MVP TEST INSTRUCTION

SETUP ENVIRONMENT(LOCAL):

1. Download Spark from :http://spark.apache.org/downloads.html (version:1.5.2)

please choose the approriate package type according to Hadoop version.

2. To build Spark and its example programs, run: build/mvn

3. Install python 2.7(or Python 3)

RUN ON CLUSTER:

1. Login Cluster:

ssh honeycomb@128.2.7.38 (password: ask teammates)

2. Copy files into CLuster local host:

sudo scp source\_file\_name honeycomb@128.2.7.38:/home/honeycomb/SparkTeam

e.g:

sudo scp /Users/jacobliu/PySpark.py honeycomb@128.2.7.38:/home/honeycomb/SparkTeam

3. Put files into HDFS:

HADOOP\_USER\_NAME=hdfs hdfs dfs -put LOCAL\_FILE\_PATH HDFS\_FILE\_PATH

e.g:

hdfs dfs -put /home/honeycomb/SparkTeam/sample\_multiclass\_classification\_data\_test.txt /user/spark/input

4. Put PySpark.py and train/test dataset into HDFS and run command line:

YOUR\_SEVER\_SPARK\_PATH/spark-submit PySpark.py YOUR\_TRAIN\_DATA\_HDFS\_PATH YOUT\_TEST\_DATA\_HDFS\_PATH YOUR\_OUTPUT\_HDFS\_PATH

e.g:

/bin/spark-submit /home/honeycomb/SparkTeam/PySpark.py /user/spark/input/sample\_multiclass\_classification\_data.txt /user/spark/input/sample\_multiclass\_classification\_data\_test.txt /user/spark/out/

RESULT: SUCCESS!

