DOCUMENTATION

DISCLAIMER: Until i received this project, I had no idea about NLP. I had heard of it as an algorithm but never worked with it. I read about NLP, tried my best to implement it and came up with the following:

After importing the required libraries, I read the train and test data.

Pre Processing:

Cleaning the data:

- Converting everything to lower case
- Tokenization
- Removing Punctuations
- Removing Stopwords
- Lemmatization

I pre processed train and test data.

MODEL 1:

Extracting Features:

Using the Count Vectorizer, I fit tranformed my training data.
Then i transformed the testing data.
I extracted 6210 features.

After converting it to a dataframe, i splitted the data (training data) into test and train and checked for the mean squared error value.
It came out to be: **1.4294775816346494**

MODEL 2:

I loaded the data from all_prompts.csv file and pre processed it.

Adding Features:

Since the previous model has no relation to the prompts provided, I made another dataframe with the following features:

Score: I calculated how many words from the promp matched the essay
Words: Total number of words in the original essay - let this be w
Terms: Total number of uniquewords in the original essay - let this be t
Ratio: Ratio of w/t
Textual_div: Reflects the textual creativity and kind of vocabulary used

I made this dataframe for the preprocessed test and train dataset. I splitted the data (training data) into test and train and checked for the mean squared error value.
It came out to be: **0.8959161882549452**

Since in the second model MSE value was lower, I went ahead and predicted the values using model 2.

Ideally, both these models should have been used. However, i wasn’t sure on how to achieve that.
Thank you, I learnt a lot from this assignment!