Assignment: Practicing Graph Analytics in Neo4j With Cypher | Coursera

In order to truly understand how to apply analytics techniques to graph networks, you need to go through the necessary steps to create a query using Cypher, submit that query to Neo4j, and interpret the results. You've seen this -- and maybe even followed along yourself -- in videos. This assignment is a chance to solidify your knowledge by practicing the same tasks on a new data set!

Learning Outcomes

After completing this assignment you will be able to:

1. Create your own Cypher queries by adapting existing queries.
2. Submit queries to Neo4j and interpret the results.
3. Describe the advantages and disadvantages of methods for analyzing graphs using different approaches including path analytics and connectivity analytics.

Introduction

The best way to understand graph analytics and the Cypher query language is to begin with existing query examples and modify them to apply to different datasets. In this assignment you will take existing query examples and adapt them to apply to the example dataset we provided containing the first 50,000 rows of the larger gene-gene association dataset.

Resources

You will need to have Neo4j installed and running in your browser.

You will also need to download the provided zip file containing the dataset(s) for this assignment (see step 1 in 'What you will do' below).

You will need to download the text documents containing supplementary resources for the lectures in this Module.

What you will do

1. Download the gene-gene association dataset by clicking the link below and extracting (unzipping) the contents.

[NOTE: this dataset may take 10-15 minutes to load into Neo4j. Please be patient. We have tried to choose a reasonably large dataset to give you a flavor of working with Big Data but which also will load in a reasonable amount of time.]

The dataset contains fewer rows and columns than the complete dataset available in the combined zip download.

Be sure that the gene_gene_association_50k.csv dataset contains 50,000 rows and three columns of data.

2. Modify an existing script to import the dataset into Neo4j.
In the Importing Data Into Neo4j - Supplementary Resources Reading you were shown how to write a Cypher script to import a dataset of a simple road network graph. We include that script here for your convenience:

```
LOAD CSV WITH HEADERS FROM "file:///C:/coursera/data/test.csv" AS line
MERGE (n:MyNode {Name:line.Source})
MERGE (m:MyNode {Name:line.Target})
MERGE (n) -[:TO {dist:line.distance}]-> (m)
```

[NOTE: replace any spaces in your path with %20, i.e. "percent twenty". In other words, replace a folder named 'my data' with 'my%20data'.Also, If you have problems loading the 50k rows CSV subset, try a 'subset of the subset' by creating a file with only 100 rows. This will load quickly and at help to validate that your script is working. However, you will then need to clear your database before loading the 50k rows CSV. Here are commands to clear your Neo4j database:

```
match (n)-[r]-() delete n, r
match (n) delete n
```

The first row deletes all edges and their corresponding nodes, the second command deletes all nodes with no edges.]
Modify the above script according to the following:

- Load the gene_gene_association_50k.csv (instead of the text.csv),
- Define the node type to be TrialGene,
- Add a Name property to the source node and assign the OFFICIAL_SYMBOL_A column values to it,
- Add a Name property to the target node and assign the OFFICIAL_SYMBOL_B column values to it,
- Define the edge type to be AssociationType,
- Give each edge a property named AssociatedWith and assign the content of the column in the dataset with the heading EXPERIMENTAL_SYSTEM.

[Need help? Review the following video and reading:

3. Perform the following analyses and document your results in order to answer the questions in the accompanying quiz (e.g. You will answer a question for each numbered item below. We suggest you write down that answer as you go along, then enter it in the quiz when you are done):

1. Calculate number of nodes in the graph.

2. Calculate the number of edges in the graph.

3. Calculate the number of loops in the graph.

4. Submit the following query and report the results.

```
match (n)-[r]->(m) where m <> n return distinct n, m, count(r)
```

5. Interpret the results of the query in Step 4 above.

6. Submit the following query and report the results:

```
match (n)-[r]->(m) where m <> n return distinct n, m, count(r) as myCount
order by myCount desc limit 1
```

7. Run the following query and interpret the results:

```
match p=(n {Name:'BRCA1'})-[:AssociationType*..2]->(m) return p
```

[Need help with Questions 1-7? Review the Basic Queries Part 1 reading and hands-on, and Basic Queries Part 2 reading and hands-on]

8. Count how many shortest paths there are between the node named 'BRCA1' and the node named 'NBR1'. *[Need help? Review the Path Analytics reading and hands-on.]*

9. Find the top 2 nodes with the highest outdegree. *[Need help? Review the Connectivity Analytics reading and hands-on.]*

10. Modify one of the Cypher queries we provided and create the degree histogram for the network, then calculate how many nodes are in the graph having a degree of 3. *[Need help? Review the Connectivity Analytics reading and hands-on.]*