# Wine Quality Prediction Tool User Guidance

Fan Yu

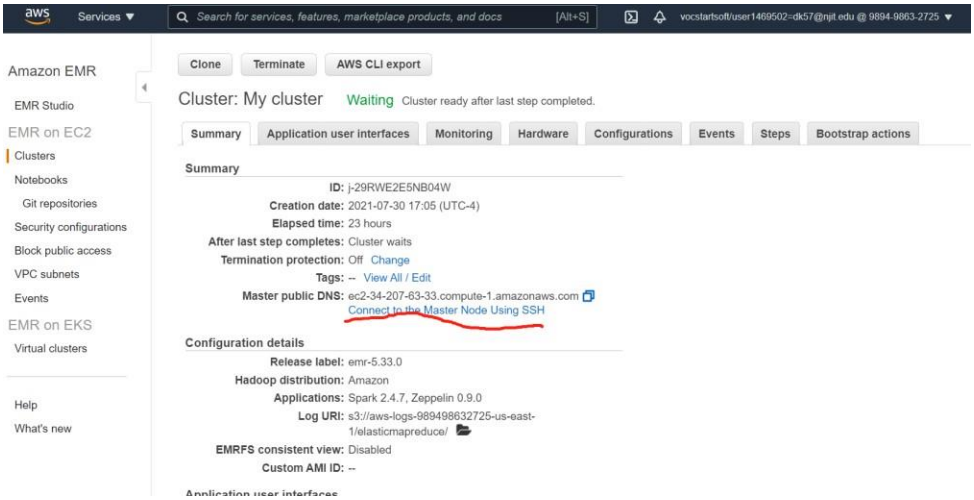
Github link for the code and dockerfile

https://github.com/kakatoto1/CS643.git

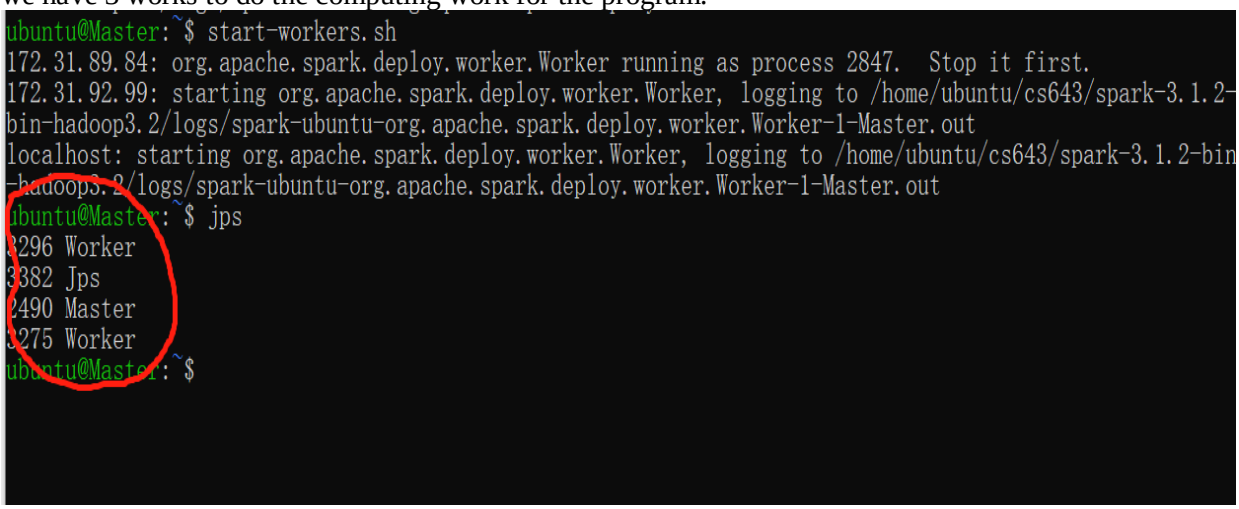
DockerHub link for the docker image
https://hub.docker.com/repository/docker/konaer/cs643

**Part1. Spark, AWS emr cluster training**

1. In AWS, build EMR cluster to run spark, open ssh connection in the cluster in order to connect locally




2. Use ssh connection to upload training.py and training data, from the connection, we could see that we have 3 works to do the computing work for the program.



3. Run training.py, we have the trained_model file build, for us to run prediction latterly. From the running, we could also see that the F-measure for our algorithm is around 0.59.

**Part2. Run predicting.py with our trained model and the data in your hands.**

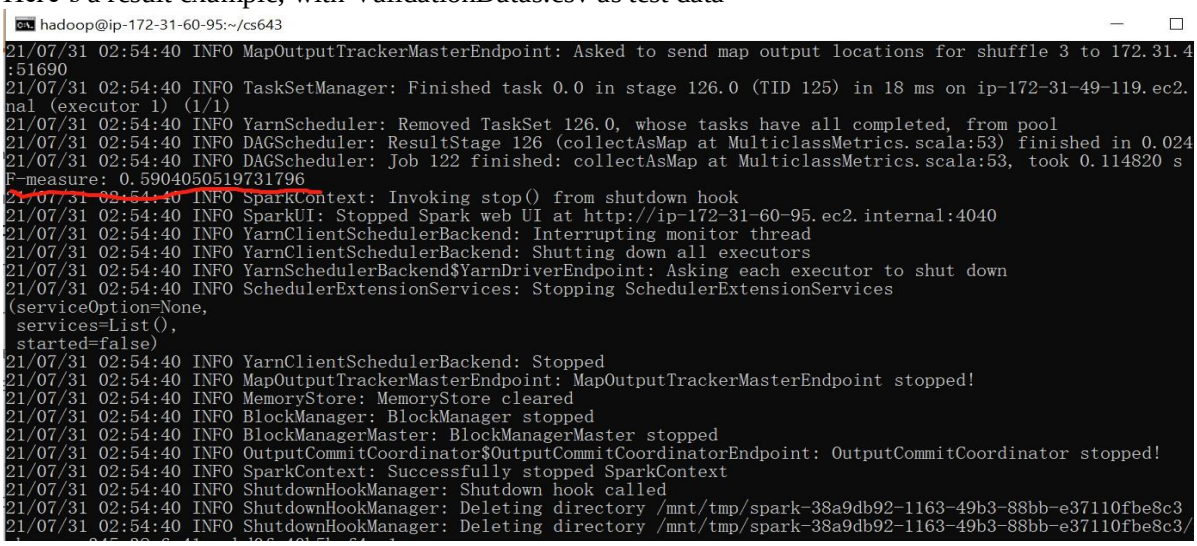
**A. With docker**

1. Please prepare your testdata.csv file to a specific directory that you would like to run.

2. Use following commend line to pull docker image to the same directory of your data file.
   $docker pull konaer/cs643:Dejing

3. Use the following commend line to run docker and test result.
   $docker run konaer/cs643:Dejing <yourTestFileName.csv>

4. Here's a result example, with ValidationDatas.csv as test data




**B. Without Docker**

If you would like to run without docker, you need to follow steps below.

1. Install Java 8 or later
2. Install python 3.6+
3. Install PyNumpy
4. Install Apache Spark from official web or use pip pyspark to install pyspark
5. From github, pull training.py , predicting.py, TrainingDataset.csv to a specific directory
6. Run [python training.py TrainingDataset.csv] first, so that trained_model file could generated.
7. Run [python predicting.py yourTestData.csv], so that you could see prediction result and F measure.