Data Mining HW2 – Kaggle competition

胡祐瑄　110065531 NTHU ISA

I do this Kaggle competition in 3 files, which are pre-processing, bert sentiment prediction , and post-processing.

In the pre-processing file, I let the dataset become the result-like dataset, and it contains file index, sentences, and emotions, and these three components can help me to finish the work. And I also tried lots of pre-processing steps, like exclude the punctuation marks or use tf-idf. But the experiment shows that the original one is the best data for this task.

In the bert sentiment prediction file, I use google colab to finish the work because my cuda did not work. In this part, I use the largest model in bert for pre-training model and trained for 2 epochs in fine-tuning because I found that it would over-fitting if I trained too much epochs. However, in the validation step , I found that some class which has little sample would be not accurate. So I wanted to try balance the sample from the original dataset, but I have no time to do it.

In the post-processing file, I just did simple stuffs like tidy up the data to the submission-like file.

And that's all what I did in this homework.