Group Members:

In our CS412 machine learning project group, The members are **Sai Venkata Sri Teja Nagubandi, Rohit Reddy Kesireddy, Aryan Rao Neelagiri, and Deeksha Nandal**. We are collaborating to work on this project. We are planning to work on a problem - “Finding the Code Similarity”.

DataSet:

A comprehensive benchmark dataset and open challenge focused on code intelligence. CodeXGLUE encompasses a wide array of code intelligence tasks and provides a dedicated platform for evaluating and comparing various models. The dataset comprises 14 distinct datasets, each targeting 10 different aspects of code intelligence.

For our specific project, we are making use of the POJ-104 dataset, which is a part of CodeXGLUE.The POJ-104 dataset consists of 32,000 example code segments written in C++, which are distributed across 64 different coding problems. This dataset will serve as the foundation for our experimentation in the field of code similarity.

Machine Learning Task

The machine learning task for code similarity, particularly with the POJ-104 dataset, typically falls under the supervised learning paradigm. This dataset is specifically designed for code similarity tasks and contains labeled data, making it well-suited for supervised machine learning approaches.

In the case of the POJ-104 dataset, you have pairs of code segments and associated labels that indicate the level of similarity between the code snippets. Therefore, you can use this labeled data to train a supervised machine learning model to recognize and measure code similarity. The labels in the dataset provide the necessary supervision for the model to learn and generalize from the provided examples.

Machine Learning Techniques:

We are planning to do the following techniques

- **Text-Based Technique** - Using Doc2Vec on the POJ-104 dataset, we generate code segment embeddings for semantic similarity analysis, revealing hidden code relationships.
- **Tree-Based Technique** - Graph Neural Networks (GNNs)
    - Using pre-trained models on large codebases and fine-tuning them on the POJ-104 dataset for code similarity tasks.
- **Transformers with GNN**

Convert input into graphs and then use graph embeddings as input to transformers.

Use transformers as it has the multi-head attention technique that can be used to find the similarity between the code.