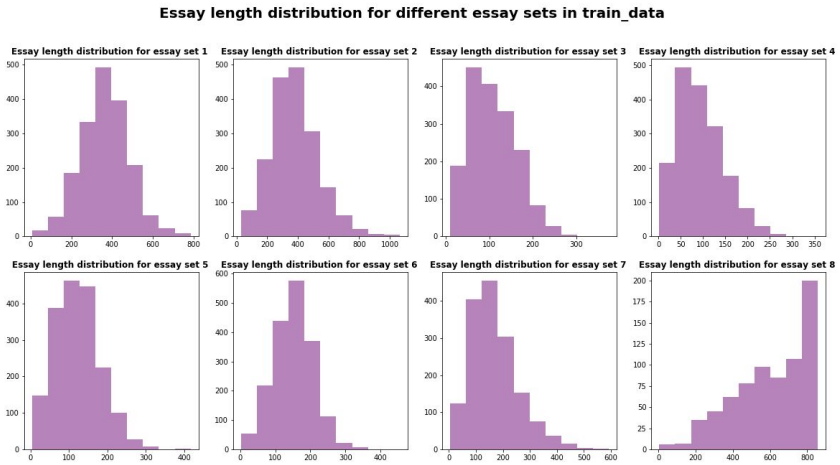
Milestone 3 - Team Anonymous

Wenjie(Bianka) Gu, Shaoling Han, Jason Zhang, Erin Yang

EDA Summary

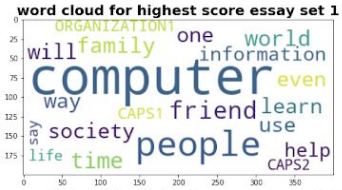
Our dataset has a combination of essays from 8 different essay prompts. After some data explorations, we find that each essay prompt has about 1700 graded essays but they have different score ranges and distributions. Here is the visualization:

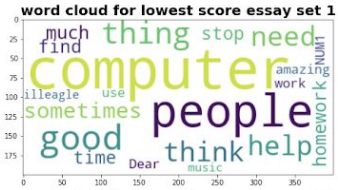
Although the data volume of essays of each essay prompt should be sufficient to do a transfer learning, we are still interested in whether we can combine all data together to train a single model so that we can have a more generalized essay grader without specifying the essay prompt of the essay to grade. To check whether such approach may work or not, we decided to check the essay length of distribution of each essay prompt and we also created word clouds for the highest and lowest scored essay of each prompt to see if they have enough similarity. Here are the visualizations:

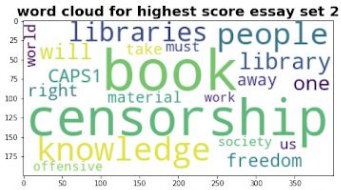


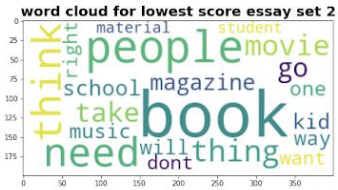
Within each essay set, we create word clouds for essays that receive either high or low scores. When examining within the essay set, it seems like there is no clear difference in the top words generated from both high and low score essays. It makes sense to us since the word cloud is created based on word frequency, and the most frequent words tend to be topic words mentioned in the prompt. For instance, for essay set 1 that asks people to '*state your opinion on the effects computers have on people*', words such as computer and people are highlighted in the word cloud.

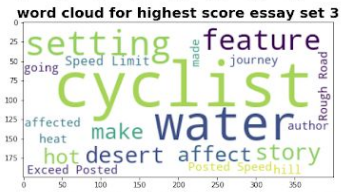
When examining essay similarity across essay sets, we found that since the essay sets ask about different topics in life, the top keywords for different essay sets are drastically different. Thus, we think it may not be a good idea to rescale the scores of all the essay sets and generate a single model for all the sets since the essays answering to different prompts could be inherently incomparable. Also, practically speaking, a real-life essay scoring system should not adopt a one-for-all evaluation criterion for essays that are of different topics.

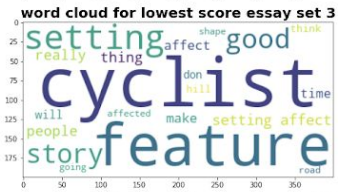


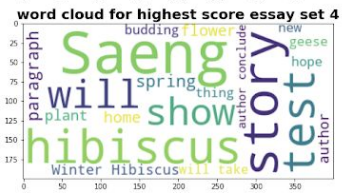


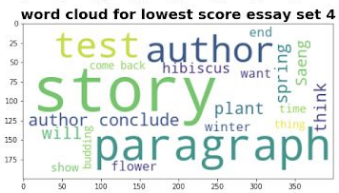


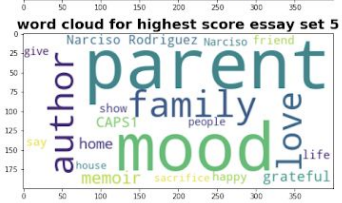


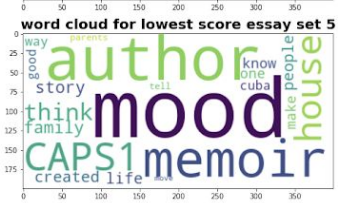


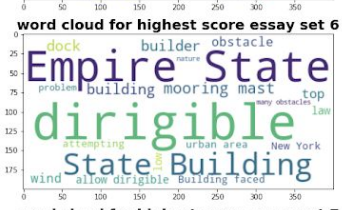


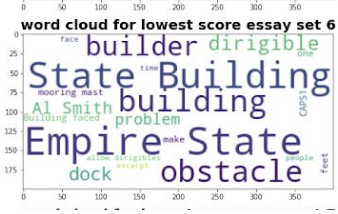












Data Pipeline

For this project, we're focusing on column 'essay' as input and 'domain1_score' as output for this dataset. Due to different grading scales for different essay sets, we decided to train the baseline models separately for each essay set. We first split the data into 8 groups based on their essay set belongingness. We then standardized and vectorized the text input and prepared a TF dataset for each group.

Baseline Model

We take a simple FFNN (feed forward neural network) model and a language model with embedding and 1d convolutional layer as our baseline models. The input is text of

essays and the output is categorical classification of score. We show the validation accuracy in the following table.

	essayset_1	essayset_2	essayset_3	essayset_4	essayset_5	essayset_6	essayset_7	essayset_8
FFNN	0.3081	0.4583	0.3497	0.4535	0.5069	0.4278	0.1401	0.1862
Embedding + CNN 1D	0.5602	0.6778	0.6474	0.6254	0.6565	0.5667	0.1656	0.2483

Among all datasets, we can see the models with embedding outperform the simple FFNN models, as expected since FFNN models do not capture actual meaning of words. We can also see the accuracy of the embedding model is higher in essay set 2, 3, 4, 5, where the number of unique score levels is relatively small, and accuracy is low in essay set 7 and 8, where there are dozens of unique score levels. This implies that it doesn't make much sense to build a multi-classification model, especially when there are lots of score levels. We need to find a better way to depict the ordinal relation among different score levels.

Plan

We plan to experiment with pre-trained word embedding such as Word2Vec and pretrained language models such as BERT to improve the accuracy of our essay rating system. Furthermore, we want to build a single model for all the essay sets and develop a web-app where we can input a sample essay and the web-app returns the extracted features from the essay as well as its rating.