# Naive Bayes clustering using Apache Mahout

**Step 1:**
**Create a working directory for the dataset and all input/output on your local drive.**

```
export WORK_DIR=/tmp/mahout-work-${USER}
mkdir -p ${WORK_DIR}
```

In my case it will create mahout-work-cloudera in /tmp folder

**Step 2:**
**Download and extract the 20news-bydate.tar.gz from the 20newsgroups dataset to the working directory from moddle.**

```
chmod 777 20news-bydate.tar.gz
mkdir -p ${WORK_DIR}/20news-bydate
cd /tmp/mahout-work-cloudera
tar -xzf 20news-bydate.tar.gz
```

**Step 3:**
**Move both the folder to 20news-bydate**

```
mv 20news-bydate-test/ 20news-bydate
mv 20news-bydate-train/ 20news-bydate
```

**Step 4:**
**Upload the folder 20news-all on hdfs at path "/user/cloudera/20news-all"**

```
hdfs dfs -put ${WORK_DIR}/20news-bydate /user/cloudera/20news-all
```

OR

```
hdfs dfs -put ${WORK_DIR}/20news-all /user/cloudera/20news-all
```

**Step 5:**
**Convert the full 20 newsgroups dataset into a < Text, Text > SequenceFile**

```
mahout seqdirectory -i /user/cloudera/20news-all -o /user/cloudera/20news-seq
```

**Step 6:**
**Convert and preprocesses the dataset into a < Text, VectorWritable > SequenceFile containing term frequencies for each document.**

```
mahout seq2sparse -i /user/cloudera/20news-seq -o /user/cloudera/20news-vectors -lnorm -nv -wt tfidf
```

**Step 7:**
**Split the preprocessed dataset into training and testing sets.**

```
mahout split -i /user/cloudera/20news-vectors/tfidf-vectors --trainingOutput /user/cloudera/20news-train-vectors --testOutput /user/cloudera/20news-test-vectors --randomSelectionPct 40 --overwrite --sequenceFiles -xm sequential
```

**Step 8:**
**Check all folders created in HDFS.**

```
hdfs dfs -ls l /user/cloudera/
```

**Step 9:**
**Train the classifier.**

```
mahout trainnb -i /user/cloudera/20news-train-vectors -el -o /user/cloudera/model -li /user/cloudera/labelindex -ow -c
```

**Step 10:**
**Test the classifier.**

```
mahout testnb -i /user/cloudera/20news-test-vectors -m /user/cloudera/model   -l /user/cloudera/labelindex -ow -o  /user/cloudera/20news-testing -c
```

**Step 11:**
**Delete the folders and files from HDFS.**

```
hadoop fs -rmr /user/cloudera/20news-seq
```