SirionLabs

Document classification is one of the core classic problems which many of the businesses face in their day to day work. The NLP community is actively pursuing this problem and you'll find new state-of-the-art models being published for the same time to time. Recently, there have been some major breakthroughs like LSTM and transformers which tried solving the problem contextually. These models have proved themselves very effective in document classification as well as in many other downstream tasks. A lot of work has been done on English language but for other languages, the task is still tricky, specifically if we talk about classifying the text without losing its context.

In this assignment we want to implement a document classification algorithm which can classify the Chinese language data. Given a set of text and their respective categories we want to predict the class of a random text from test set with highest possible confidence.

A small dataset containing Chinese news articles along with their respective category is provided. The data contain 2500 news articles and is spread across 10 categories. You are free to choose the methodology of your choice (one or multiple) to go ahead with the problem. You will be expected to share your work in the form of a jupyter notebook and the trained model in pickle/h5 format so that we can test it with the unseen data which we will be holding.

The assignment will be evaluated on the basis of:

1. Data Pre-processing
2. Feature Engineering
3. Modeling Approach
4. Accuracy
5. Coding Standard

In case you are using any large size library/pretrained model, you can share your work using google colab notebook as well. You'll be expected to provide brief explanation of the steps/approach and the reasoning behind it, which can be done either in the jupyter notebook itself or in a separate presentation. You should be able to finish the task in 4-5 days.