University of Virginia

DS 5559: Big Data Analytics

Assignment: Working with GraphFrames

Last Updated: Oct 20, 2019

INSTRUCTIONS

In this assignment, you will run the GraphX code below to answer the questions. The value *None* is used as a placeholder.

TOTAL POINTS: 8

In [ ]:

```
from pyspark.sql import SQLContext
sqlContext = SQLContext(sc)
from graphframes import *

# Vertex DataFrame; contains identifier field "id"
v = sqlContext.createDataFrame([
  ("1", "Adam", "koala"),
  ("2", "Callie", "flamingo"),
  ("3", "Elle", "panda"),
  ("4", "Jacqui", "fox")
], ["id", "name", "favorite_animal"])

# Edge DataFrame; contains source field "src" and destination field "dst"
e = sqlContext.createDataFrame([
  ("1", "2", "dad"),
  ("1", "3", "husband"),
  ("1", "4", "son_in_law"),
  ("2", "1", "daughter"),
  ("2", "3", "daughter"),
  ("2", "4", "granddaughter"),
  ("3", "1", "wife"),
  ("3", "2", "mom"),
  ("3", "4", "daughter"),
  ("4", "1", "mother_in_law"),
  ("4", "2", "grandmother"),
  ("4", "3", "mom")
], ["src", "dst", "relationship"])
```

1) (1 PT) Create a GraphFrame

```
In [1]: g = None
```

2) (1 PT) Show the vertices

```
In [ ]:
```

3) (1 PT) Compute and print the number of grandmother relationships in the graph *g*

```
In [2]:
```

4) (1 PT) Run PageRank for 20 iterations with a reset probability 0.25. Next, print the vertices.

```
In [ ]: results = None
        vertices = None
        print(vertices)
```

PageRank
i. In the cell below, copy the vertex and edge dataframe code
ii. Modify the dataframes and build a new graph to produce *pagerank* values which are not all the same. These values are shown in the *results.pagerank* field

```
In [ ]: # Enter vertex and edge data here
```

5) (1 PT) Enter PageRank code here

```
In [ ]:
```

6) (1 PT) Print the results, showing values that are not all the same

```
In [ ]:
```

7) (2 PTS) Explain your results. Do they make sense?