K{r}eeda Labs

Assignment Report

1. Introduction- The objective of this assignment is to evaluate the candidate's ability to extract information from research paper data, create a knowledge graph using natural language processing (NLP) techniques, and build a machine learning model to extract valuable insights from the knowledge graph.
2. Data Acquisition and Preprocessing

* Data Acquisition-
* Data Preprocessing- Removed stop words, punctuation, and special characters to clean the text data. Tokenized the text into words or phrases for further analysis. Applied Lemmatization to normalize words.
* Data Exploration- We started by analyzing the overall word frequency distribution across all research papers in the dataset and also identified common keywords within the research papers to understand the primary focus areas of the dataset.
* We created a word cloud visualization to provide an intuitive representation of the most frequent words in the research papers.

1. Entity recognition and relationship extraction

* Extracted entities like authors, organization and keyword from the “url” given in dataset. With the help of Beautifulsoup and HTML parser.
* Extracted relationship between author, organization and keywords.
* Extract relationships between entities within the text. For example, identify which authors collaborated on the same papers, which organizations are affiliated with which authors, and the key terms associated with specific papers.

1. Construction of the knowledge graph

* We create a directed graph using NetworkX (**nx.DiGraph()**), where edges have a direction indicating the collaboration relationship.
* For each paper, we iterate through the list of authors and add edges between them to signify collaboration.
* The **nx.spring\_layout** function is used for layout purposes to arrange nodes in a visually appealing manner.

1. Feature engineering

* Feature engineering involves selecting and creating meaningful features that can enhance the performance of a machine learning model. In the context of extracting insights from a knowledge graph representing author collaborations.
* Degree centrality, betweenness centrality, and other graph-based metrics. Number of collaborations and publication count for each author.
* Collaboration strength (e.g., based on the number of joint publications).
* Weighted indicators of the strength of the relationship between entities.

1. Model selection and training

* As this is a classification task, so we used random forest classifier to train our model.
* Splitted the dataset into test and train.

1. Evaluation of the NLP model

* Evaluated the performance of NLP model using appropriate metrics.
* Accuracy= 0.98