Project Proposal for PLP Certificate	
Project Title	A Chinese Traditional Poetry Analyzing System Based on Deep Learning
Project Members	YAN JIAHUAN, LIANG ZIJIAN, LI QINYUAN
Overview	Our project aims to develop a poetry analysis application that addresses the problem of incompatibility between the language systems of classical Chinese poetry and contemporary colloquial Chinese. There exist significant differences in grammar, vocabulary, and expression between ancient and modern Chinese, resulting in numerous challenges for readers when reading classical poetry. Therefore, this project is dedicated to corpus analysis of traditional poetry and will present the results from three perspectives: translation into colloquial language, imagery painting, and allusion analysis. This project will be applied in cultural tourism and educational training fields, such as museum guides and cultural explanations, where multiple perspectives can be utilized to showcase the meanings and backgrounds of poems, enhance the presentation effect, reduce comprehension difficulties, lower the barriers for the dissemination of traditional culture, increase people's understanding and cognition of Chinese culture, and enhance the entertainment and tourism experience. This project can also be widely implemented in primary and secondary school Chinese language classrooms, using more vivid methods to help students better understand and learn these classic cultures, enhance their humanistic literacy and language proficiency, and lay a solid foundation for their future development.
Design	1. Word representation module: Chinese poetry, particularly from ancient times, is known for its refined language structure. Certain phrasing methods like parallelism, metaphor, personification, and antithesis can reflect emotions and events of that era, while simply using the word embedding method from GLOVE or Bert will destroy the semantic structure and lead to information loss and distortion. *Methodology: Retrain Chinese-Bert-wwm 2. Vernacular translation module: The original context of Chinese poetry is hard to understand, especially for non-native Chinese poetry lovers, thus we will use seq2seq method to translate the structured and refined poetry to plain and understandable Chinese text. *Methodology: Retrain seq2seq based LSTM model 3. Allusion explanation module: It is common to see allusions in Chinese poetry, which stands for the background/event that happens at the current or has happened before. For example, ‘凭谁问，廉颇老矣，尚能饭否’ written by the famous ancient writer Xin Qiji. This background story is actually from ‘Historical Records: Biography of Lian Po and Lin Xiangru’, which is another famous literary work in China. We will use allusion recognition model to capture the allusions based on the textual-level information, while we hope to further explain what is the allusion about, thus we will use powerful ChaGPT API to help us demonstrate what is the background story of the allusions. *Methodology: Allusion recognition based on Bert/POS Tagger + ChatGPT API 4. Painting association module: This module is determined to visualize the imagery, emotion, and event in the poetry. We will extract the key components from poetry by using Name Entity Recognition (NER) and POS tagging, the emotion will be extracted by another sentiment/emotion classification model. We will use the captured contents from

	aforementioned simple functions as key words, and simultaneously feed into a image generation model (API) to get the corresponding picture. The image generation object is not the main focus in this project; hence we will only focus on how to extract the most reliable key words from original poetry/translated poetry. *Methodology: NER + ENIRE 1.0
Scope of Work	A. Database 1. The Chinese-poetry database[1] is a comprehensive database of Chinese poetry, comprising a total of 55,000 Tang poems, 260,000 Song poems, and 21,000 Song lyrics and other forms of poetry. The database includes works by nearly 29,000 ancient poets from the Tang and Song dynasties. The data includes information such as titles, authors, and the contents of the poems, and is stored in JSON file format. 2. The Classical-Modern database[2] is a parallel corpus consisting of literary and modern languages. It provides comprehensive coverage of traditional ancient writings and was derived from online sources. The corpus comprises a total of 960,000 sentence pairs, which were generated through scripted sentence division and manual proofreading. B. Techniques and Measurement We divide the project into four parts. 1. For the first part, we implement the word embedding of traditional poems. We use the first dataset to implement this part. We will retrain Chinese-Bert based on whole word mask (wwm) methodologies to get the ideal relatively coarse-grained word tokenization and embedding. This method can solve the problem of slicing the meaningful word like '明月' into words with information loss like '明' and '月'. We currently do not possess the ideal output test dataset; thus means we have no way to measure the performance mathematically. However, as tokenization can be human-eye validated, we will manually choose some Chinese poetries to judge whether the outcome is acceptable or not. 2. For the second part, we implement the translation from traditional poetry to plain text. We use the second dataset to generate word vectors of traditional poems. We use the word separation results from the first part to generate word vectors of traditional poems by using the Word2Vec model, and then we use the Seq2Seq model to implement the translation. We will use BELU to measure the translation results. 3. For the third part, we implement the allusion parsing of traditional poems. We will first try to do allusion recognition, however, this task is a data-oriented one, so if we failed to find available data, we will try entity recognition instead. We will call ChatGPT to view the allusion parsing, analysis the role ChatGPT plays in it, and discuss the pros and cons of using ChatGPT. 4. For the fourth part, we will present the traditional poem as an image. We use the results of the translation in the second part to perform entity recognition and sentiment analysis on the plain text. We use the obtained words and sentiment analysis results to call the drawing API to visualize the text and present the traditional poem as an image.
Effort Estimates	1. Word representation module: a. Rough Task1 (Data Preparation/Cleaning): It is hard to find an ideal Chinese poetry public dataset, we will need to recollect our own dataset based on other people's effort, and do traditional Chinese replacement (e.g., '漢' to '汉', '語' to '语').

	b. Rough Task2 (Data Preparation/Cleaning): Some data is crawled from Wikipedia, and the data quality is not guaranteed, we will need to do supplementation to 'untitled', 'unnamed' and 'counterfeit' data. c. Rough Task3 (word embedding): As Chinese poetry is a relatively structured and refined text, it is not appropriate to tokenize it based on word level, thus will bring the significant information loss issue. For example: '九州生气恃风雷，万马齐喑究可哀', the general representation method will tokenize word by word to ('九', '州', '生', '气', '恃', '风', '雷'…), and do vectorization based on it, while the ideal output should be ('九州','生气', '恃', '风雷'). To alleviate the world-level ambiguity problem, we decide to re-train the Bert model based on SiKuQuanShu and utilize the whole word masking algorithm. 2. Vernacular translation module: Rough Task (Data Preparation/Cleaning): Till now, the only Chinese poetry translation data we can find is actually based on ancient text, which is not Chinese poetry strictly speaking. 3. Allusion explanation module: Rough Task (Data Preparation/Cleaning): This idea actually from one of the paper published in 2019, but they actually did not release the dataset. We will do NER recognition to get the events if the original method is not doable because of lack of data. 4. Painting association module: Rough Task (image generation): The test image generation model we tried previously is based on Baidu Ernie, which actually performs not so ideal if the input words is not semantically relevant (e.g., '秋天，直升机，狗'). And this actually requires us to ensure the keywords extracted from text is relevant, so that the image generation can be smoother and more acceptable. 5. Integration: Our modules is rich and comprehensive in function, hence, it will not be an easy task to concatenate them together.

[1] https://github.com/chinese-poetry/chinese-poetry

[2] https://github.com/NiuTrans/Classical-Modern