Mohammad Hashemi

CU_seha6047

The process of creating additional features is as follows:

1. I added the confusion matrix to the code.

2. I printed the words which are labeled wrong.

3. In the beginning I just ran the default code to find out how it works.

4. I saw the train set and I understand that last sentence can help to detect the label of the text. lots of them had a pattern (for 10 points …) and there are lots of key word related to that class in last sentence, so I decided to only count the words in last sentence, but it doesn't make the accuracy much more better.

5. Then I decided to remove some words such as the, a, an, am, is, he, she, etc. which doesn't help to find the label of a text. I tried to store only the adjectives, adverbs, nouns and verbs. For this purpose, I used word tokenizer in nltk library. This step took a lots of time to run. So I decided to make it parallel and run this code on multiple machine. Then I stored these words in multiple files as features(i).csv

6. By counting nouns, verbs, adjectives and verbs the accuracy became better.

7. For next feature, I tried to use PorterStemmer from nltk library. This stemmer lets me to count such words like "goes" and "go" as one word. Again, adding this feature made accuracy better.

8. The last feature that I added is counting two words together in the original sentence. So for example in "I am here" it will count "I_am" and "am_here". I added this feature because it makes a better relation between sentence parts. Adding this feature again increased the accuracy.

My last submission on Kaggle gave me 0.80777 accuracy.

The train accuracy is: 0.999808