

Automated Essay Grading Milestone #4

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

In the creation of our baseline model, we proceeded by creating the training and validation sets, and cleaning the data—particularly, we cleaned the essay by converting all of the words to lowercase and discarding punctuation so that we were left with just the words of the essay. In doing so, we were able to create a baseline model that correlated ngram frequency to essay score.

To do this, we used the TfidfVectorizer, with an ngram range of (1,1)—for the range of (2,2), our computer's computation power was not enough, so we will seek to ameliorate this in the next steps. Furthermore, we used a Logistic Regression with L1 and L2 losses. Finally, we used the Spearman coefficient to analyze the model (standardizing and de-standardizing the essay scores along the way).

Our analysis concluded that, for ngram range of (1,1), the Logistic Model with L1 was more successful than L2, but Spearman favored the L2 than the L1. However, both were highly successful according to Spearman, with ~ 0.91 correlations