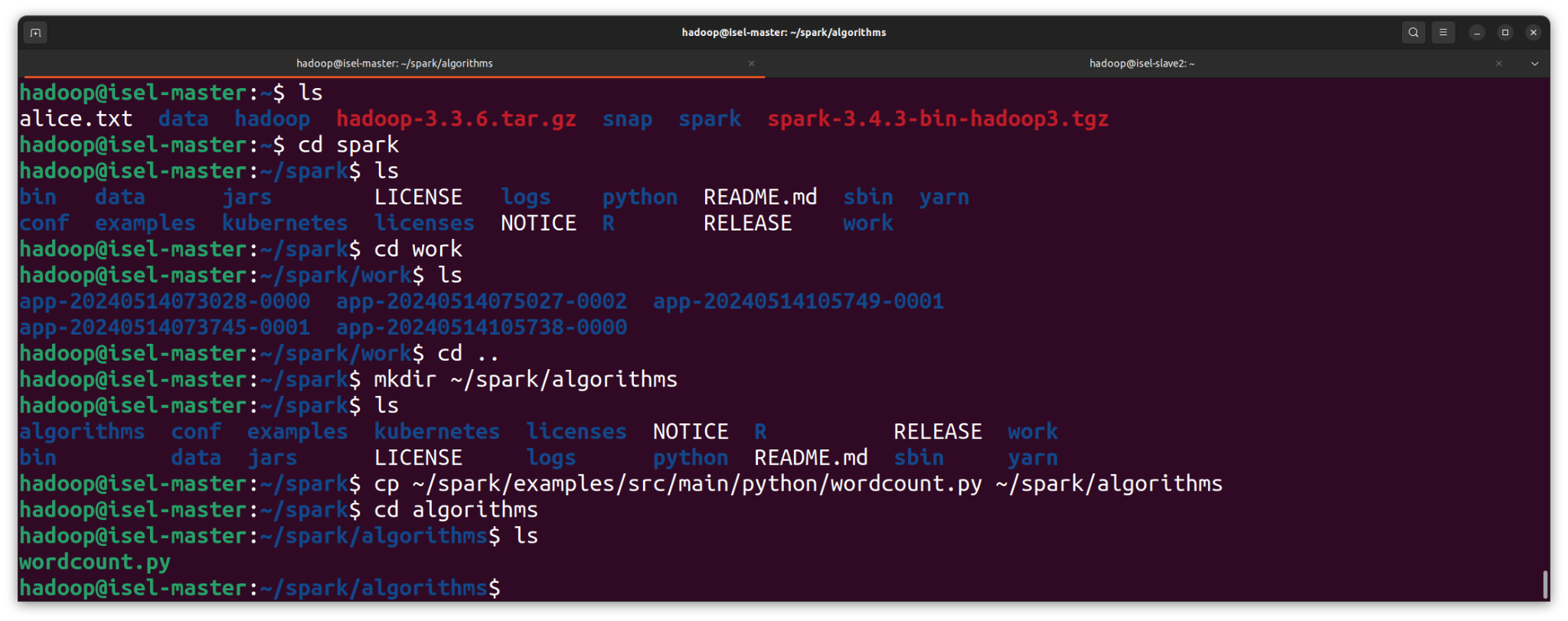
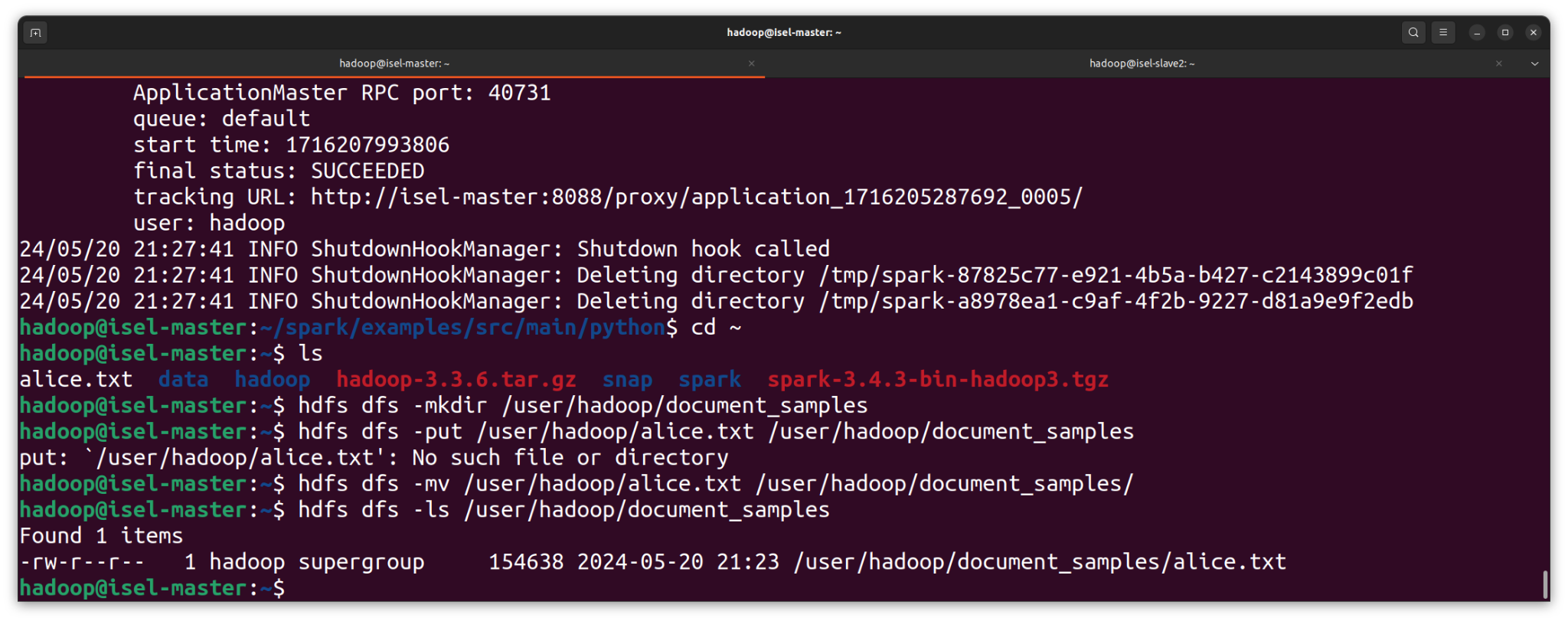
**Submit Spark Jobs**

I have created a directory for keeping the algorithms we submitted onto Spark

*~/spark/algorithms/*

And also a directory for keeping the samples (just for now to test) in HDFS

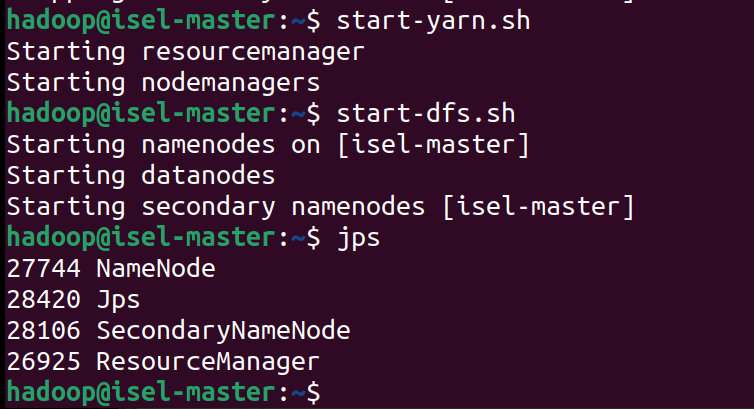
*/user/hadoop/document\_samples/*



***Spark-submit can be run anywhere as the configuration settings***

**Starting Nodes**

Start YARN, DFS and ensure name/datanodes are running



***start-yarn.sh***

***start-dfs.sh***

***Jps***

# **Analysis**

**For non-documents related algorithms**

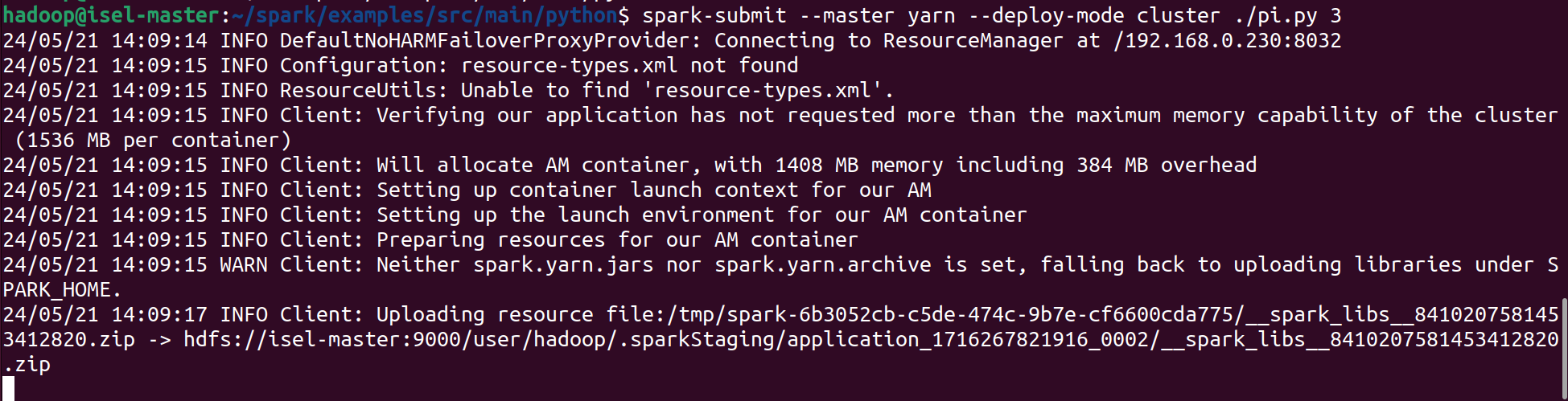
*Just for testing purposes*

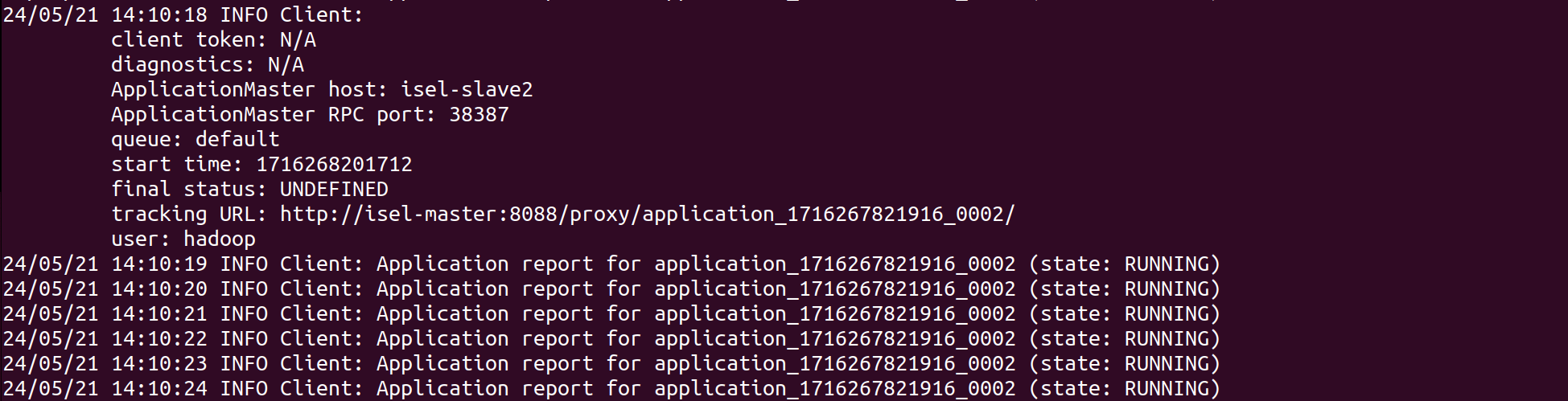
Run example:

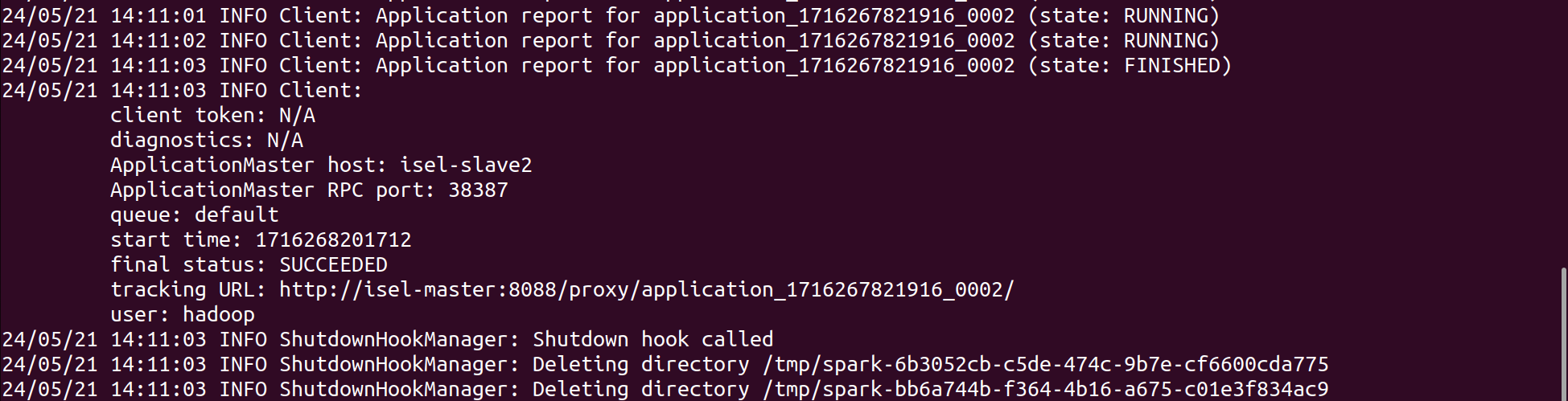
***spark-submit --master yarn --deploy-mode cluster ./pi.py 3***

Put location of algorithm (~/spark/algorithms…) instead of ./pi.py example

Add arguments after if necessary



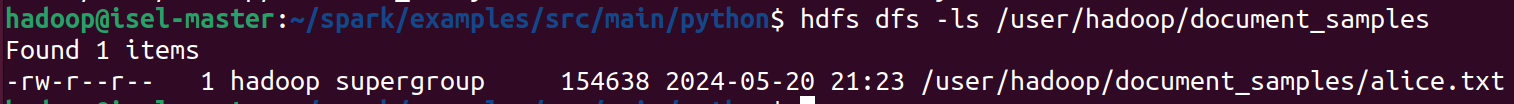




**For document related analysis**

First, check if document is successfully uploaded into HDFS

***hdfs dfs -ls /user/hadoop/document\_samples*** \*directory path



Then run program and provide file path of the document in HDFS (hdfs://…) :

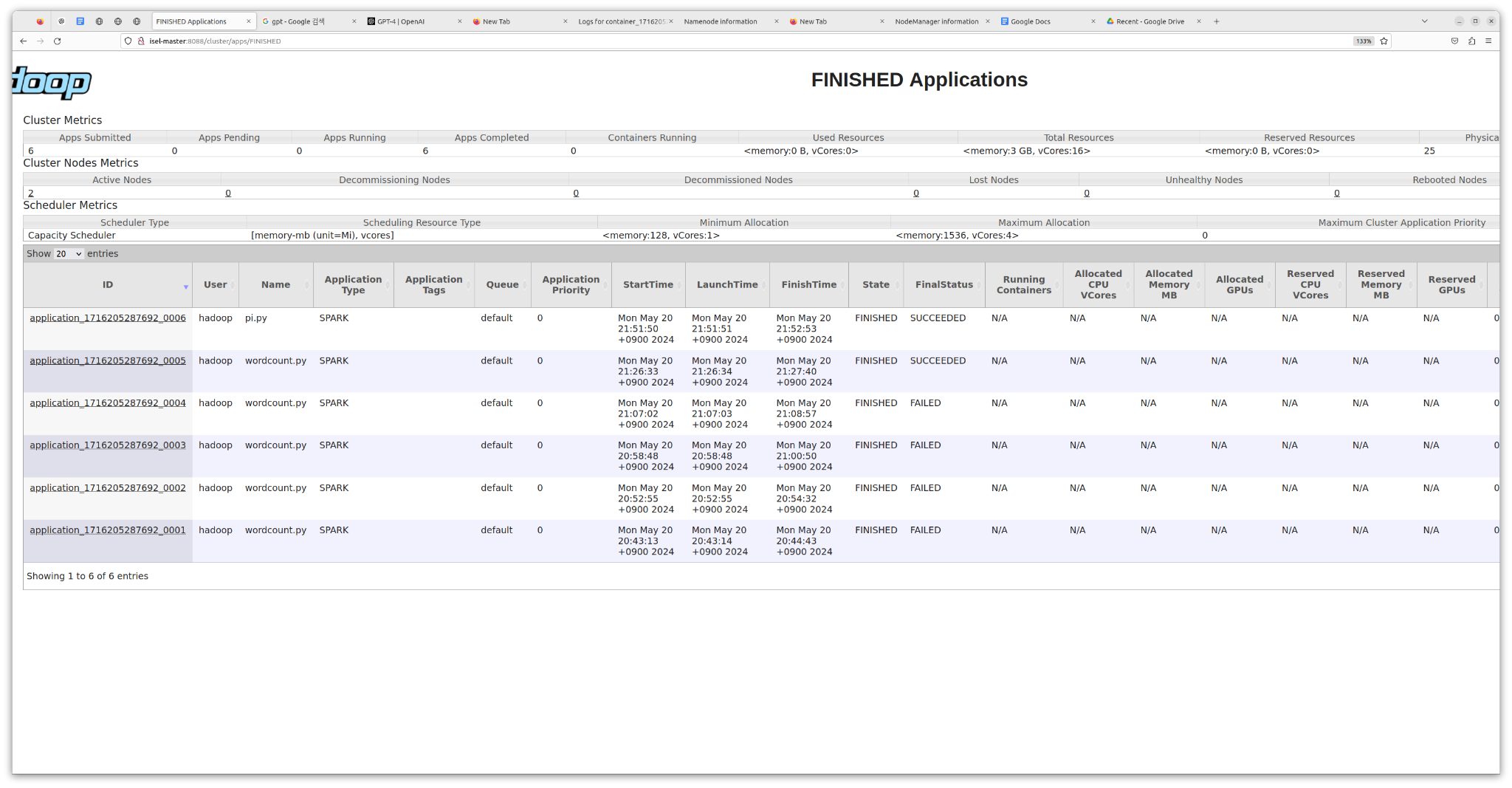
***spark-submit --master yarn --deploy-mode cluster ~/spark/algorithms/wordcount.py /user/hadoop/document\_samples/alice.txt***



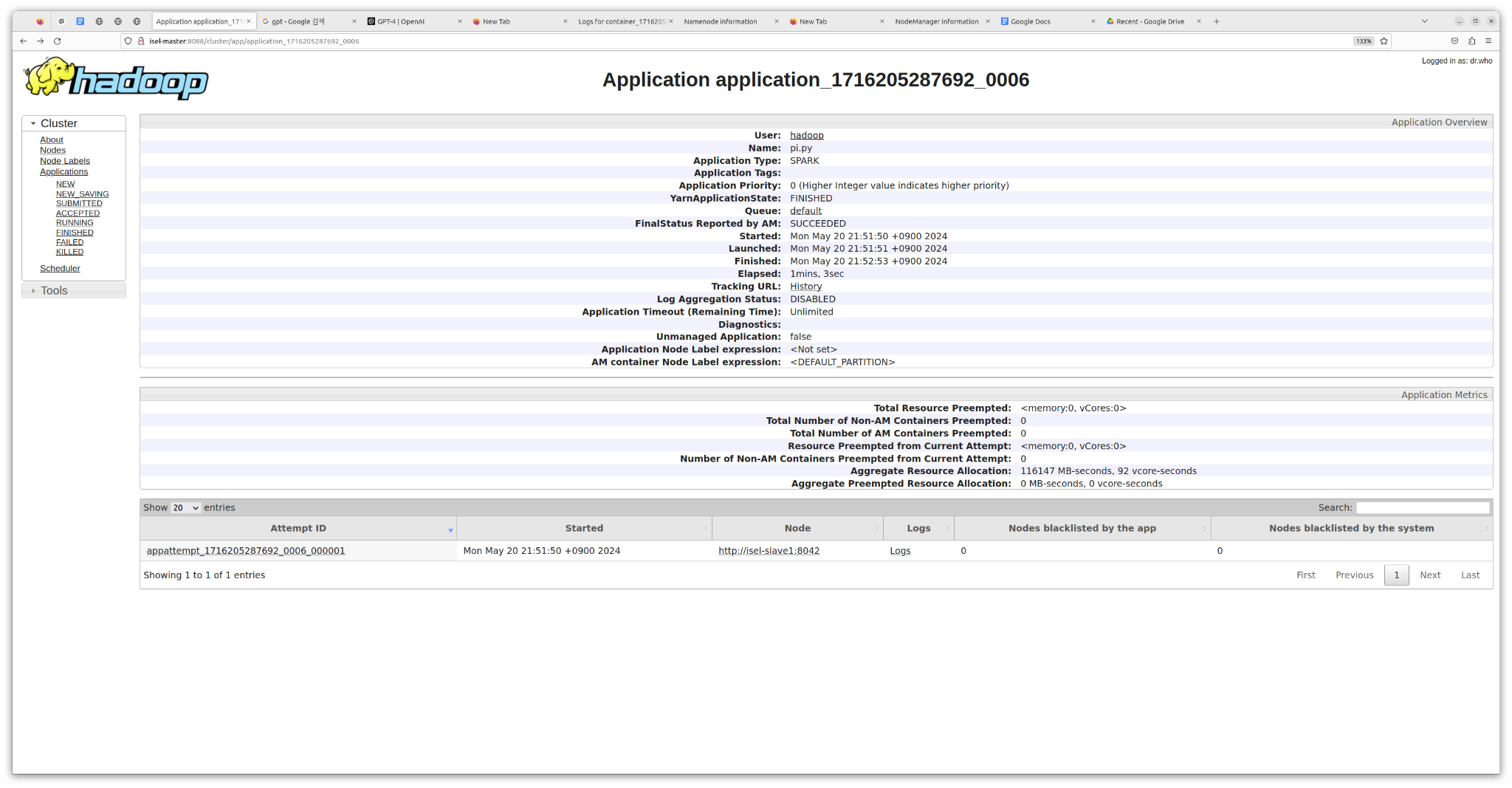
**View Results**

Go to Hadoop web URI:

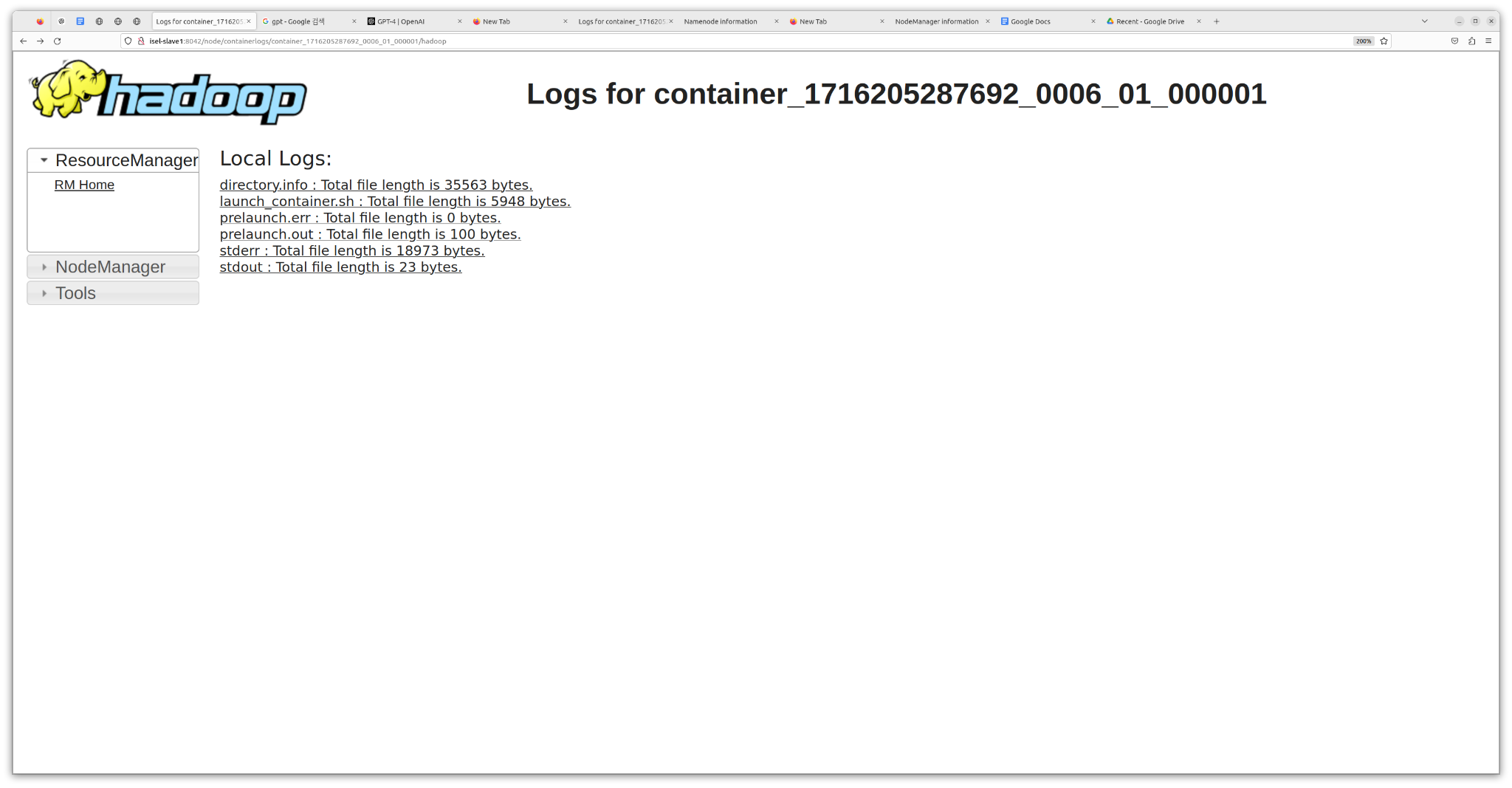
[**http://isel-master:8088/cluster**](http://isel-master:8088/cluster)

****

Click on the Spark application:

****

Click on Logs under the categories



Select stdout

