Next Word Prediction

You've certainly heard about the GPT language models that are used by many AI tools such as ChatGPT. What these models do, in simple terms, they try to accurately predict what the most logical next word would be given an input. In this project, you will try to implement a similar model, of course, much less advanced. Select a set of training data with a lot of real-life text (e.g. a book) and use it to train a neural network which, given input with a couple of words, returns the list of the most probable words that would make for a logical continuation.

Reading and Splitting the Text Data:

Read the text and split it into individual words.

Remove any unnecessary characters (e.g., punctuation).

Exploratory Data Analysis (EDA):

Analyze the text to find the most and least frequently occurring words.

Identify common bigrams (pairs of consecutive words).

Preparing Training Data:

Decide on a fixed input length

X

X (e.g., 5 words).

Create pairs of input sequences and the next word.

Encoding the Data:

Encode words as integers or one-hot vectors.

Split data into training, testing, and validation sets.

Building the Neural Network:

Construct a simple neural network using TensorFlow/Keras.

Experiment with different architectures and hyperparameters.

Training and Evaluating the Model:

Train the model on the training set.

Evaluate its performance on the test set.

Test the model with custom input sequences.