Authorship Attribution

Assignment 4

Text and Multimedia Mining

LET-REMA-LCEX06-2020

This project has been done by Python language program. The classification method that has been applied is support vector machine. The most important applied libraires are numpy, matplotlib, sickit-learn, and nltk. Following this I have tried to explain how I elicited the features.

Step1. Loading data

In this part I defined a function which is named *load_data* to load text training and test files. I used the load_file of sklearn library and the output are two numpy array include authors' files.

Step2. Extracting features

In this part I tried to extract from both each document and all documents. in this respect, information extracting from a specific file is not related to other files. I called this function *extract_features_lexical_syntactical* that its output are the extracted features. To start, I defined an array to collect the features, then removing white spaces from the beginning and end of the texts with function named *strip*. The extracted features are listed below:

1. The number of characters of each text (len(text))
2. The number of English characters of each text (isalpha())
3. The number of digits of each text (isdigit())
4. The number of punctuations of each text (string. punctuation)
5. The number of sentences of each text (sent_tokenize from nltk)
6. The average of characters of each sentence
7. The average of words of each sentence (word_tokenize from nltk)
8. The number of sentences whose the first characters are lower case (not isupper())
9. The number of tokens (unigrams) of each text (word_tokenize from nltk)
10. The number of tokens which are not stop words or punctuations
11. The number of above words (from stage 10) whose the first characters are not vowels (a,e,I,o,u)
12. The repetition of alphabets (function: frequency_alphabet)
13. The number of tokens based on its part of speech (function: pos_tag from nltk)

Through each of 1 to 11, I extracted 1 feature, but with 12 and 13, I got 26 and 31 ones, respectively. Overall, there are 68 features.

After focusing on each document, I defined a function named *extract_BoW* and using *countvectorizer* function from sklearn library in this body. Within this function each document is

turned in to a vector with that model our data has been trained, so the output also includes the model.

Step3. Training the model

In this section I have tried to use the support vector machine and k-fold method to train the model. Besides, to evaluate my model I used f1-score measure. In this regard, in the Run file I defined a function named *classify* whose input are texts of documents and the output are the mean of accuracy and the most accurate model.

Step4. Model evaluation:

I defined a function named *evaluate* in Run file. In this function I used f1-score from sklearn and saved the output in a variable called *result*.

To avoid getting confusion in the next referring I defined a *config* file in which all parts can be changed.

Feature Groups

1. **Frequency of letters Group:** It includes 26 English letters that I figured out it is a typical criterion in text mining. I chose it because I felt there is a specific pattern in writing of authors even in letters which they use.
2. **POS tags Group:** it includes 31 part of speech of each token. I think authors have a unique habit to write matters in text. May be some people like use adjectives or adverbs more than others.
3. **Function words frequency Group:** after omitting the stop words and punctuations, some words would be remained that indicate the style of authors. I believe authors persist to use of some specific directory of words. It makes sense for me that the number of words, sentences, tokens and other related features in this group just goes back to the style of each author.
4. **Character features Group:** these are happened in level of characters. It means the number of punctuations, digits and the other language characters is a preference for an author. Some people tent to overuse some of the mentioned characters.

The support vector machine model has been evaluated four times with each time neglecting one of the groups. So, the result is depicted in figure1.

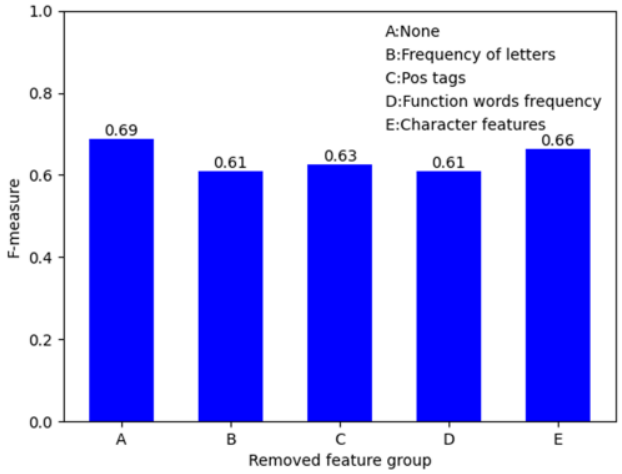

Figure1

The average of accuracy with conserving all groups has been computed 0.69 that is the highest amount of accuracy. In addition, if we do not consider each of two groups (Frequency of letters and Function words frequency) in our model we get the lowest quality of detection. Moreover, group E (Character features) has the lowest influence on our accuracy among these groups.

Evaluation the classifier

As far as I was supposed to get a F1-score higher than 0.5, I would indicate my classifier works well because it shows 0.76 accuracy. The result is reported as a graph in figure2.

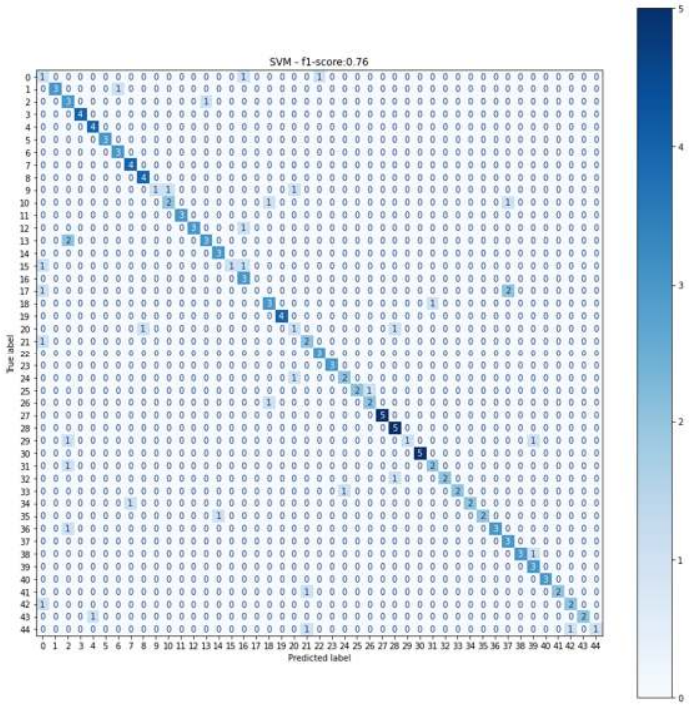

Figure2

Partially, I can mention that using feature selection was a useful way to get the above result. To be more specific, I selected the Support Vector Machine to detect authors because this model performs well when the dimensions of features are higher in corresponding the quantity of observations. The feature selection is a good method when we encounter a considerable amount of data because that items with negligible influence on the model would be eliminated, so time and memory could be managed.

Problems and limitations

1. Through the processing, I realized when some features were added, the model was not working as well as before. Thus, these current features were the best combination that I could reach.
2. I think if the quantity of authors' texts were more, we could have got access more precise model.
3. One of the other problems was related to the length of texts. Those seem shorter than I expected. I was thinking with longer texts could lead to more productive model.
4. Increasing the number of features could have been one of my options if I had additional time with designing more sophisticated model.
5. Semantic analysis is the other feature that could be helpful that I have more time I will probably think about that. Considering some emotional or sentiment features is a way to distinguish writers.
6. The main challenge in this procedure might be judging writers based on their writings. Some authors are inspired by some specific writings, then they start to write the own version but in similar words. Detecting this type of similarity can be challenging.