CS 777 - Assignment Project

1. Dataset

The dataset is the same as the assignment 4 and 5. The small dataset (37.5 MB of text) is used for training and testing of the model locally. The large training dataset (1.9 GB of text) is used for training the model in the cloud, the large test dataset (200 MB of text) is for test the model in the cloud. The urls of these dataset are listed as follows.

Dataset	Google Cloud Storage
Small Training Dataset (37.5 MB of Text)	gs://metcs777/SmallTrainingData.txt
Large Training Dataset (1.9 GB of Text)	gs://metcs777/TrainingData.txt
Large Text Dataset (200 MB of Text)	gs://metcs777/TestingData.txt

2. Model

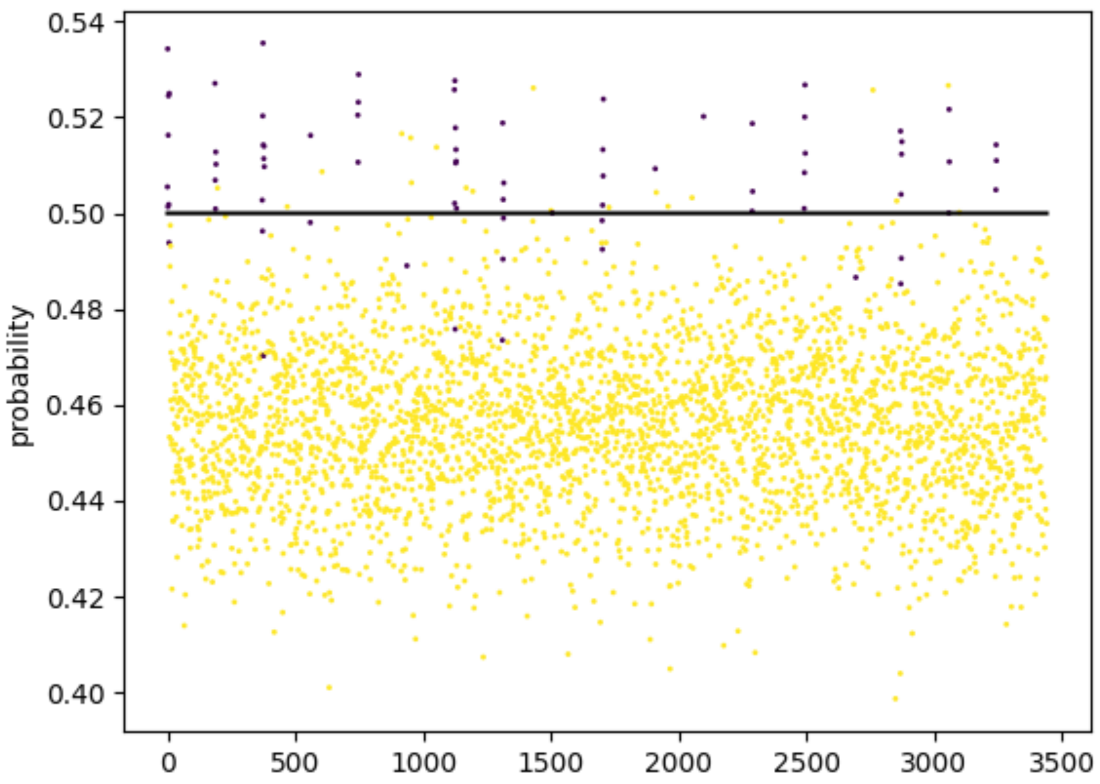
A 2-layer fully connected neural network model has been implemented based on spark. The input unit is 10,000, which is equal to the dimension of the feature; the hidden unit is adjustable and the output unit is set to 1 because this is a binary classification problem. Finally, this neural network classifier can automatically figure out whether a text document is an Australian court case. To run the model, just changing the directory of the dataset and setup the hidden unit parameter.

3. Result

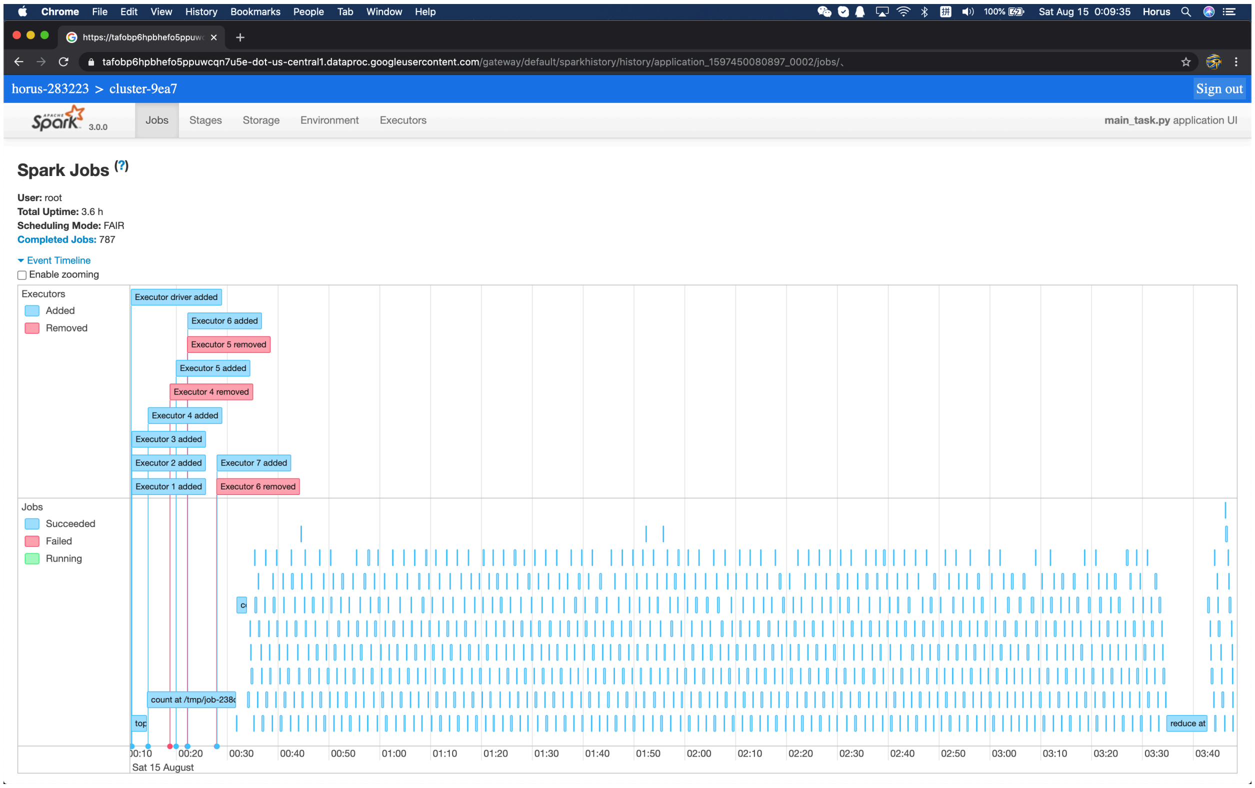
The following table shows some conclusion when the model can give a 0.99+ accuracy and 0.78+ f1 score on small dataset.

Number of Hidden Layer	Stopped at Iteration	Final Cost
1	232	82.3349775453756
4	214	84.9832883962226
10	174	87.4260026633332
32	140	90.8662707029069
100	112	94.4300906893284

Probability plot for small dataset:



Spark history on large dataset:


Chrome File Edit View History Bookmarks People Tab Window Help
100%
Sat Aug 15 0:09:35
Horus
https://tafobp6hpbhefo5ppuw
tafobp6hpbhefo5ppuwcqn7u5e-dot-us-central1.dataproc.googleusercontent.com/gateway/default/sparkhistory/history/application_1597450080897_0002/jobs/
horus-283223 > cluster-9ea7
Sign out
Spark 3.0.0
Jobs
Stages
Storage
Environment
Executors
main_task.py application UI

Spark Jobs (?)

User: root
Total Uptime: 3.6 h
Scheduling Mode: FAIR
Completed Jobs: 787

Event Timeline
☐ Enable zooming