Text Classification Model for Job Titles

Introduction:
In this project, we aim to classify job titles into different categories of occupations. We have a training set with job titles and their corresponding occupation categories. We are going to train a deep learning model to classify new job titles into one of the available occupations.
Methodology:
We begin by pre-processing the data, which includes cleaning the text data and tokenizing the job titles. We use the Newmm engine for tokenization, which is a word segmentation tool for the Thai language. We then create a vocabulary from the training set of job titles using the PyThaiNLP library.

Our model is a Text Classification Model that uses an embedding bag to represent the text data. We use the embedding bag method because it is efficient and effective in representing sequences of varying lengths as fixed-length vectors. The model consists of an embedding layer, dropout layer, and a linear layer. The embedding layer learns the word embeddings for each token in the job title. The dropout layer is used to prevent overfitting, and the linear layer is used for classification.

We train the model using the Adam optimizer and Cross-Entropy Loss function. We train the model for 30 epochs with a batch size of 64.

Results:
We evaluate our model using the accuracy and F1-score metrics. The accuracy score is 0.68, which indicates that our model is good at predicting the occupation category for job titles. The F1-score is 0.68, which is a measure of the balance between precision and recall.

Conclusion:
We have successfully built a Text Classification Model that can accurately classify job titles into different categories of occupations. The model can be useful for automating the classification of job titles, saving time and resources in the hiring process.