Athens University of Economics and Business

MSc in Business Analytics

Assignment 2

Deadline: 4/7/2021

Group assignment (groups of up to 3 people)

The assignment corresponds to 25% of the total grade of the course.

Discussions between groups are recommended but collaborating on the actual solutions is considered cheating and will be reported.

There will be no extension of the assignment deadline!

Professor: Y.Kotidis (kotidis@aueb.gr)

Assistant responsible for this assignment: I.Filippidou (filippidoui@aueb.gr)

Assignment Description: Neo4j Graph database

Dataset:

You are given a subset of the high energy physics theory citation network, which contains authors, articles, journals and citations between articles. In particular, the dataset contains 29555 articles with id, title, year and abstract, 15420 authors with names, 836 journals with names and 352807 citations among papers. You can download the dataset (Citation Dataset) from moodle in csv format.

The dataset files are:

ArticleNodes.csv: Contains info about Article nodes (id, title, year, journal and abstract).

AuthorNodes.csv: Contains article id and the name of the author(s).

Citations.csv: Contains info about citations between articles (articleId,--[Cites]->, articleId).

Property graph model

You are asked to model the data as a property graph by designing the appropriate entities and assigning the relevant labels, types and properties. For your modeling, you need to study the details of all the files that describe the citation network and represent **accordingly all attributes as properties on nodes and edges of a graph**. In your model you should include only the attributes that describe each node and edge type, without repetitions of elements (e.g. same property being displayed on both a node and an edge). Finally, nodes should not be connected when this is not required by the model.

Importing the dataset into Neo4j

Based on your model, **you should create a graph database on Neo4j and load the citation network elements (nodes, edges, attributes).** You can load the dataset directly from the provided csv files, by using either the neo4j browser or the neo4j import tool, or any programming language that is supported by neo4j. To speed up loading and query response times, you could also create proper indexes on your model properties.

Querying the database

After the creation of your database, **you are asked to write and execute the following queries using the Cypher language**.

Queries:

1) Which are the top 5 authors with the most citations (from other papers). Return author names and number of citations.
2) Which are the top 5 authors with the most collaborations (with different authors). Return author names and number of collaborations.
3) Which is the author who has wrote the most papers without collaborations. Return author name and number of papers.
4) Which author published the most papers in 2001? Return author name and number of papers.
5) Which is the journal with the most papers about “gravity” (derived only from the paper title) in 1998. Return name of journal and number of papers.
6) Which are the top 5 papers with the most citations? Return paper title and number of citations.
7) Which were the papers that use “holography” and “anti de sitter” (derived only from the paper abstract). Return authors and title.
8) Find the shortest path between ‘C.N. Pope’ and ‘M. Schweda’ authors (use any type of edges). Return the path and the length of the path. Comment about the type of nodes and edges of the path.
9) Run again the previous query (8) but now use only edges between authors and papers. Comment about the type of nodes and edges of the path. Compare the results with query 8.
10) Find all authors with shortest path lengths > 25 from author ‘Edward Witten’. The shortest paths will be calculated only on edges between authors and articles. Return author name, the length and the paper titles for each path.

Assignment handout

Your deliverable should be a compressed file that you will upload to moodle and include:

1. Report.pdf

a. Detailed description of your graph model using a chart and a verbal description of the elements.
b. The commands you used in order to import the files to the database.
c. The Cypher code for the required queries with their **respective results**.

2. The program/script you implemented: for any step of this assignment.

3. queries.cy: A text file with the queries you expressed in Cypher language.