**Next Word Predictor Project**

**Surender Reddy Erigela (100941347)**

**Rohith Gundlapally (100941348)**

**Sai Siddharth Mothe (100947571)**

**Project Overview**

The Next Word Predictor project aims to build a language model using a neural network to predict the next word in a sequence of text. This documentation provides an overview of the project, its components, and the steps taken to develop the Next Word Predictor.

**Code Overview**

The core functionality of the Next Word Predictor is implemented using TensorFlow and Keras. The main components of the code include:

1. **Loading Data:**
   * The text data is loaded from a file (**data2.txt**).
2. **Tokenization:**
   * Tokenization is performed using the **Tokenizer** class from Keras. It converts the text into sequences of integers.
3. **Input Sequence Generation:**
   * Input sequences are generated by considering progressively longer sub-sequences from the tokenized sentences.
4. **Data Preprocessing:**
   * Pad sequences to ensure uniform length.
   * Split sequences into input (x) and output (y) for training.
5. **Model Definition:**
   * A Sequential model is defined using Keras.
   * It includes an Embedding layer, an LSTM layer, and a Dense layer with softmax activation.
6. **Model Compilation:**
   * The model is compiled using the categorical cross-entropy loss function and the Adam optimizer.
7. **Model Training:**
   * The model is trained on the input sequences and corresponding one-hot encoded labels for a specified number of epochs.

**Model Architecture**

The neural network model consists of the following layers:

1. **Embedding Layer:**
   * Converts integer-encoded vocabulary into dense vectors.
2. **LSTM Layer:**
   * Long Short-Term Memory layer with 150 units.
3. **Dense Layer:**
   * Output layer with softmax activation to predict the next word.

**Training**

The model is trained for 10 epochs using the Adam optimizer and categorical cross-entropy loss.

**Usage**

The trained model can be used to predict the next word in a given sequence of text.

**Future Improvements**

Possible enhancements for the Next Word Predictor project include:

* Fine-tuning hyperparameters for better performance.
* Exploring different neural network architectures.
* Expanding the dataset for improved language modeling.