Classification Demo
Ruichi Yu

The 20 newsgroups dataset 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 the Mahout CBayes classifier to create a model that would classify a new document into one of the 20 newsgroups.

CBayes:
http://mahout.apache.org/users/classification/bayesian.html

Part 1: Use existed model

1. Download Mahout:
https://mahout.apache.org/general/downloads.html

2. Download Mahout-trunk:
git clone git://git.apache.org/mahout.git mahout-trunk

3. For Maven users please include the following snippet in your pom under mahout-trunk forder:

```
<dependency>
    <groupId>org.apache.mahout</groupId>
    <artifactId>mahout-core</artifactId>
    <version>${mahout.version}</version>
</dependency>
```

4. If running Hadoop in cluster mode, start the hadoop daemons by executing the following commands:

```
$ cd $HADOOP_HOME/bin
$ ./start-all.sh
```

Running locally:

```
$ export MAHOUT_LOCAL=true
```

5. Before running, please make sure you have already set up javahome

```
export JAVA_HOME=/Library/Java/Home
```

6. In the trunk directory of Mahout, compile and install Mahout:

```
    $ cd /Users/Rich/Documents/Courses/Fall2014/BigData/mahout-distribution-0.9/mahout-trunk
```

the above is your $MAHOUT_HOME

```
    $ mvn -DskipTests clean install
```

7. Run the 20 newsgroups example script by executing:

```
$ ./examples/bin/classify-20newsgroups.sh
```

8. Please select the algorithm you would like to use. Here we choose 1.

Then you can see the results.

Part 2: Train your own model

1. Set up your path:(very important)

```
export MAHOUT_HOME=/Users/Rich/Documents/Courses/Fall2014/BigData/mahout-distribution-0.9/mahout-trunk/bin

export MAHOUT_CONF_DIR=/Users/Rich/Documents/Courses/Fall2014/BigData/mahout-distribution-0.9/mahout-trunk/src/conf
```

2. Build your working directory

```
export WORK_DIR=/Users/Rich/Documents/Courses/Fall2014/BigData/mahout-distribution-0.9/WorkDir

mkdir -p ${WORK_DIR}
```

3. Download and extract the 20news-bydate.tar.gz from the 20newsgroups dataset to the working directory:

```
curl http://people.csail.mit.edu/jrennie/20Newsgroups/20news-bydate.tar.gz -o ${WORK_DIR}/20news-bydate.tar.gz

$ 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
```

4. Convert the full 20 newsgroups dataset into a < Text, Text > SequenceFile:

```
Important Hint here:
Please use the full path of mahout!!

/Users/Rich/Documents/Courses/Fall2014/BigData/mahout-distribution-0.9/mahout-trunk/bin/mahout seqdirectory -i ${WORK_DIR}/20news-all -o ${WORK_DIR}/20news-seq -ow
```

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

```
/Users/Rich/Documents/Courses/Fall2014/BigData/mahout-distribution-0.9/mahout-trunk/bin/mahout seq2sparse -i ${WORK_DIR}/20news-seq -o ${WORK_DIR}/20news-vectors -lnorm -nv -wt tfidf
```

6.Split the preprocessed dataset into training and testing sets:

```
/Users/Rich/Documents/Courses/Fall2014/BigData/mahout-distribution-0.9/mahout-trunk/bin/mahout split -i ${WORK_DIR}/20news-vectors/tfidf-vectors --trainingOutput ${WORK_DIR}/20news-train-vectors --testOutput ${WORK_DIR}/20news-test-vectors --randomSelectionPct 40 --overwrite --sequenceFiles -xm sequential
```

7.Train the classifier:

```
Important Hint here:
abc is the path you store the labelindex. You can change it to other name

/Users/Rich/Documents/Courses/Fall2014/BigData/mahout-distribution-0.9/mahout-trunk/bin/mahout trainnb -i ${WORK_DIR}/20news-train-vectors -el  -o ${WORK_DIR}/model -li ${WORK_DIR}/abc -ow -c
```

8. Test the classifier:

```
/Users/Rich/Documents/Courses/Fall2014/BigData/mahout-distribution-0.9/mahout-
```

```
trunk/bin/mahout testnb -i ${WORK_DIR}/20news-test-vectors -m ${WORK_DIR}/model -l
${WORK_DIR}/abc -ow -o ${WORK_DIR}/20news-testing -c
```