Athens University of Economic and Business

MSc in Business Analytics

Data Mining Techniques – Assignment 2

Deadline: 9/7/2018

Group assignment (groups of up to 2 people).

The assignment corresponds to 20% 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: I.Filippidou (filippidoui@aueb.gr)

Description

You are given a file that contains the data from 684 chess games played in world chess tournaments. Specifically, the file includes the players' details, the tournaments, the date, the result of the game, as well as all the moves of each game and the positions that occur in the chessboard with each move. You can find the file with all games in moodle (chessData.txt) along with the description for each component of the file (dataDescription.txt).

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 this task, you will need to study the chessData.txt file details as described in dataDescription.txt. From this data, all elements of a chess game must be represented as nodes and edges, and especially all the positions and moves depicted in each instance of the file. All positions on the chessboard are unique (same positions may occur in several games), in your database you should create position nodes uniquely described by their FEN property. When you design your model, in each different node and edge type, you should include only the elements that describe it, there should be no repetitions of elements (for example the 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.

Based on the model you have designed, you should then create a graph database instance on Neo4j and load the data that you are given. In particular, you should create a small program in a language of your choice, which will accept the file with the data given and create new files (csv) suitable for loading into neo4j. Finally, after entering the data in your database, you are asked to write and execute the following queries using the Cypher language. To speed up loading and query response times, you could also create the right indexes.

Queries

1. In how many games (count) the position with FEN: r1bqkbnrpppp1ppp2n51B2p34P35N2PPPP1PPPRNBQK2R has appeared and what was the percentage that white wins.
2. For all games containing position FEN: r1bqkbnrpppp1ppp2n51B2p34P35N2PPPP1PPPRNBQK2R, how many times (count) won the white, won the black or the game was draw.
3. Which was the event that had the most games, and in how many of these games had played "Karpov Anatoly" with white or black.
4. Which player had played most games with "Ruy Lopez" opening.
5. How many games had the sequence of moves (in the exact order) "Nc6", "Bb5", "a6", and which was the players of these games.
6. Display all game details, event, players and moves of the game with GameNumber: 636.
7. Display all chess games that include the position with FEN: r1bqkbnrpppp1ppp2n51B2p34P35N2PPPP1PPPRNBQK2R and after this position the move "a6" was not played. Also display the alternative moves and the game result.

Assignment handout:

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

1. Report.pdf:
   a. The names and registration numbers of the team members.
   b. Detailed description of the data model you have created using a diagram and a verbal description.
   c. Description of the files you have created for importing into neo4j, as well as the commands you used to import the files to the database.
   d. All cypher queries and the results from neo4j.
2. The program / script, you have used in order to create the csv files for neo4j.
3. queries.cy: A text file with the queries you expressed in Cypher language.