CS4225/CS5425 BIG DATA SYSTEMS FOR DATA SCIENCE

Tutorial 3: Streaming and Graphs

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

Answer:

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

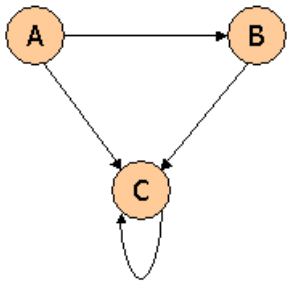
sensorID	Event Time
id1	11:54
id1	12:02
id1	12:13

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:

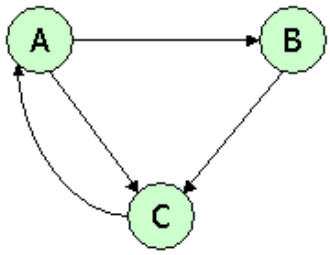
3. Consider three Web pages with the following links:



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.

Answer:

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.

Answer: