Assignment 2: Knowledge Graph based Question Answering System

1. Data Preprocessing and Graph Construction

Data Cleaning: Addressed missing values in key columns such as Revenue and Metascore. Processed categorical columns like Actors and Genres for better representation.

Graph Construction: Constructed a knowledge graph using NetworkX. Nodes in the graph represent entities like Titles, Years, Directors, Actors, and Genres. Edges capture relationships such as 'year' and 'actor.'

2. Embeddings Generation

Node2Vec Embeddings: Employed Node2Vec to generate embeddings for nodes in the knowledge graph. Considered parameters like dimensions, walk length, and number of walks for effective node representation.

BERT Embeddings: Utilized SentenceTransformer with the 'bert-large-nli-stsb-mean-tokens' model to generate embeddings for textual data, enhancing the contextual understanding of entities.

3. Downstream Task and Fine-Tuning

Downstream Model: Designed a neural network for a downstream task with linear layers. The model takes BERT embeddings as input and predicts Node2Vec embeddings as output, bridging the textual and structural representations.

Training: Fine-tuned the downstream model using Mean Squared Error (MSE) loss and the Adam optimizer. Conducted 100 epochs to optimize the model's performance.

4. Integration for Question Answering

Question Processing: Leveraged spaCy for Named Entity Recognition (NER) to identify entities such as movie titles and directors in input questions.

Subgraph Selection: Utilized NER to extract a relevant subgraph from the knowledge graph, focusing on entities identified in the input question.

Top-k Matching Nodes: Developed a mechanism to identify the top-k nodes in the knowledge graph that best match the predicted embeddings for a given question, streamlining the search process.

5. Evaluation and Testing

Performance Metric: Evaluated the system's performance using Precision@5, Precision@15, and Precision@25 as metrics.
P@5 = 0.2375
P@15 = 0.39375
P@25 = 0.50625