Bill.com Challenge

Description: Bill.com is a platform that connects businesses through payments. These connections create a large network of dense and sparse connections. At Bill.com, we are interested in using this network to help our customers find new businesses to connect to, search our database, and provide recommendations, to name a few. Therefore, to emulate the challenges we face at Bill.com, we have provided a toy dataset that contains information at the entity level as well as the relationships that make up a network.

Data set: The dataset is a network, represented as a graph, from a social media website. Each node, represented by a number id, is a page from that website, and an edge exists between two nodes if both pages link to each other. In addition, we have provided the page's description and the page type.

The data consists of five files listed below:

1. training_graph.csv - This dataset contains the edges from the training graph. Each row contains a pair of nodes, representing the fact that an edge exists between the two nodes.
2. node_features_text.json - This dataset contains the text descriptions for each of the nodes. Each node id maps to a one-hot encoded text sequence.
3. node_classification.csv - This dataset contains the page type for each of the nodes. One column contains the node id, while the other contains the page type for that node.
4. isolated_nodes.csv - We have purposefully removed all of the edges from some of the nodes in the graph. They are listed here for your reference.
5. test_edges.csv - Your goal is to classify whether an edge exists between these pairs of nodes.
6. test_labels.csv - A binary feature indicating whether an edge between two nodes exists or not in the test dataset. This dataset will be withheld until approximately 2pm on Saturday.

Goals: Your goal is to create a model that predicts whether two pages link together utilizing the graph data, the page descriptions, and the page type. We have provided pairs of nodes in test_edges.csv and your job is to predict whether an edge exists between these pairs of nodes.

Please report your test accuracy and any other relevant metrics/visualizations for test_labels.csv. When judging, we are interested not just in the raw performance of your algorithm in terms of test accuracy, but are also looking for interesting/novel problem-solving approaches, insightful exploratory data analysis and visualizations, possibilities for future work, and the potential challenges your model may face if deployed in a real-world setting.