Mark Kim

11/4/2022

Homework Assignment #10 Writeup

A cursory examination of the models resulting from using 2-fold, 10-fold, and 20-fold cross-validation does not seem to show much (if any) real difference between them. Looking at the classification report for each shows that F-1 scores are identical between each model. With closer inspection, we indeed find that 2-fold and 10-fold perform identically. We only find a change when we look at the 20-fold cross-validation. Looking at the 20-fold cross-validation classification report, one can see that the classification is more consistent across the board. Instead of seeing higher and lower precision and recall depending on the author, with 20-fold cross-validation, the model has consistent precision and recall with all authors, which is the desired result. Overall, however, we cannot be sure if this improvement is statistically significant without rigorous analysis.

Looking at the feature weights for determining each author prediction, we can observe that the weights and thus the importance of each respective feature varies for each author. The graph on the following page (figure 1) illustrates that the top and bottom features by importance is different. It is not surprising that for all three authors, the least and most important features were unigrams. The least important features (all unigrams) were:

* Edgar Allen Poe: “pilgrimage”, “twin”, “nondescript”
* H.P. Lovecraft: “distaste”, “loses”, “uplifted”
* Mary Shelley: “mimic”, “posts”, “declares”

The most important features (also all unigrams) were:

* Edgar Allen Poe: “raymond”, “perdita”, “madame”
* H.P. Lovecraft: “heart”, “love”, “gilman”
* Mary Shelley: “raymond”, “perdita”, “adrian”

Even more surprising are the repetitions of “raymond” and “perdita” as high importance for predicting sentences for both Edgar Allen Poe and Mary Shelley. My intuition says that the words that are found exclusively in texts written by a particular author would hold the most

Chart, bar chart

Description automatically generated

Figure : Top and Bottom 10 Features

weight, so it seems a little counter-intuitive that two of the most important three features were shared by two authors. Diametrically, my conjecture for the least important features is that they have low weights because those unigrams are found often in all the texts. But again, there is the counter-intuitive fact that the three authors do not share the same list of low-weight unigrams. Nevertheless, I suspect that it is the *combination* of factors that really form the basis of predictions. Hence, we actually cannot consider just the most or least important features. Indeed, I believe the word “important” may not even be meaningful when describing the weights used for each feature. It seems that the phrase “can’t see the forest for the trees” describes what we are seeing here.

When scrutinizing a random sample of incorrect predictions (figure 2), we cannot be entirely positive the reason for the errors, but we can certainly speculate. Table

Description automatically generatedOne pattern we can see is that the average word length is similar for all of the cases. Another item of note is the fact that almost all incorrect predictions were for Mary Shelley and Edgar Allen Poe. Related to this is that every sentence attributed to Mary Shelley was incorrectly predicted as Edgar Allen Poe. These two items drive me to venture the guess that they both have similar writing styles. Finally, the model puts a relatively high weight for the number of words in a sentence for both Edgar Allen Poe and H.P. Lovecraft. Notice that the four sentences with the most words in the sample were incorrectly attributed between EAP and HPL.

Figure : Wrong Predictions