CSC 482 Project 1: Grading Fourth Grade Essays with NLP

Lauren Nakamichi, Kathirvel Gounder

Lead:

To begin scoring the leads of the essays, we first split the leads from the rest of the essays. To do this, we took the first paragraphs out of each of the essays. Some of the essays contained titles or tables of contents, so we filtered those out by choosing the first paragraph with at least 4 sentences. If no such paragraph exists, then we reduced the threshold to 3 sentences, and then 2, etc. until a match was found.

To predict the score for the lead paragraphs, we used five different features: number of sentences, the average number of words per sentence, vocabulary size, range of sentence lengths, and longest sentence length. These features helped us to evaluate the quality of the writing in the lead paragraph statistically.

To measure the average number of words per sentence, we calculated the mean sentence length within the lead paragraph. We defined the sentence length to be the number of tokens returned by the NLTK sentence tokenizer.

To approximate the vocabulary size, we measured the size of a set of all of the tokens after they have been converted to lowercase and filtered only to tokens comprised of letters of the alphabet. One flaw of this method is that contractions (containing apostrophes) and other words containing symbols would have been wrongly represented as two tokens, but since we are measuring the size of a set, additional occurrences of the same word would not increase the impact.

The range of sentence length was calculated by finding the difference between the length of the longest sentence and the length of the shortest sentence. We once again measured the length of the sentence by the number of tokens in the sentence.

The hardest part for not just the lead but for all categories was using the right type of classifier. Initially we tried to use naive bayes and decision trees which both ended up being somewhat poor choices. The accuracies were wildly fluctuating and we instead decided to be more conservative and use Linear Discriminant Analysis or if the training set cooperated quadratic discriminant analysis. In less fancy jargon, we just fit gaussian (normal) distributions to each class (the random vector consisting of our above features) using the MLE estimate for the mean (and covariance if we did QDA, LDA assumes the same covariance for all classes which is why we get linear decision boundaries).

Ending:

To predict the scores for the ending, we first removed the endings from the essay by extracting the last paragraph with at least the threshold number of sentences, like when we were extracting the lead paragraph.

The features we chose to look at were number of sentences, average number of words per sentence, vocabulary size, average word length, sentence length range, longest sentence length, and the syllable to word ratio.

Number of sentences, average number of words per sentence vocabulary size, and sentence length range were calculated the same way as they were for the lead, but in the ending paragraph.

Average word length was calculated by filtering out the tokens that were comprised of characters from the alphabet and finding the mean length in number of characters.

When choosing a classifier to interpret and analyze the values for these features, we followed the same process as the one used to predict the lead score, ultimately landing on Linear Discriminant Analysis.

Choice (Spelling):

The third section from the rubric we chose to grade for this assignment was the “spelling” section. For this section, we chose to examine the entire essay, rather than a portion since the grading criteria was not restricted to a specific portion of the essay like the previous two.

Like for the lead and ending, we chose to use linear discriminant analysis.

We chose three features for the classifier: vocabulary size, average word length, and the percentage of the characters from the text that are vowels. We chose these features because vocabulary size and average word length could have pointed to the level of difficulty in the spelling taken on by the writer. Average word length and vowel percentage could have pointed to the accuracy of the spelling of the words in the essay.

For the vocabulary size measurement, we used the same method as we did for the lead and ending grades, but across the entire essay.

To measure the average word length, we divided the number of characters in the essay by the number of tokens returned by the tokenizer. Once again, contractions and other words containing symbols were separated and would have been returned as multiple tokens, but mitigating the impact of this was left out of scope.

To measure the percentage of characters that were vowels, we flattened the lists returned by the tokenizer to a list of characters and divided the count of letters that were vowels (‘a’, ‘e’, ‘i’, ‘o’, ‘u’) by the count of total characters. Again, symbols would have skewed this count.

Another approach we tested involved checking a dictionary for tokens from the essays. For the dictionary, we used the Unix dictionary at /usr/share/dict/words. With this approach, we found the following results:

Score	Mean Percentage of Words in Unix Dictionary
0.0	0.7662190865651625
1.0	0.7685299087365429
2.0	0.786654544411373
2.5	0.7371273712737128
3.0	0.7642327944171283
4.0	0.7416147075101567

Since there seemed to be no pattern in the mean percentage of words from the essay in the Unix dictionary, we did not use it as a feature in the grader. We believe this occurred because the Unix dictionary was unfit for this situation, so we chose to omit this feature.

Evaluation:
We did an approximate 80% to 20% training to test split, trained our classifier and reported accuracy. Our baseline was a bit different from what Professor Khosmood recommended as we believed that we wanted our classifier to generalize to an unseen test set of essays, we just never wanted to dip below 14.28 percent as that would be worse than random guessing in the long run. So what we did was run a thousand trials of our classifier, where each time we would train on a random different 80% of the data and evaluate on the corresponding random test set. This way we got a good idea of the variability of our classifier (overfitting) and bias from the accuracy.

For our Lead:

We got a classifier accuracy of 0.421925 and standard deviation of 0.071

For Ending:

We got a classifier accuracy of 0.345 and a standard deviation of 0.063

For Spelling:

We got a classifier accuracy of 0.7513 % and a standard deviation of 0.062

I actually think Discriminant analysis would be a good choice for this problem in general (with more data and better features) as a majority of students are around the satisfactory mark. QDA and LDA are bayesian approaches so they make good use of prior information and original class frequencies hence our good accuracies despite small number of features. If we had more time, we would get a few more discriminatory features and try and collect more data, that way we can fit the distributions better.