Eric Johnson / Ching-Yung Lin
efj2106@columbia.edu

Naïve Bayes Classifier

Prepare the Files

export WORK_DIR=/tmp/mahout-work-ericjohnson
mkdir -p ${WORK_DIR}
cd /usr/local/Cellar/hadoop/2.5.1/bin/
./hadoop dfs mkdir /hw3
./hadoop -mkdir -p ${WORK_DIR}/20news-bydate
cd ${WORK_DIR}/20news-bydate && tar xzf ../20news-bydate.tar.gz && cd .. && cd ..
mkdir ${WORK_DIR}/20news-all
cp -R ${WORK_DIR}/20news-bydate/*/* ${WORK_DIR}/20news-all

Copy to HDFS

cd /usr/local/Cellar/hadoop/2.5.1/bin/
./hadoop dfs -put ${WORK_DIR}/20news-all /hw3/
cd /usr/local/Cellar/hadoop/2.5.1/sbin/mahout-trunk/bin/
./mahout seqdirectory -i /hw3/20news-all -o /hw3/20news-seq –ow

Create Vectors

./mahout seq2sparse -i /hw3/20news-seq -o /hw3/20news-vectors -lnorm -nv -wt tfidf

Split Vectors (Training / Testing)

./mahout split -i /hw3/20news-vectors/tfidf-vectors --trainingOutput /hw3/20news-train-vectors --testOutput /hw3/20news-test-vectors --randomSelectionPct 40 --overwrite --sequenceFiles -xm sequential

Train File – Naïve Bayes

cd /usr/local/Cellar/hadoop/2.5.1/sbin/mahout-trunk/bin/
./hadoop dfs -mkdir /hw3/modelNB
export JAVA_HOME=/System/Library/Java/JavaVirtualMachines/1.6.0.jdk/Contents/Home/ ./mahout trainnb -i /hw3/20news-train-vectors -el -o /hw3/modelNB -li /hw3/labelindex –ow

Testing – Output Confusion Matrix

./mahout testnb -i /hw3/20news-test-vectors -m /hw3/modelNB -l /hw3/labelindex -ow -o /hw3/20news-testingNB