# MAIS 202 Deliverable #1

## Choice of Dataset

The dataset I will be using is the Shakespearify dataset available on Kaggle at https://www.kaggle.com/datasets/garnavaurha/shakespearify. The reason I will be using this dataset is because it contains the complete plays of William Shakespeare as both the original text and as a modern translated text created by Shakecleare. This dataset is complete and contains entries for a phrase-by-phrase translation, which will make data preprocessing significantly easier.

## Methodology

In this project, I will be creating a Shakespeare-English machine translator. Users will be able to input full English sentences into a text prompt and receive as output a Shakespearian translation of the sentence. I will be approaching this by using an encoder-decoder model with a recurrent neural network which will be trained on a modern English translation of Shakespeare's works.

### Data Preprocessing

As the data on this dataset is already fairly clean, to preprocess data I will first clean up punctuation by adding spaces so they function as separate words, check for text artefacts, create words representing the start of the sentence and the end of the sentence (SOS and EOS), tokenize the words in each sentence (assign a numerical value to each unique word in the dataset for proper processing), and split the dataset into fixed-length blocks by padding shorter blocks and possibly excluding longer blocks.

### Machine Learning Model

I will be using a simple many-to-many encoder-decoder model using RNN cells to build this translator, as explained on https://towardsdatascience.com/language-translation-with-rnns-d84d43b40571. This will take an input sentence, encode the meaning into an internal state, which the decoder will then use to output a translated state. There are more advanced models that use several other additional types of layers to address problems like dealing with longer sentences and rarer vocabulary, but for this project, I will stick with a simpler model, as it would be more difficult to implement these other techniques, which might not generalize well with a comparatively small dataset. I want to predict translations of novel sentences using this model, where any modern English sentence can be inputted, and a Shakespearean sentence can be outputted.

### Evaluation Metric:

In this project, a translation is judged based on whether the predicted word was accurate to the target word. Therefore, this project will be evaluated on a loss function.

## Application

In my web application, I will have an input text box where the user can input any modern English sentence. Upon clicking a translate button, the user will see a Shakespearean translation appear below. I plan to implement this using Flask, which I have used a couple of times before.