Authorship Classification

Yazhou Zhang

Data set:

I downloaded 10 English books from Gutenberg.org. For each of these 10 books, I choose every 2000 tokens/strings as one text segment, which can be viewed as a sample for training or testing. In total, I have 2720 samples for all the 10 books.

Train-Test Split:

For the all 2720 samples, I use 5 fold cross validation to split them to 5 folds. Every time 4 folds for training, the remaining 1 fold for testing. And after training and testing in all 5 different cases, I average the accuracy results to get the final result to evaluate my model.

Feature Extraction:

To extract the useful features which can be helpful for authorship identification, here I extract 4 different features from each text segments: Lexical feature, punctuation feature, Bag Of Words feature and Part Of Speech features. Details about these features can be found the markdown sections in the jupyter notebook file.

Classification Method:

Here I choose supervised machine learning method to classify these text segments. I have tried Support Vector Machine (SVM), Random Forest Classifier(RF), K-Nearest Neighbors(KNN) and Multilayer Perceptron(MLP) as the classifier, however, Multilayer Perceptron(MLP) outperforms other classifier in this situation.

The Answers to the questions in the PDF:

1. How would you evaluate the performance of this system?

Answer: To evaluate the performance of my model, I use 5-fold cross validation to split the data set, then train and test this model for independent 5 times, then average the results.

This cross validation can not only show how accurate this predictive model works on an independent and unknown data, but also avoid overfitting.

2. What if the target language is Chinese? Do you need any other data preprocessing?

Answer: The significant difference between Chinese and English is that Chinese has no space between words/characters, so we need to segment the chunks of characters into words first. (There are some Chinese words Segment toolboxes like StandfordSegmenter) After that, we can treat Chinese as English since we can now tokenize them using NLTK.

3. What are good features if you were to solve it in a machine learning fashion?

Answer: I think good features for this problem should be:

(1) Able to capture the unique and distinctive aspects of someone's writing style;
(2) As variant as possible among different authors;
(3) Consistent in different chapters or when the author is writing on different topics

4. Is it possible to avoid manually defining features?

Answer: Yes, it is possible. With the deep learning algorithms, like Convolutional Neural Network(CNN) or Recurrent Neural Network (RNN, LSTM), the deep learning model can automatically extract useful information and learn the sequential/spacial information of the given text data.

Here is a example for RNN based authorship classification:

https://cs224d.stanford.edu/reports/YaoLeon.pdf

5. What if you want to generate a random text that looks like a specific author's work?

Answer: We can use Variational Autoencoder (VAE) for text generation, where both the encoder and decoder are RNNs.

The details can be found in this paper:

https://arxiv.org/pdf/1702.02390.pdf