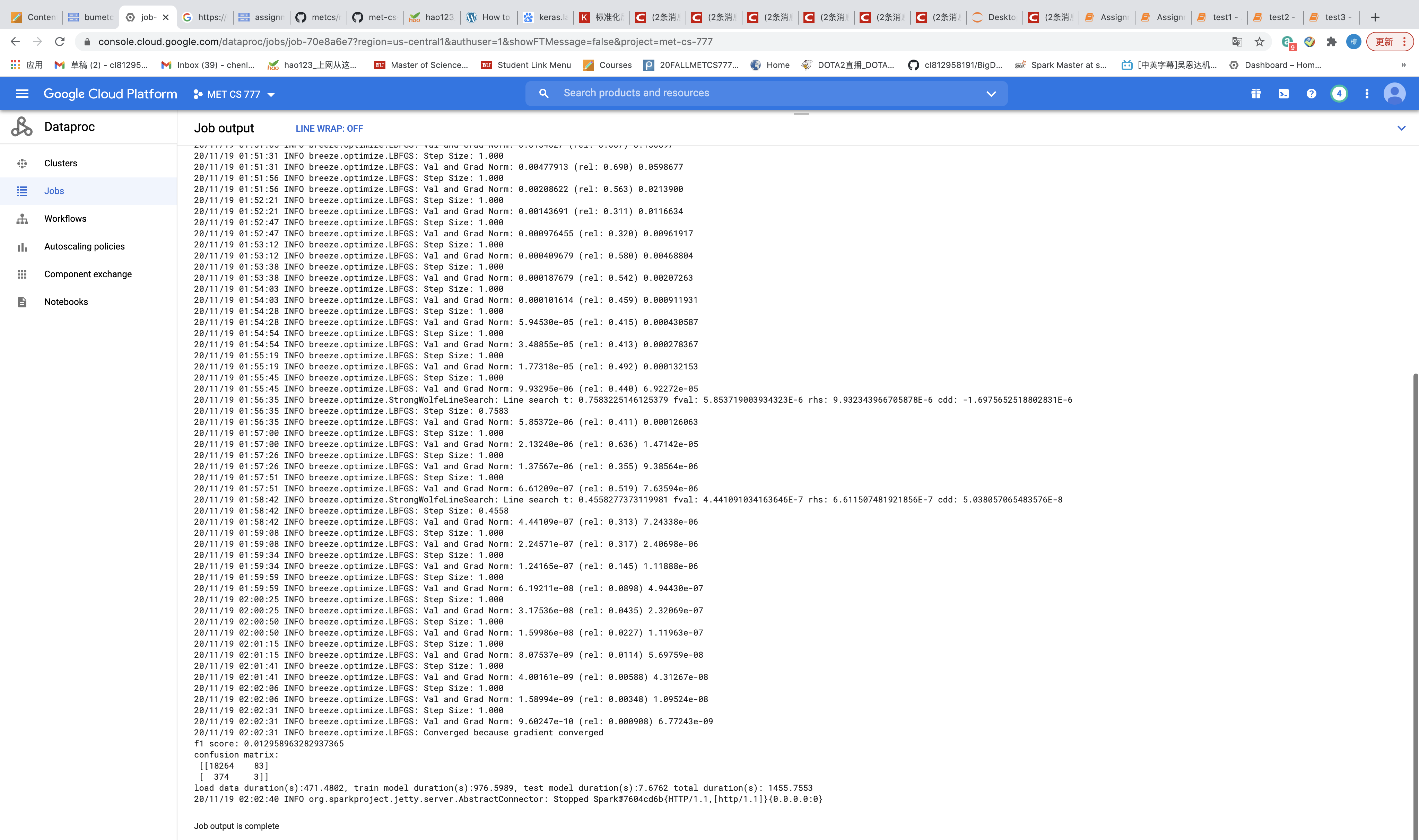
CS777\_Assignment5\_report\_Lin\_Chen

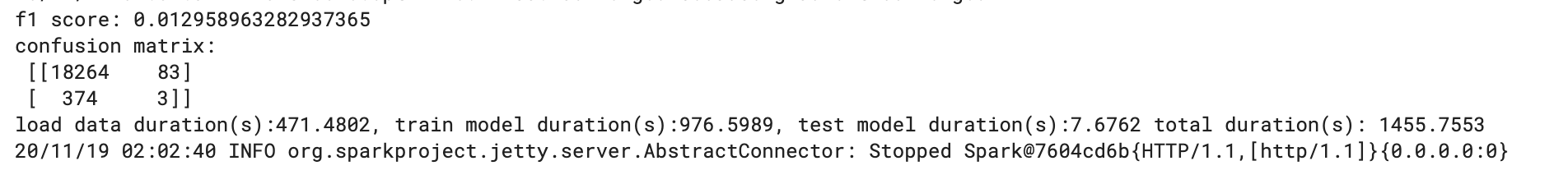
Task1:

I got the TF vectors of the big dataset and the Test dataset firstly.

Then, I use the LogisticRegressionWITHLBFGS in pyspark mllib classification to train the model and test based on test set

The result is as follows:





We can see clearly that the f1 score is only 0.013, I think it is because the TF vector is not as good as TF-IDF vector as features. And we can see that the true negative is low because of the imbalance dataset.

And the time duration is as follows:

Load data: 471.48s (7.86 min)

Train model: 975.6s (16.28min)

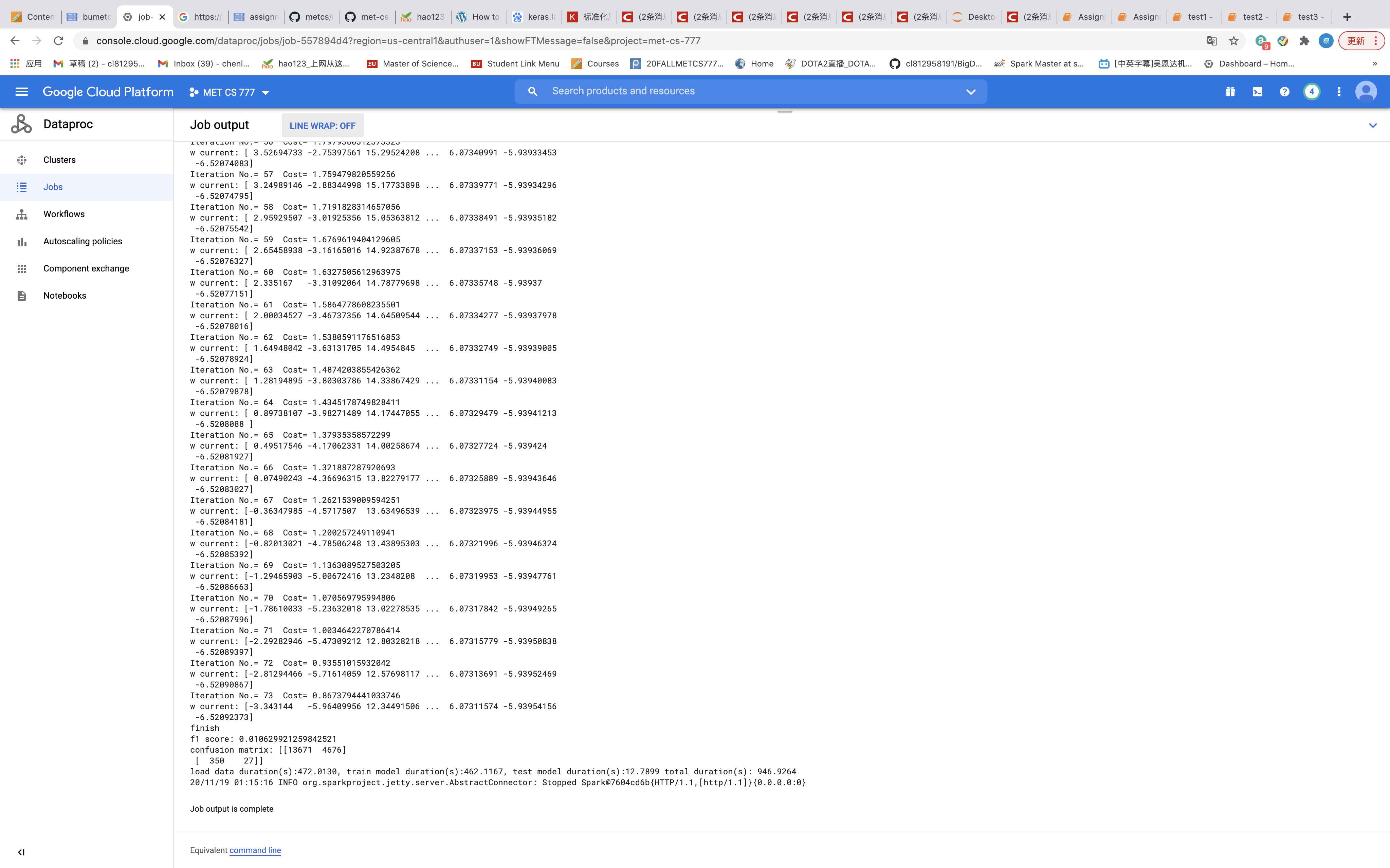
Test mode: 7.68s

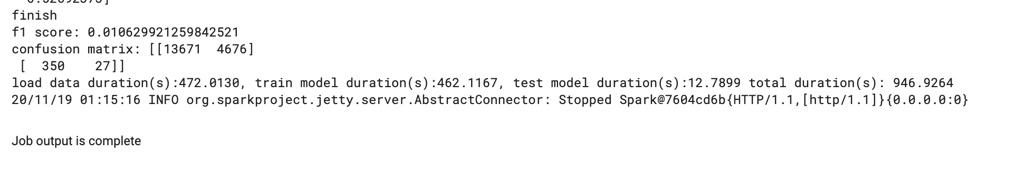
Total duration: 1455.76s (24.26min)

Task2

I set the parameter lambda (regularization coefficient) as 100 and learning rate=0.5, the stop condition difference<=0.00005 and the total epoch is 300.

Output:





We can see that the f1 score is 0.01 because it contains imbalance dataset.

For time duration:

Load data: 472s (7.87min)

Train model: 462s(7.7min)

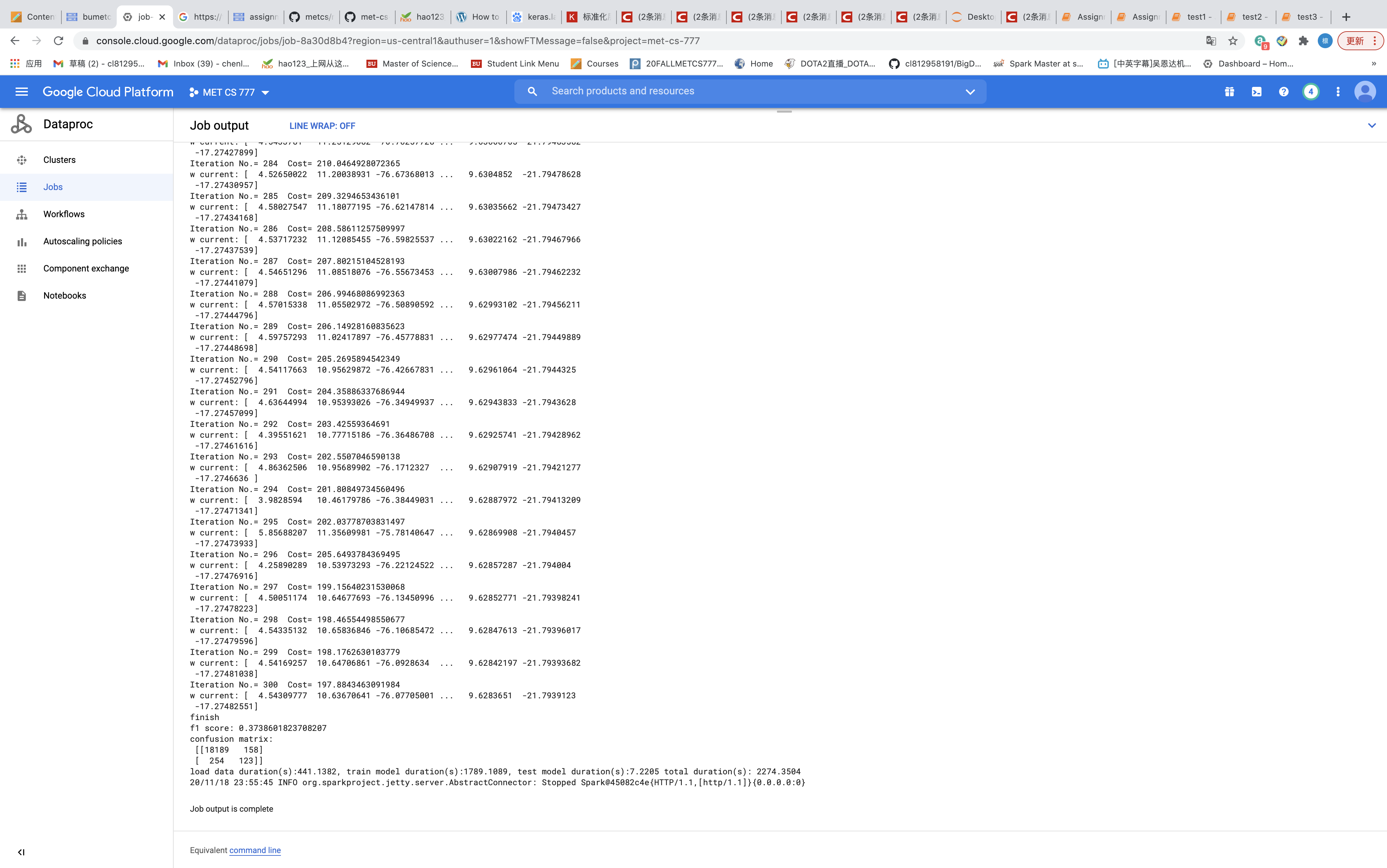
Test model: 12.79s

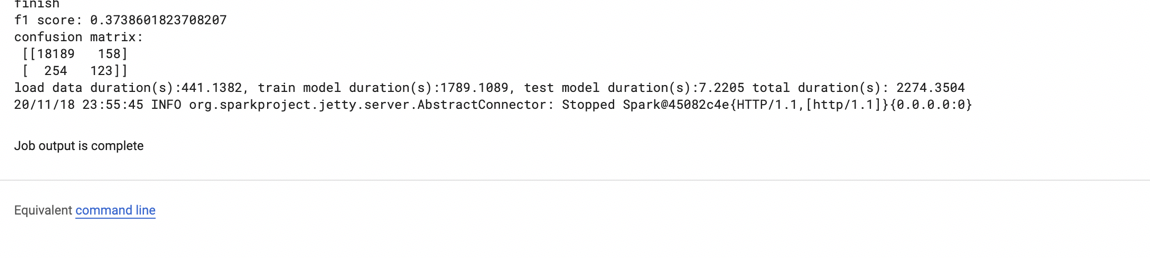
Total duration: 942.93s (15.72min)

Task3:

I use the weighted loss of SVM to train the model:

Output:





We can see that at this time, the f1 score is 0.37, which has increased a lot.

For the time duration:

Load data: 441.14s (7.35min)

Train model: 1789.1s (29.82min)

Test model: 7.22s

Total duration: 2274.35s (37.9min)