The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup

documents, partitioned (nearly) evenly across 20 different newsgroups. The 20 newsgroups collection has become a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering. We will use Mahout Bayes Classifier to create a model that would classify a new document into one of the 20 newsgroup.

Data Preprocessing Steps:

Generate the input data set for training classifier

$MAHOUT\_HOME/bin/mahout prepare20newsgroups -p

/home/hadoop/Documents/20news-bydate/20news-bydate-train -o 20news-train -a

org.apache.mahout.vectorizer.DefaultAnalyzer -c UTF-8

Generate the input data set for test classifier

$MAHOUT\_HOME/bin/mahout prepare20newsgroups -p

/home/hadoop/Documents/20news-bydate/20news-bydate-test -o 20news-test -a

org.apache.mahout.vectorizer.DefaultAnalyzer -c UTF-8

Run the complementary naïve bayes algorithm:

Train a CBayes Classifier using bi-grams

$MAHOUT\_HOME/bin/mahout trainclassifier --input 20news/20news-bydate-train --

output 20newsmodel --classifierType cbayes

Test a CBayes Classifier using bi-grams

$MAHOUT\_HOME/bin/mahout testclassifier --model 20newsmodel -d 20news/20newstest --classifierType bayes --method mapreduce