which node to teleport to. For the other terms like 0.7(a/2), note that a splits its PageRank between b and c, while b gives all of its to c, and c keeps all its own. However, all PageRank is multiplied by 0.7 before being sent, as the probability of taking regular steps (i.e. non-teleport) is 0.7.

Solving the equations: we get a = 0.1. Then, the 2nd equation gives b = 0.7*0.1/2 + 0.1 = 0.135. Finally, since the 3 weights sum up to 1, the remaining weight of 1 – 0.1 – 0.135 = 0.765 must lie with node c (you can check that this also satisfies the last equation).

4. Consider three Web pages with the following links:



Suppose we compute PageRank with β=0.85. Write the equations for the PageRanks a, b, and c of the three pages A, B, and C, respectively.

Following the same procedure as above, we have:

a = .85c + 0.05
b = .425a + 0.05
c = .85b + .425a + 0.05

(i.e. every node receives 0.15 / 3 = 0.05 teleport probability, and a receives all the pagerank from c; b receives half the pagerank from a, and c receives all the pagerank from b and half from a, before applying the teleport probability).

Note that if the question doesn't ask to solve the equations, you don't have to solve them. Also note that it is also fine to use the matrix form of the PageRank equations:

$$M = \begin{bmatrix} 0 & 0 & 1 \\ \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 1 & 0 \end{bmatrix}, N = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}, A = \beta M + (1-\beta)N = \begin{bmatrix} 0.05 & 0.05 & 0.9 \\ 0.475 & 0.05 & 0.05 \\ 0.475 & 0.9 & 0.05 \end{bmatrix}$$

The PageRank equation is then $r = Ar$, where r is PageRank, which could also be written as:

$$a = 0.05a + 0.05b + 0.9c$$

$$b = 0.475a + 0.05b + 0.05c$$

$$c = 0.475a + 0.9b + 0.05c$$

Note that these are equivalent to the equations in the 1st solution (so both are correct answers), since a+b+c=1.

Tutorial Solutions