**Project Proposal**

**Next Word Prediction**

**1. Title of the Project:**

"Next Word Prediction using Natural Language Processing"

**2. Brief on the Project:**

Next word prediction is a significant task in natural language processing (NLP) that aims to predict the next word in a sequence of words or sentences. This project intends to develop a next word prediction model using machine learning and NLP techniques. The primary goal is to improve user experience in various applications such as text completion, auto-suggestion, and typing efficiency.

**Project Type:**

This project falls under the domain of NLP and machine learning.

**Problem Statement:**

The problem we plan to address is the inefficiency and inconvenience users face while typing or composing text. By predicting the next word accurately, we can assist users in completing their sentences faster and with fewer errors.

**Motivation:**

The motivation behind this project is to enhance user experience in applications like search engines, text messaging, and word processors. Accurate next word prediction can save time, reduce typing effort, and improve overall productivity.

**Previous Work:**

Several studies and models exist in the field of next word prediction, including Markov models, neural networks (such as LSTM and Transformers), and probabilistic language models (like N-grams). These models have shown promising results in predicting the next word based on context.

**Approach:**

Our tentative approach involves:

* Preprocessing the text data to handle tokenization, cleaning, and normalization.
* Building a language model, possibly using a recurrent neural network (RNN) or Transformer architecture.
* Training the model on a large corpus of text data.
* Evaluating the model's performance using metrics such as perplexity, accuracy, and F1 score.
* Fine-tuning the model based on feedback to improve predictions.

**3. Deliverables of the Project:**

**List of Questions:**

* Can we predict the next word accurately based on the context of the previous words?
* How does the size of the training corpus affect prediction accuracy?
* What impact does the choice of language model architecture have on performance?

**Model Details:**

The model will utilize:

* Tokenization techniques to convert words into numerical representations.
* Embedding layers to capture semantic meaning.
* Recurrent layers or attention mechanisms for sequence modeling.
* Softmax layer for predicting the probability distribution of the next word.

**Expected Observations and Outcome:**

We expect the model to demonstrate:

* Improved prediction accuracy as it learns from more data.
* The ability to suggest relevant words based on context.
* Faster text completion for users, especially in applications like messaging and email.
* We got the highest accuracy 85.63%.

**4. Resources:**

**Data Set Source:**

<https://www.kaggle.com/code/ysthehurricane/next-word-prediction-bi-lstm-tutorial-easy-way/notebook>

**Software:**

* Python will be the primary programming language for implementing the model.
* Libraries such as TensorFlow for building and training the neural network.
* NLP libraries like Tokenizer for text preprocessing and analysis.

This project aims to contribute to the field of NLP by developing an efficient and accurate next word prediction model. Through experimentation and evaluation, we plan to demonstrate the effectiveness of the model in improving user typing experience and text completion tasks.

**Individual Details:**

**Name: -** Nitesh Kumar

**E-mail:-** [niteshk834@gmail.com](mailto:niteshk834@gmail.com)

**Phone Number:** - 7014532218