



CSE 3021 – Social And Information Network

Deep Learning on Graph Database using Neo4j (Review-1)

Team Members

Reg. No.	Name
15BCE1326	Osho Agyeya
15BCE1136	Pushpit Bagga
15BCE1252	Tushar Pahuja
15BCE1257	Hargur Bedi

Faculty : Dr.L.Jani Anbarasi
VIT, Chennai.





ABSTRACT

Neo4j is a graph database management system developed by Neo4j, Inc. Described by its developers as an ACID -compliant transactional database with native graph storage and processing, Neo4j is the most popular graph database according to DB-Engines ranking, and the 22nd most popular database overall. Keras is an open source neural network library written in Python. It is capable of running on top of TensorFlow, Microsoft Cognitive Toolkit or Theano. Designed to enable fast experimentation with deep neural networks, it focuses on being user-friendly, modular, and extensible. The project is aimed to connect a Neo4j graph database to Keras. The main objective of the project is to create a neural network for a review prediction task. The model is trained on the set of review scores given by the persons for products. The database is created in Neo4j and it is injected into Keras learning model to make predictions regarding reviews of products.



MOTIVATION

The main motivation behind the project was to study the Machine Learning and Deep Learning concepts and learn how to apply them to real life applications. The aim was to study how people give ratings to different products based on various factors and how Deep Learning using Keras could be used to perform the prediction of ratings. Neo4j software would be used to represent the data in the form of a graph with people and products as nodes and study the relationship between them.



EXISTING SYSTEM

Various regression models using machine learning algorithms like Random Forest, Decision Tree and Support Vector Regression have been used for movie rating prediction tasks.

Pros

- Simple to train the model on the dataset.
- Ability to handle huge amounts of data.

Cons

Acts as a black box to the user due to the lack of proper visual understanding provided by the models.



PROPOSED SYSTEM

A Deep Learning Neural Network model is used for this project. The model has two dense layers, of width 6, with tanh activation layers. These give it room to separately combine the corresponding elements from the style preference of the persons and style of the movie vectors. Then an output layer of width one is applied, also with tanh activation. The network at this point has generated a single value, the prediction for the review score. With our model built, we compile it to use the popular Adam optimizer and mean squared error as the loss function.

Pros

- The model used along with Neo4j graph database allows the learning system to explore more of your data.
- The neural networks are well known for being good at memorizing huge amounts of data which in our case would lead to high accuracy of prediction.



SOFTWARE & HARDWARE REQUIREMENTS

- **SOFTWARE AND LANGUAGES**

- Neo4j, Python, Cypher, Keras

- **HARDWARE USED**

- ASUS GL552VW ROG, NVIDIA GeForce GTX 960M with Compute capability 6 running Windows Operating System



Neo4j

It is a graph database management system developed by Neo4j, Inc. Described by its developers as an ACID-compliant transactional database with native graph storage and processing, Neo4j is the most popular graph database according to DB-Engines ranking, and the 22nd most popular database overall.

Python

It is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales. In July 2018, Van Rossum stepped down as the leader in the language community after 30 years. Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library.



Cypher

Cypher is a declarative graph query language that allows for expressive and efficient querying and updating of a property graph. Cypher is a relatively simple but still very powerful language. Very complicated database queries can easily be expressed through Cypher. This allows users to focus on their domain instead of getting lost in database access.

Keras

It is an open source neural network library written in Python. It is capable of running on top of TensorFlow, Microsoft Cognitive Toolkit or Theano. Designed to enable fast experimentation with deep neural networks, it focuses on being user-friendly, modular, and extensible.



LITERATURE SURVEY

- The main aim of the project is to study the relationship between the people and the reviews of products written by them.
- The main challenge of the project is to predict **what review score a person will give to a product.**
- In order to study the relationship, a simple product review graph would be generated. The graph would have the following nodes with properties:
 - **Person:** with a style_preference vector of width 6 that one-hot encodes with product style they like.
 - **Product:** with a style vector of width 6 that one-hot encodes which style that product is.
 - **Review:** between **person** and **product**, which a score floating point number.



How review scores are calculated?

Review scores are calculated as the dot product between a person's style_preference and a product's style.

For example, a person with style_preference $[0, 1, 0, 0, 0, 0]$ reviews product Nintendo Switch with style $[0, 1, 0, 0, 0, 0]$, and her review score is 1.0. When she reviews a PS4, with style $[1, 0, 0, 0, 0, 0]$, her review score is 0.0.

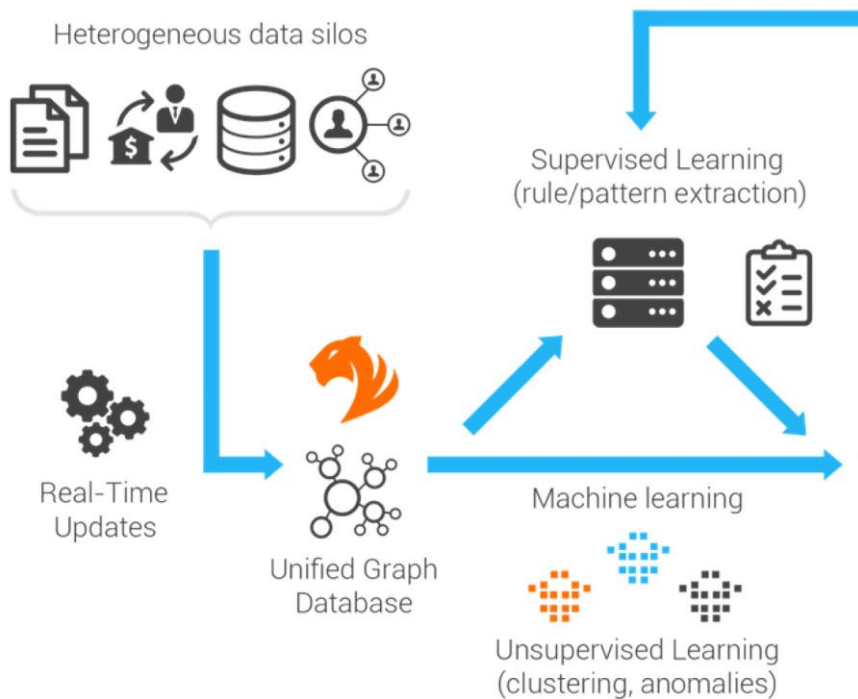


SYSTEM ARCHTECTURE

A **Deep Learning Neural Network** model is used for this project. The model has two dense layers, of width 6, with tanh activation layers. These give it room to separately combine the corresponding elements from the style preference of the persons and style of the movie vectors. Then an output layer of width one is applied, also with tanh activation. The network at this point has generated a single value, the prediction for the review score. With our model built, we compile it to use the popular Adam optimizer and mean squared error as the loss function.



SYSTEM DIAGRAM

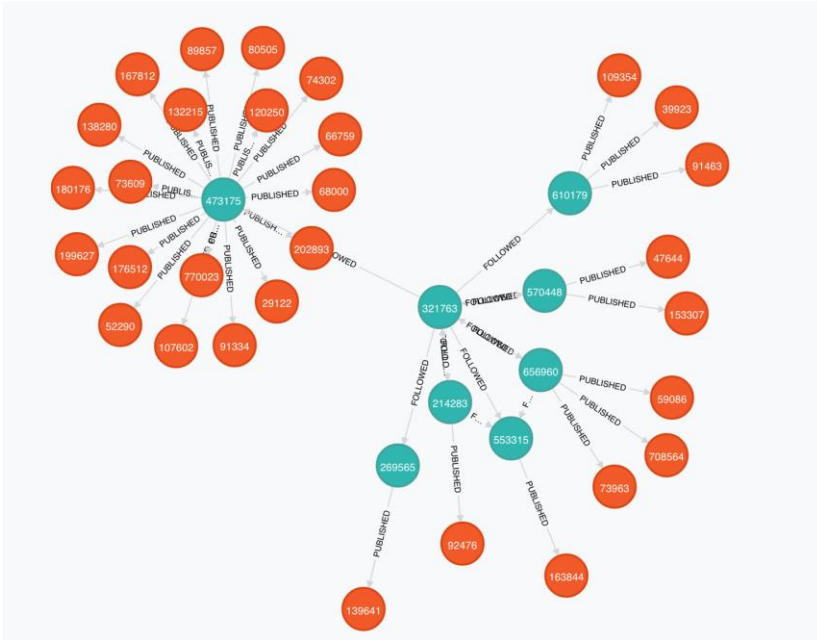




MODULES

1. Creation of graph database manually in Neo4j
2. Exploratory data analysis
3. Establishing connection between Neo4j and Keras
4. Create a machine learning model
5. Training session of the model
6. Hyperparameter tuning after observing results of training session
7. Finalizing the model
8. Making final predictions

1. Creation of graph database manually in Neo4j



- This involves converting structured database into graph database. When deriving a graph model from a relational model, we should keep the following guidelines in mind:

A row is a node

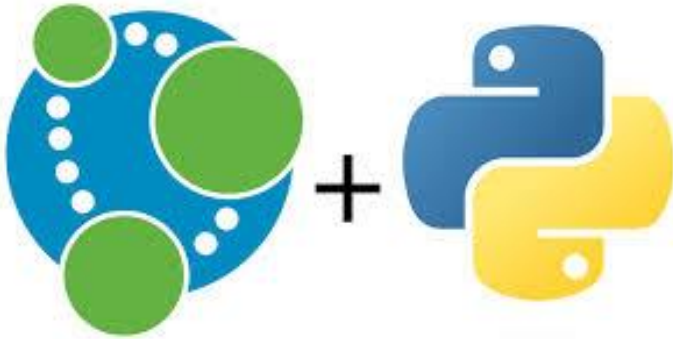
A table name is a label name

2. Exploratory data analysis

		ACQUIRED	
		KNOWN	UNKNOWN
AWARENESS	KNOWN	The things we are aware of knowing Beware complacency	The things we are aware of not knowing Deductive reasoning
	UNKNOWN	The things we are unaware of knowing Acknowledge & retrieve	The things we are unaware of not knowing Inductive reasoning

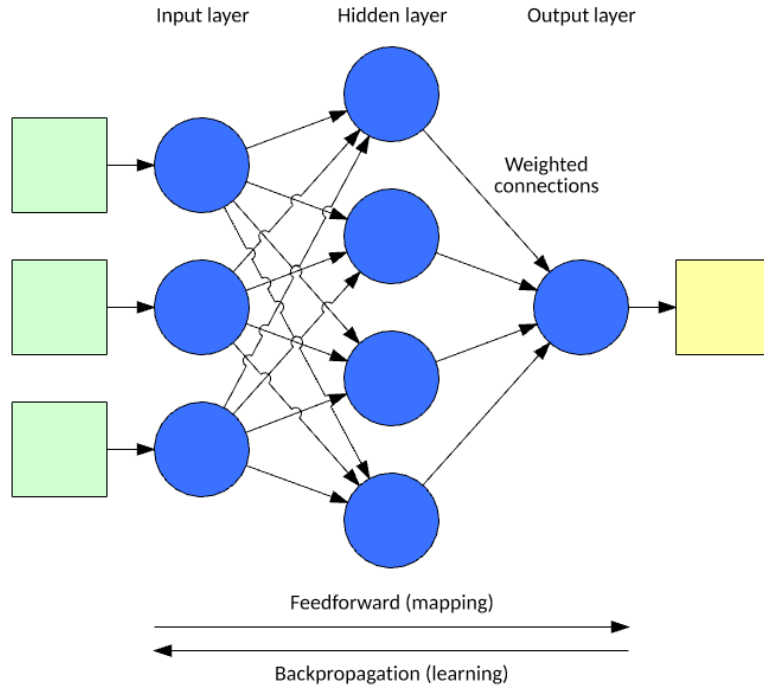
- In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. This involves studying the data to find patterns in it in order to find clusters and related data.

3. Establishing connection between Neo4j and Keras



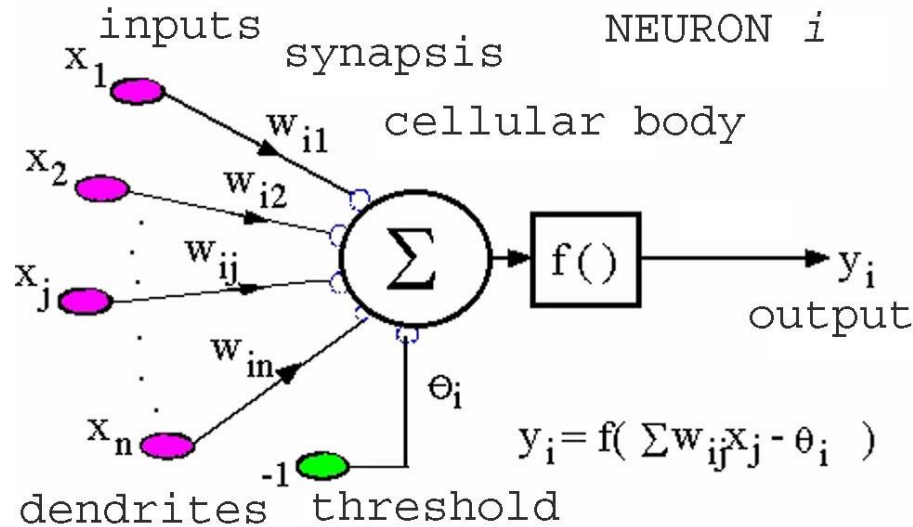
- After having studied the data, we are supposed to connect the graph database to the Keras code which we want to execute by the usage of Neo4j APIs.

4. Create a machine learning model



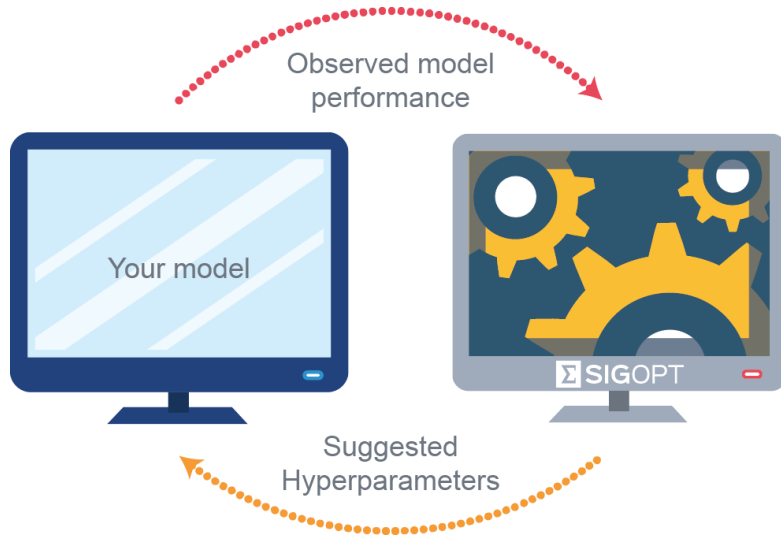
- This involves defining the layers of the neural network along with the loss function and optimizer. The no. of layers in the output layer varies according to the input statement.

5. Training session of the model



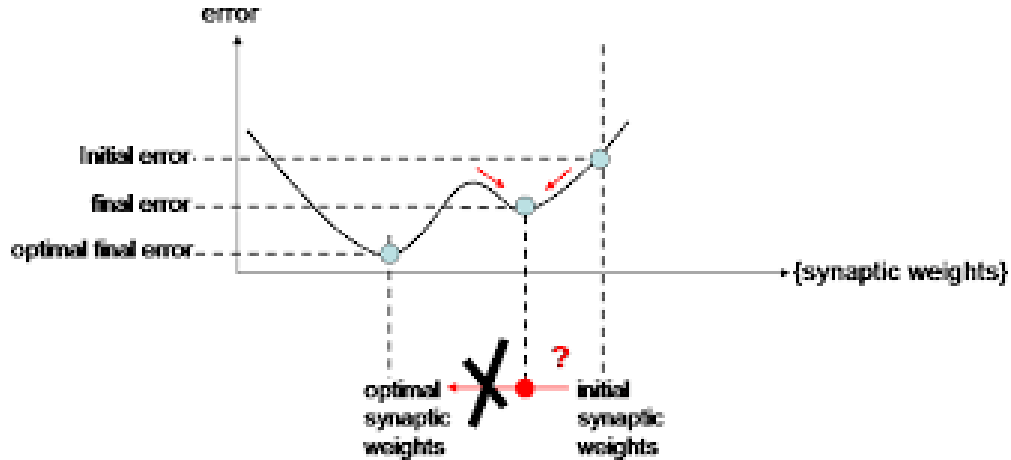
- This involves changes in the initial values of the network according to the prediction and true value. The neurons try to adjust themselves according to the data that it received and predicted output.

6. Hyperparameter tuning after observing results of training session



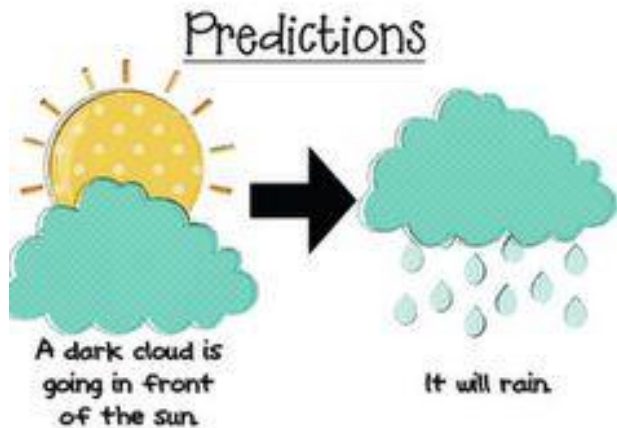
- The way in which the neural network trains itself goes on to decide whether the value of the learning rate, no of layers, no. of neurons in each layer should be changed or not to cause faster converging.

7. Finalizing the model



- After finalizing the values of the hyperparameters, the neural network is finalized and retrained from scratch.

8. Making final predictions



A **prediction** is a good guess about what will happen in a story using clues from the story.

- Final predictions are made for unseen data and these predictions are stored for later use.



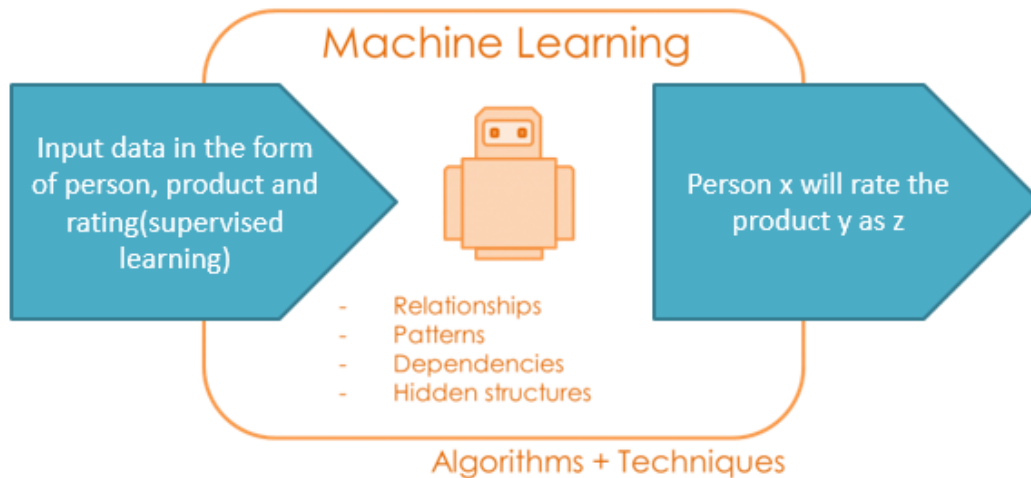
DESIGN & ANALYSIS

- **SOFTWARE AND LANGUAGES**
 - Neo4j, Python, Cypher, Keras
- **HARDWARE USED**
 - ASUS GL552VW ROG, NVIDIA GeForce GTX 960M with Compute capability 6 running Windows Operating System



EXPECTED PROJECT OUTCOME

- The model is supposed to predict the **review score a person will give to a product** using the knowledge learnt from the dataset.





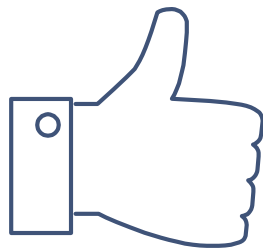
WORK PLANS FOR 2ND REVIEW

- Complete graph database creation
- Exploratory data analysis
- Data cleaning
- Deep learning model generation
- Training
- Initial results



REFERENCES

- <https://neo4j.com/docs/developer-manual/3.4/>
- <https://keras.io/>
- Python Machine Learning by Sebastian Raschka
- Neural Networks and Deep Learning by Coursera.com



THANKS!