Members: Kevin Loughlin, Annie Hwang, Paul Lisker, and Anmol Gupta

Milestone 5

In our Milestone 4, we considered word counts for each ngram in our logistic regression to calculate the expected scores of the essays. Word count is the most basic characteristic of an essay that many people positively correlate to higher essay scores, which was seen for most of the essay sets in our data exploration phase. However, as mentioned before in milestone #3, we want to consider more features that can increase the prospects of our modeling scores. In order to this, we are going to add the number of sentences, average sentence lengths, number of paragraphs, average paragraph lengths, number of grammatical errors, and number of punctuations used as extra predictors for our logistic regression model. But in addition to including these factors, we will experiment with other models we have learned, such as LDA, QDA, Random Forest, and Decision Trees. Incorporating cross-validation, as well, with these models, we should be able to use Spearman's rank correlation coefficient to score these various methods, and ultimately, find the most accurate model.