CS4225/CS5425 BIG DATA SYSTEMS FOR DATA SCIENCE

Tutorial 4: Test Practice

1. True/False: Personalized PageRank will generate the same ranking for any set of web pages.

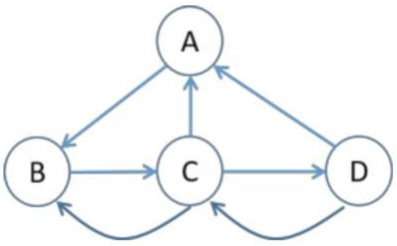
Answer: False. It depends on the teleport set.

2. True/False: In the Spark runtime, RDD cannot reside in the hard disk.

Answer: False. RDD can also be in the disk if out of memory. ("spill to disk")

3. Set up the topic-specific PageRank equations for this graph, with teleport set {A} and β = 0.8 (jump probability = 1−β).

Note: you do not need to solve the equations for r(x).



Answer:

$$r = A \cdot r, \text{ where } A = \beta M + (1-\beta)U$$

A	0	0	⅓	½
B	1	0	⅓	0
C	0	1	0	½
D	0	0	⅓	0

1	1	1	1
0	0	0	0
0	0	0	0
0	0	0	0

4. In HDFS, each chunk is replicated for three times by default. In contrast, in Spark, RDD uses lineage for reliability. What are the major problems if Spark also uses replications for reliability?

Answer: Consumes a lot of memory.

5. Describe a system issue caused by skewed (i.e. highly imbalanced) data distributions on the following systems.

Pregel: computing PageRank on a 1 billion vertex graph, where the degrees of nodes are highly skewed (i.e. some nodes with very large degree)

MapReduce: computing word count, when the frequencies of different keywords is highly skewed (some keywords appearing many times)

Answer: Pregel: High memory requirement in some workers OR: high network I/O for some workers.
MapReduce: High memory requirement in some workers

6. Show pseudocode for the compute() function for the PageRank over vertices algorithm in Pregel / Giraph. You can (if you choose) use the functions: getValue(), setValue(), getNumVertices(), getSuperStep(), getOutEdgeIterator().

Answer:

```
Compute(v, messages):
if getSuperStep() >= 1:
  sum = 0
  for m in messages:
    sum += m
  v.setValue(0.15 / getNumVertices() + 0.85 * sum)
if getSuperStep() < 30:
  sendMsgToAllEdges(v.getValue() / len(getOutEdgeIterator()))
else:
  voteToHalt()
```