Introduction to Graph Databases

Alejandro Vaisman

Course Project

1 Introduction

Consider the graph **courseproject2025.db**.The graph contains many Movie nodes where the imdb rating is missing (imdbRating property). Also note that normally the imdbId property should contain a "tt" prefix, which is also missing. Some other inconsistencies may exist, and they must be fixed if necessary for this project.

Remark 1. The project can be done in groups of at most two students (of course, it can also be done individually).

Remark 2. All items must be solved to get the project approved. That means, no item can be left undone.

2 Part I - Graph Preparation

You must first perform the following data preparation tasks.

1. Load the graph **courseproject2025.db** in a Neo4j database (use version 5.25).

2. Fix at least 5,000 nodes where the imdbRating property. That means, obtain the movies' imdb rating. Since the IMDB API requires an AWS account, you can use a free API like for example https://omdbapi.com/.

3. Create a user node for yourself in the graph, and rate at least 200 movies.

3 Part II

3.1 Recommendation

You will compute metrics that allow obtaining recommendations of movies that are similar to the ones you have rated.

1. Compute similar movies based on their imdbRating, year and duration. Choose the metric that works best for the database content. Preferably, apply a weighted combination of at least two metrics.

2. Also apply some similarity computation based on non-mumeric properties, like genre for example, and define a score based on this.

3. With this information, compute a similarity score and create an edge in the original graph with such score as a property. Only create the edges for the 10% most similar movies.

4. Finally, compute the movies that you have not rated, and that are most similar to the ones yo rated with a score higher than 8.

3.2 Community Detection

This is a self-research work. You will investigate two community-detection algorithms and apply these methods to the recommendation graph.

1. Classify users using the Louvain clustering method.

2. Do the same using k-means clustering.

You need to clearly explain the procedures and report the results.