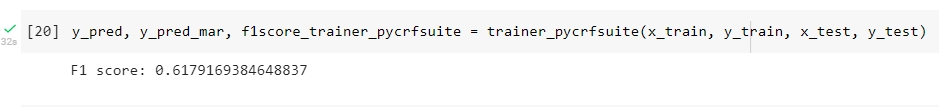
AIMAS Homework 3

F74085319 - 高金生



**Operating environment**

Google Colab with Python 3.8.16

**Program process**

1. Load the data from gdrive
2. Data preprocessing (from sample\_data.txt) into CRF model input format (processed\_data.txt)
3. Choosing the model to train (I use sklearn-crfsuite, python-crfsuite, and pytorch-crfsuite)
4. Splitting data into training and testing
5. Train the data
6. Output the result

**Characteristics of the application**

The characteristic of the program is structured like a machine learning code which are gathering data, preparing data, choosing a model, training, prediction, and evaluation.

**Self-modification**

Here in homework3, I train the data with 2 other models which are python-crfsuite as well as pytorch-crfsuite. In python-crfsuite, I first create a trainer object from CRFSuite.Trainer class and then I feed the trainer object with my training data using trainer.append. After that, I set the parameters and train the model. Using this model I need to calculate the F1score, and I got error when calculating the f1score. After debugging, I found out that I need to convert my label to one-hot encoded format using MultiLabelBinarizer() and then perform the f1\_score function from sklearn.metrics. The second model that I have tried using is pytorch-crfsuite, but I still have error implementing it.

**Experimentation**

Using different models and tweaking with the parameters to use, I got an f1score of 0.62 with the pytorch-crfsuite. The parameters I used here are as follows

Text

Description automatically generated

**Experience**

Through this homework I learned that NER or Named Entity Recognition is a subset of NLP and it involves identifying and classifying named entities in text into predefined categories. On the other hand, Conditional Random Fields or (CRF) is one approach for NER and it is a model that can be used to predict a label sequence given an input sequence.

In this homework I didn’t manage to successfully debug the pytorch-crfsuite. For me, this homework is quite hard because I didn’t used to do NLP or machine learning. But I thought it was a nice experience and awesome skill to know as a computer science student.