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

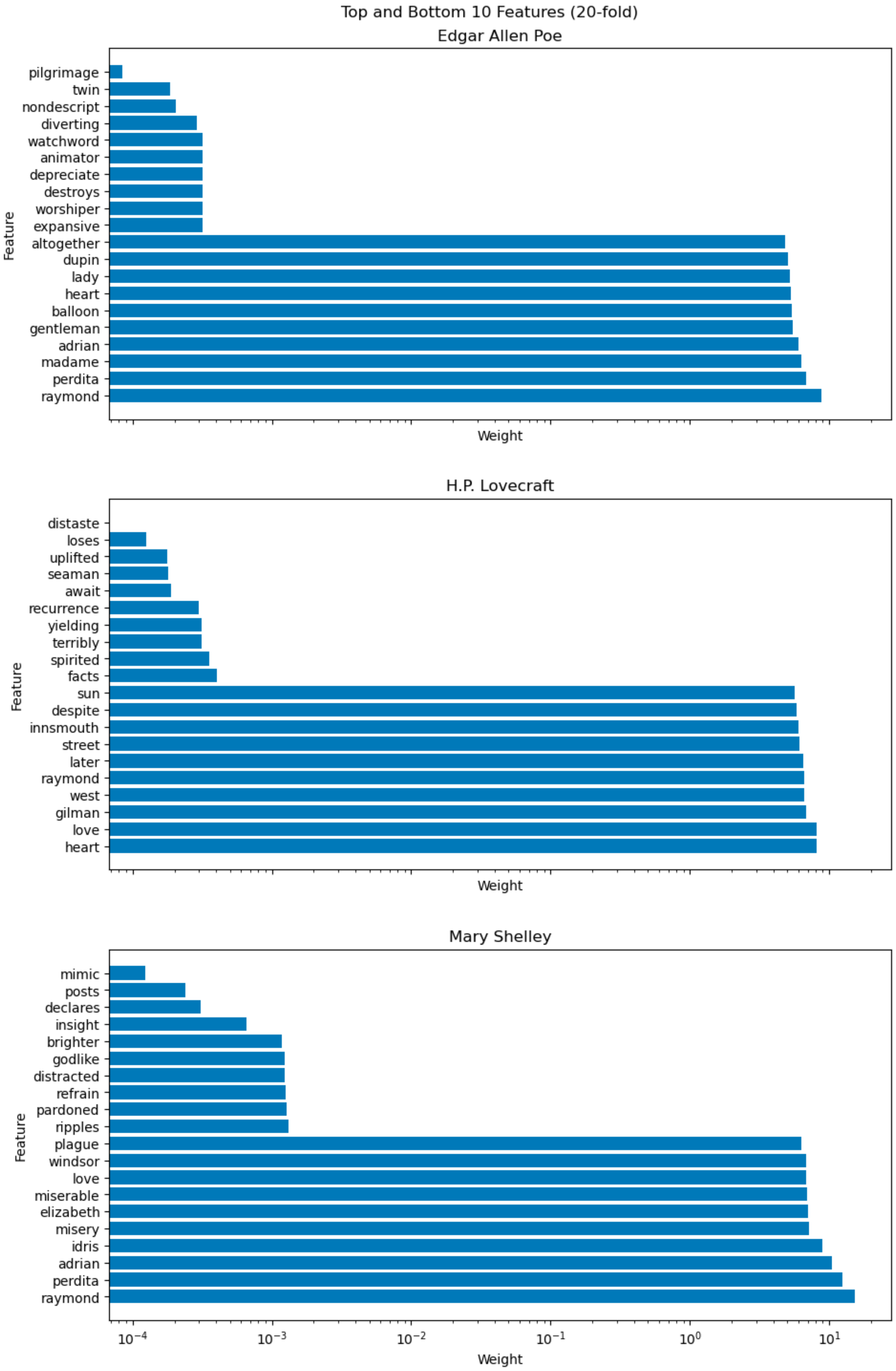
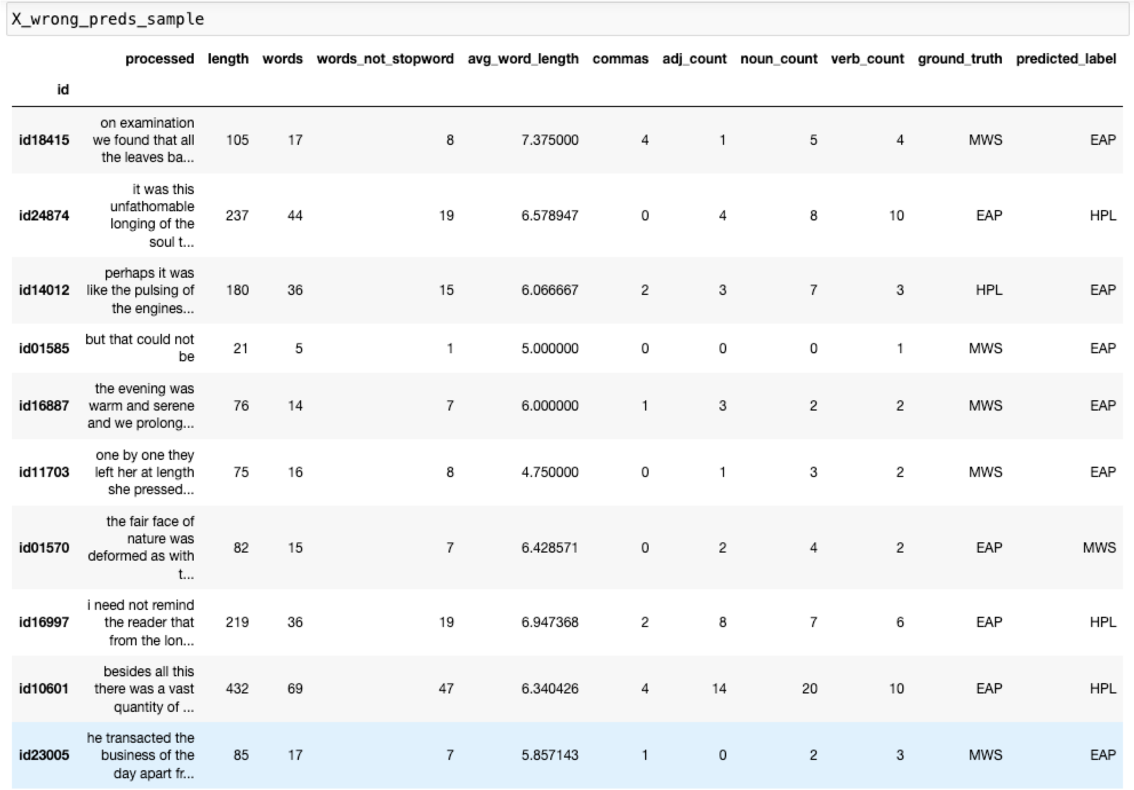


Figure 1: Top and Bottom 10 Features

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. One 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



```
X_wrong_preds_sample
```

	processed	length	words	words_not_stopword	avg_word_length	commas	adj_count	noun_count	verb_count	ground_truth	predicted_label
id											
id18415	on examination we found that all the leaves ba...	105	17	8	7.375000	4	1	5	4	MWS	EAP
id24874	it was this unfathomable longing of the soul t...	237	44	19	6.578947	0	4	8	10	EAP	HPL
id14012	perhaps it was like the pulsing of the engines...	180	36	15	6.066667	2	3	7	3	HPL	EAP
id01585	but that could not be	21	5	1	5.000000	0	0	0	1	MWS	EAP
id16887	the evening was warm and serene and we prolong...	76	14	7	6.000000	1	3	2	2	MWS	EAP
id11703	one by one they left her at length she pressed...	75	16	8	4.750000	0	1	3	2	MWS	EAP
id01570	the fair face of nature was deformed as with t...	82	15	7	6.428571	0	2	4	2	EAP	MWS
id16997	i need not remind the reader that from the lon...	219	36	19	6.947368	2	8	7	6	EAP	HPL
id10601	besides all this there was a vast quantity of ...	432	69	47	6.340426	4	14	20	10	EAP	HPL
id23005	he transacted the business of the day apart fr...	85	17	7	5.857143	1	0	2	3	MWS	EAP

Figure 2: Wrong Predictions

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.