Mahout 介绍、安装与应用案例

该系列课程是应邀实验楼整理编写的，这里需要赞一下实验楼提供了学习的新方式，可以边看博客边上机实验，课程地址为 https://www.shiyanlou.com/courses/237

【注】该系列所使用到安装包、测试数据和代码均可在百度网盘下载，具体地址为 http://pan.baidu.com/s/10PnDs，下载该 PDF 文件

1 搭建环境

部署节点操作系统为 CentOS，防火墙和 SElinux 禁用，创建了一个 shiyanlou 用户并在系统根目录下创建/app 目录，用于存放 Hadoop 等组件运行包。因为该目录用于安装 hadoop 等组件程序，用户对 shiyanlou 必须赋予 rwx 权限（一般做法是 root 用户在根目录下创建/app 目录，并修改该目录拥有者为 shiyanlou(chown –R shiyanlou:shiyanlou /app)。

Hadoop 搭建环境：

- 虚拟机操作系统： CentOS6.6 64 位，单核，1G 内存
- JDK：1.7.0_55 64 位
- Hadoop：1.1.2

2 Mahout 介绍

Mahout 是 Apache Software Foundation（ASF） 旗下的一个开源项目，提供一些可扩展的机器学习领域经典算法的实现，旨在帮助开发人员更加方便快捷地创建智能应用程序。AMahout 包含许多实现，包括聚类、分类、推荐过滤、频繁子项挖掘。此外，通过使用 Apache Hadoop 库，Mahout 可以有效地扩展到云中。

Mahout 的意思是大象的饲养者及驱赶者。Mahout 这个名称来源于该项目（有时）使用 Apache Hadoop —其徽标上有一头黄色的大象 —来实现可伸缩性和容错性。

Mahout 项目是由 Apache Lucene（开源搜索）社区中对机器学习感兴趣的一些成员发起的，他们希望建立一个可靠、文档翔实、可伸缩的项目，在其中实现一些常见的用于集群和分类的机器学习算法。该社区最初基于 Ng et al. 的文章 “Map-Reduce for Machine Learning on Multicore”（见 参考资料），但此后在发展中又并入了更多广泛的机器学习方法。Mahout 的

目标还包括：

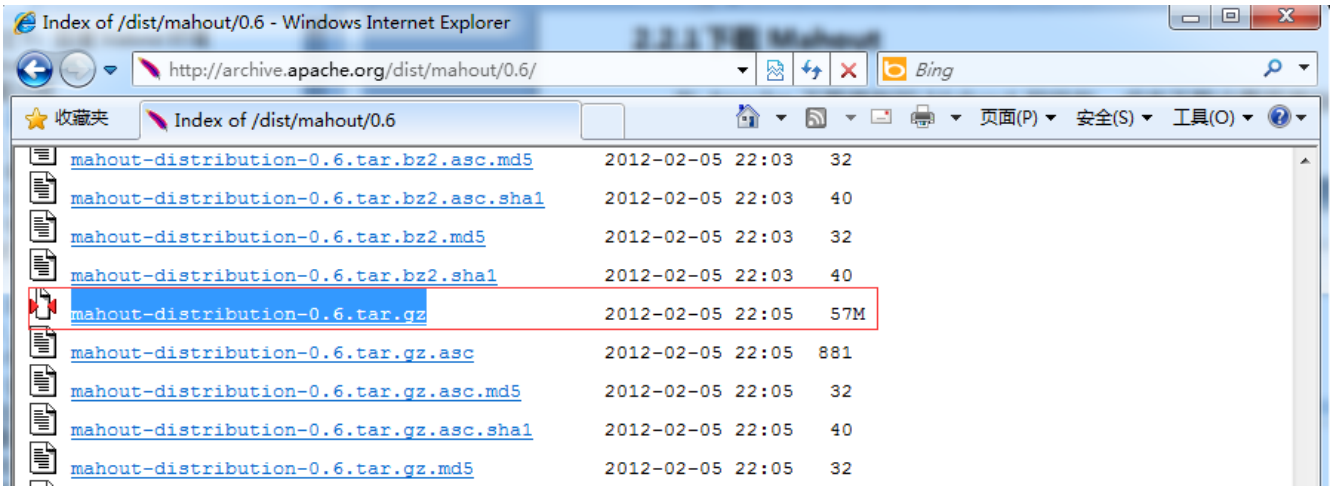
- 建立一个用户和贡献者社区，使代码不必依赖于特定贡献者的参与或任何特定公司和大学的资金。
- 专注于实际用例，这与高新技术研究及未经验证的技巧相反。
- 提供高质量文章和示例

3 搭建 Mahout 环境

3.1 部署过程

3.1.1 下载 Mahout

在 Apache 下载最新的 Mahout 软件包，点击下载会推荐最快的镜像站点，以下为下载地址：http://archive.apache.org/dist/mahout/0.6/

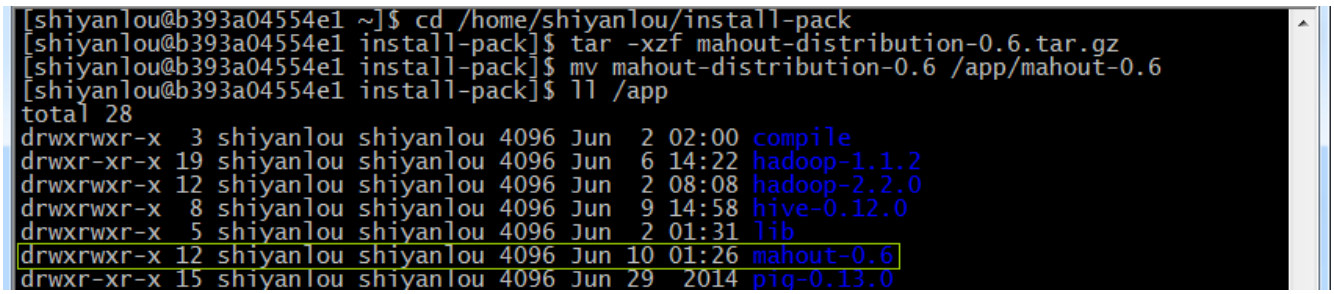


也可以在/home/shiyanlou/install-pack 目录中找到该安装包，解压该安装包并把该安装包复制到/app 目录中

cd /home/shiyanlou/install-pack

tar -xzf mahout-distribution-0.6.tar.gz

mv mahout-distribution-0.6 /app/mahout-0.6



3.1.2 设置环境变量

使用如下命令编辑/etc/profile 文件：

sudo vi /etc/profile

声明 mahout 的 home 路径和在 path 加入 bin 的路径：

export MAHOUT_HOME=/app/mahout-0.6
export MAHOUT_CONF_DIR=/app/mahout-0.6/conf
export PATH=$PATH:$MAHOUT_HOME/bin

```
export HIVE_HOME=/app/hive-0.12.0
export PATH=$PATH:$HIVE_HOME/bin
export CLASSPATH=$CLASSPATH:$HIVE_HOME/bin

export MAHOUT_HOME=/app/mahout-0.6
export MAHOUT_CONF_DIR=/app/mahout-0.6/conf
export PATH=$PATH:$MAHOUT_HOME/bin
~
```

编译配置文件/etc/profile，并确认生效

source /etc/profile
echo $PATH

3.1.3 验证安装完成

重新登录终端，确保 hadoop 集群启动，键入 mahout --help 命令，检查 Mahout 是否安装完好，看是否列出了一些算法：

mahout --help

```
[shiyanlou@b393a04554e1 ~]$ mahout --help
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
no HADOOP_HOME set, running locally
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/app/mahout-0.6/mahout-examples-0.6-job.jar!/org/slf4j/i
mpl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/app/mahout-0.6/lib/slf4j-jcl-1.6.1.jar!/org/slf4j/impl/
StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/app/mahout-0.6/lib/slf4j-log4j12-1.6.1.jar!/org/slf4j/i
mpl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
Unknown program '--help' chosen.
Valid program names are:
  arff.vector: : Generate Vectors from an ARFF file or directory
  baumwelch: : Baum-Welch algorithm for unsupervised HMM training
  canopy: : Canopy clustering
  cat: : Print a file or resource as the logistic regression models would see it
  cleansvd: : Cleanup and verification of SVD output
  clusterdump: : Dump cluster output to text
  clusterpp: : Groups Clustering Output In Clusters
  cmdump: : Dump confusion matrix in HTML or text formats
  cvb: : LDA via Collapsed Variation Bayes (0th deriv. approx)
  cvb0_local: : LDA via Collapsed Variation Bayes, in memory locally.
  dirichlet: : Dirichlet Clustering
```

3.2 测试例子

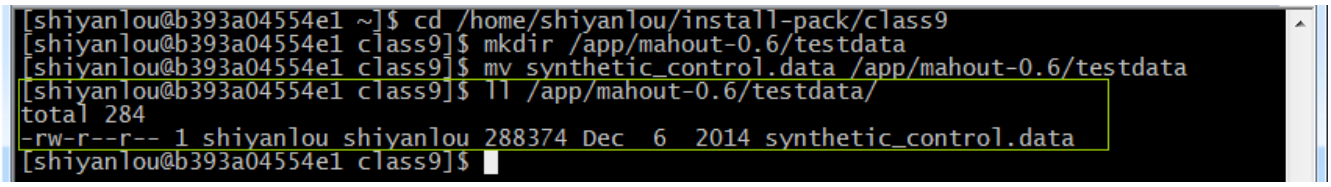
3.2.1 下载测试数据

下载一个文件 synthetic_control.data，下载地址 http://archive.ics.uci.edu/ml/databases/synthetic_control/synthetic_control.data ,也可以在/home/shiyanlou/install-pack/class9 目录中找到该测试数据文件，把这个文件放在 $MAHOUT_HOME/testdata 目录下

cd /home/shiyanlou/install-pack/class9

mkdir /app/mahout-0.6/testdata

mv synthetic_control.data /app/mahout-0.6/testdata



3.2.2 启动 Hadoop

通过下面命令启动 hadoop 并通过 jps 查看进程

cd /app/hadoop-1.1.2/bin

./start-all.sh

jps

3.2.3 使用 kmeans 算法

使用如下命令进行 kmeans 算法测试：

cd /app/mahout-0.6/

mahout org.apache.mahout.clustering.syntheticcontrol.kmeans.Job

```
[shiyanlou@b393a04554e1 ~]$ cd /app/mahout-0.6/
[shiyanlou@b393a04554e1 mahout-0.6]$ mahout org.apache.mahout.clustering.syntheticcontrol.
kmeans.Job
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
no HADOOP_HOME set, running locally
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/app/mahout-0.6/mahout-examples-0.6-job.jar!/org/slf4j/i
mpl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/app/mahout-0.6/lib/slf4j-jcl-1.6.1.jar!/org/slf4j/impl/
StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/app/mahout-0.6/lib/slf4j-log4j12-1.6.1.jar!/org/slf4j/i
mpl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
15/06/10 01:43:38 WARN driver.MahoutDriver: No org.apache.mahout.clustering.syntheticcontr
ol.kmeans.Job.props found on classpath, will use command-line arguments only
15/06/10 01:43:38 INFO kmeans.Job: Running with default arguments
15/06/10 01:43:39 INFO kmeans.Job: Preparing Input
15/06/10 01:43:39 WARN mapred.JobClient: Use GenericOptionsParser for parsing the argument
s. Applications should implement Tool for the same.
15/06/10 01:43:40 INFO input.FileInputFormat: Total input paths to process : 1
15/06/10 01:43:40 INFO mapred.JobClient: Running job: job_local_0001
15/06/10 01:43:40 INFO mapred.Task: Task:attempt_local_0001_m_000000_0 is done. And is in
the process of commiting
15/06/10 01:43:40 INFO mapred.LocalJobRunner:
15/06/10 01:43:40 INFO mapred.Task: Task attempt_local_0001_m_000000_0 is allowed to commi
t now
15/06/10 01:43:40 INFO output.FileOutputCommitter: Saved output of task 'attempt_local_000
1_m_000000_0' to output/data
15/06/10 01:43:41 INFO mapred.JobClient:  map 0% reduce 0%
15/06/10 01:43:43 INFO mapred.LocalJobRunner:
15/06/10 01:43:43 INFO mapred.Task: Task 'attempt_local_0001_m_000000_0' done.
15/06/10 01:43:44 INFO mapred.JobClient:  map 100% reduce 0%
15/06/10 01:43:44 INFO mapred.JobClient: Job complete: job_local_0001
15/06/10 01:43:44 INFO mapred.JobClient: Counters: 8
```

这里需要说明下，当你看到下面的代码时以为是错的，其实不是，原因：MAHOUT_LOCAL：设置是否本地运行，如果设置该参数就不会在 hadoop 运行了，一旦设置这个参数那 HADOOP_CONF_DIR 和 HADOOP_HOME 两个参数就自动失效了。

MAHOUT_LOCAL is not set, so we don't add HADOOP_CONF_DIR to classpath.

no HADOOP_HOME set , running locally

3.2.4 查看结果

结果会在根目录建立 output 新文件夹，如果下图结果表示 mahout 安装正确且运行正常：

cd /app/mahout-0.6/output

ll

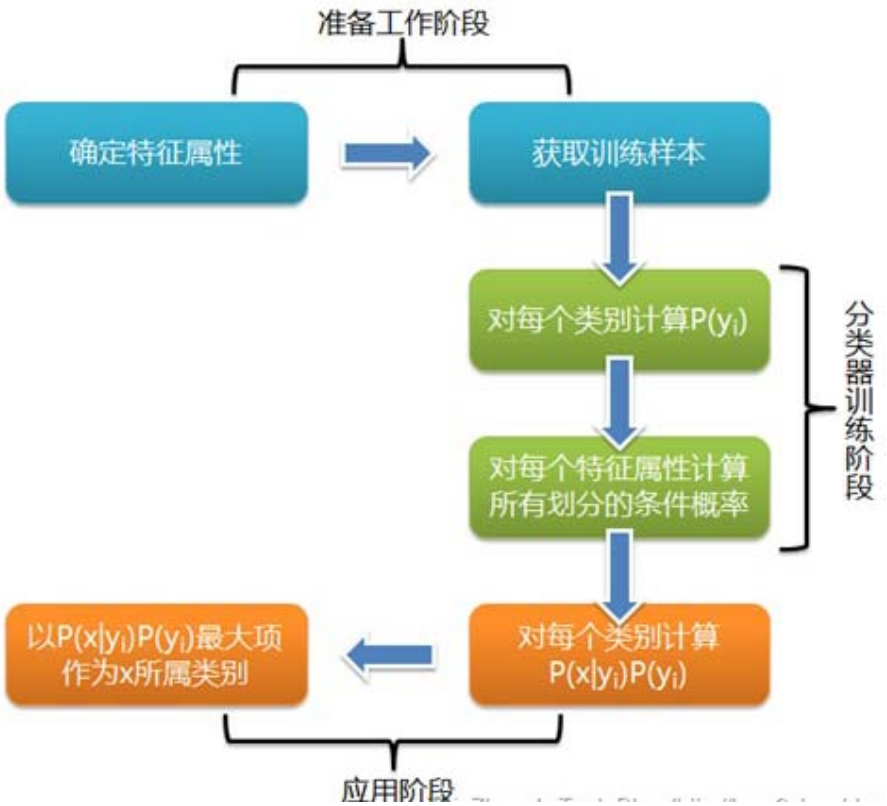
```
[shiyanlou@b393a04554e1 ~]$ cd /app/mahout-0.6/output
[shiyanlou@b393a04554e1 output]$ ll
total 52
drwxrwxr-x 2 shiyanlou shiyanlou 4096 Jun 10 01:44 clusteredPoints
drwxrwxr-x 2 shiyanlou shiyanlou 4096 Jun 10 01:43 clusters-0
drwxrwxr-x 2 shiyanlou shiyanlou 4096 Jun 10 01:43 clusters-1
drwxrwxr-x 2 shiyanlou shiyanlou 4096 Jun 10 01:44 clusters-10-final
drwxrwxr-x 2 shiyanlou shiyanlou 4096 Jun 10 01:43 clusters-2
drwxrwxr-x 2 shiyanlou shiyanlou 4096 Jun 10 01:44 clusters-3
drwxrwxr-x 2 shiyanlou shiyanlou 4096 Jun 10 01:44 clusters-4
drwxrwxr-x 2 shiyanlou shiyanlou 4096 Jun 10 01:44 clusters-5
drwxrwxr-x 2 shiyanlou shiyanlou 4096 Jun 10 01:44 clusters-6
drwxrwxr-x 2 shiyanlou shiyanlou 4096 Jun 10 01:44 clusters-7
drwxrwxr-x 2 shiyanlou shiyanlou 4096 Jun 10 01:44 clusters-8
drwxrwxr-x 2 shiyanlou shiyanlou 4096 Jun 10 01:44 clusters-9
drwxrwxr-x 2 shiyanlou shiyanlou 4096 Jun 10 01:43 data
```

4 测试例子：运行 20newsgroup

4.1 算法流程

朴素贝叶斯分类是一种十分简单的分类算法，朴素贝叶斯的思想基础是这样的：对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率哪个最大，就认为此待分类项属于哪个类别。

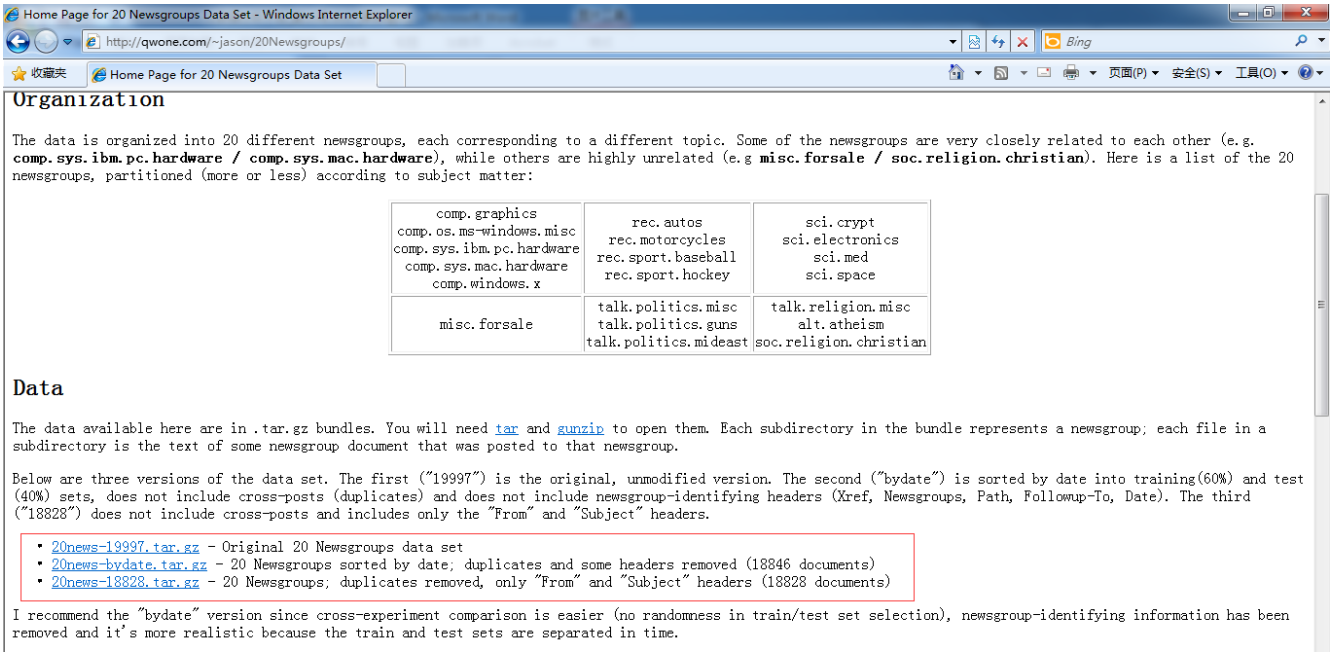
这二十个新闻组数据集合是收集大约 20,000 新闻组文档，均匀的分布在 20 个不同的集合。这 20 个新闻组集合采集最近流行的数据集合到文本程序中作为实验，根据机器学习技术。例如文本分类，文本聚集。我们将使用 Mahout 的 Bayes Classifier 创造一个模型，它将一个新文档分类到这 20 个新闻组集合范例演示



4.2 实现过程（mahout 0.6 版本）

4.2.1 下载数据并解压数据

下载 20Newsgroups 数据集，地址为 http://qwone.com/~jason/20Newsgroups/ ，下载 20news-bydate.tar.gz 数据包，也可以在/home/shiyanlou/install-pack/class9 目录中找到该测试数据文件：



解压 20news-bydate.tar.gz 数据包，解压后可以看到两个文件夹，分别为训练原始数据和测试原始数据：

cd /home/shiyanlou/install-pack/class9

tar -xzf 20news-bydate.tar.gz

```
[shiyanlou@b393a04554e1 ~]$ cd /home/shiyanlou/install-pack/class9
[shiyanlou@b393a04554e1 class9]$ tar -xzf 20news-bydate.tar.gz
[shiyanlou@b393a04554e1 class9]$ ll
total 14140
-rw-r--r--  1 shiyanlou shiyanlou 14464277 Dec  6  2014 20news-bydate.tar.gz
drwxr-xr-x 22 shiyanlou shiyanlou     4096 Mar 18  2003 20news-bydate-test
drwxr-xr-x 22 shiyanlou shiyanlou     4096 Mar 18  2003 20news-bydate-train
[shiyanlou@b393a04554e1 class9]$
```

在 mahout 根目录下建 data 文件夹 ,然后把 20news 训练原始数据和测试原始数据迁移到该文件夹下：

mkdir /app/mahout-0.6/data

mv 20news-bydate-t /app/mahout-0.6/data*

ll /app/mahout-0.6/data

```
[shiyanlou@b393a04554e1 class9]$ mkdir /app/mahout-0.6/data
[shiyanlou@b393a04554e1 class9]$ mv 20news-bydate-t* /app/mahout-0.6/data
[shiyanlou@b393a04554e1 class9]$ ll /app/mahout-0.6/data
total 8
drwxr-xr-x 22 shiyanlou shiyanlou 4096 Mar 18  2003 20news-bydate-test
drwxr-xr-x 22 shiyanlou shiyanlou 4096 Mar 18  2003 20news-bydate-train
[shiyanlou@b393a04554e1 class9]$
```

4.2.2 建立训练集

通过如下命令建立训练集，训练的数据在 20news-bydate-train 目录中，输出的训练集目录为 bayes-train-input：

cd /app/mahout-0.6

*mahout org.apache.mahout.classifier.bayes.PrepareTwentyNewsgroups *
*-p /app/mahout-0.6/data/20news-bydate-train *
*-o /app/mahout-0.6/data/bayes-train-input *
*-a org.apache.mahout.vectorizer.DefaultAnalyzer *
-c UTF-8

```
[shiyanlou@b393a04554e1 ~]$ cd /app/mahout-0.6
[shiyanlou@b393a04554e1 mahout-0.6]$ mahout org.apache.mahout.classifier.bayes.PrepareTwen
tyNewsgroups \
> -p /app/mahout-0.6/data/20news-bydate-train \
> -o /app/mahout-0.6/data/bayes-train-input \
> -a org.apache.mahout.vectorizer.DefaultAnalyzer \
> -c UTF-8
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
no HADOOP_HOME set, running locally
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/app/mahout-0.6/mahout-examples-0.6-job.jar!/org/slf4j/i
mpl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/app/mahout-0.6/lib/slf4j-jcl-1.6.1.jar!/org/slf4j/impl/
StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/app/mahout-0.6/lib/slf4j-log4j12-1.6.1.jar!/org/slf4j/i
mpl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
15/06/10 01:51:09 WARN driver.MahoutDriver: No org.apache.mahout.classifier.bayes.PrepareT
wentyNewsgroups.props found on classpath, will use command-line arguments only
15/06/10 01:51:13 INFO driver.MahoutDriver: Program took 3901 ms (Minutes: 0.0650166666666
6667)
[shiyanlou@b393a04554e1 mahout-0.6]$
```

4.2.3 建立测试集

通过如下命令建立训练集，训练的数据在 20news-bydate-test 目录中，输出的训练集目录为 bayes-test-input：

*mahout org.apache.mahout.classifier.bayes.PrepareTwentyNewsgroups *
*-p /app/mahout-0.6/data/20news-bydate-test *
*-o /app/mahout-0.6/data/bayes-test-input *
*-a org.apache.mahout.vectorizer.DefaultAnalyzer *
-c UTF-8

```
[shiyanlou@b393a04554e1 mahout-0.6]$ mahout org.apache.mahout.classifier.bayes.PrepareTwen
tyNewsgroups \
> -p /app/mahout-0.6/data/20news-bydate-test \
> -o /app/mahout-0.6/data/bayes-test-input \
> -a org.apache.mahout.vectorizer.DefaultAnalyzer \
> -c UTF-8
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
no HADOOP_HOME set, running locally
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/app/mahout-0.6/mahout-examples-0.6-job.jar!/org/slf4j/i
mpl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/app/mahout-0.6/lib/slf4j-jcl-1.6.1.jar!/org/slf4j/impl/
StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/app/mahout-0.6/lib/slf4j-log4j12-1.6.1.jar!/org/slf4j/i
mpl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
15/06/10 01:51:55 WARN driver.MahoutDriver: No org.apache.mahout.classifier.bayes.PrepareT
wentyNewsgroups.props found on classpath, will use command-line arguments only
15/06/10 01:51:58 INFO driver.MahoutDriver: Program took 2645 ms (Minutes: 0.0440833333333
33335)
```

4.2.4 上传数据到 HDFS

在 HDFS 中新建/class9/20news 文件夹，把生成的训练集和测试集上传到 HDFS 的

/class9/20news 目录中：

hadoop fs -mkdir /class9/20news

hadoop fs -put /app/mahout-0.6/data/bayes-train-input /class9/20news

hadoop fs -put /app/mahout-0.6/data/bayes-test-input /class9/20news

hadoop fs -ls /class9/20news

hadoop fs -ls /class9/20news/bayes-test-input

```
[shiyanlou@b393a04554e1 mahout-0.6]$ hadoop fs -mkdir /class9/20news
[shiyanlou@b393a04554e1 mahout-0.6]$ hadoop fs -put /app/mahout-0.6/data/bayes-train-input /class9/20news
[shiyanlou@b393a04554e1 mahout-0.6]$ hadoop fs -put /app/mahout-0.6/data/bayes-test-input /class9/20news
[shiyanlou@b393a04554e1 mahout-0.6]$ hadoop fs -ls /class9/20news
Found 2 items
drwxr-xr-x   - shiyanlou supergroup          0 2015-06-10 01:56 /class9/20news/bayes-test-input
drwxr-xr-x   - shiyanlou supergroup          0 2015-06-10 01:56 /class9/20news/bayes-train-input
[shiyanlou@b393a04554e1 mahout-0.6]$ hadoop fs -ls /class9/20news/bayes-test-input
Found 20 items
-rw-r--r--   1 shiyanlou supergroup     517657 2015-06-10 01:56 /class9/20news/bayes-test-input/alt.atheism.txt
-rw-r--r--   1 shiyanlou supergroup     647979 2015-06-10 01:56 /class9/20news/bayes-test-input/comp.graphics.txt
-rw-r--r--   1 shiyanlou supergroup     541469 2015-06-10 01:56 /class9/20news/bayes-test-input/comp.os.ms-windows.misc.tx
t
-rw-r--r--   1 shiyanlou supergroup     372728 2015-06-10 01:56 /class9/20news/bayes-test-input/comp.sys.ibm.pc.hardware.t
xt
-rw-r--r--   1 shiyanlou supergroup     360374 2015-06-10 01:56 /class9/20news/bayes-test-input/comp.sys.mac.hardware.txt
-rw-r--r--   1 shiyanlou supergroup     596442 2015-06-10 01:56 /class9/20news/bayes-test-input/comp.windows.x.txt
-rw-r--r--   1 shiyanlou supergroup     316487 2015-06-10 01:56 /class9/20news/bayes-test-input/misc.forsale.txt
-rw-r--r--   1 shiyanlou supergroup     404027 2015-06-10 01:56 /class9/20news/bayes-test-input/rec.autos.txt
-rw-r--r--   1 shiyanlou supergroup     374415 2015-06-10 01:56 /class9/20news/bayes-test-input/rec.motorcycles.txt
-rw-r--r--   1 shiyanlou supergroup     470818 2015-06-10 01:56 /class9/20news/bayes-test-input/rec.sport.baseball.txt
-rw-r--r--   1 shiyanlou supergroup     486353 2015-06-10 01:56 /class9/20news/bayes-test-input/rec.sport.hockey.txt
-rw-r--r--   1 shiyanlou supergroup     501244 2015-06-10 01:56 /class9/20news/bayes-test-input/sci.crypt.txt
-rw-r--r--   1 shiyanlou supergroup     394404 2015-06-10 01:56 /class9/20news/bayes-test-input/sci.electronics.txt
-rw-r--r--   1 shiyanlou supergroup     595095 2015-06-10 01:56 /class9/20news/bayes-test-input/sci.med.txt
-rw-r--r--   1 shiyanlou supergroup     546769 2015-06-10 01:56 /class9/20news/bayes-test-input/sci.space.txt
-rw-r--r--   1 shiyanlou supergroup     749414 2015-06-10 01:56 /class9/20news/bayes-test-input/soc.religion.christian.txt
-rw-r--r--   1 shiyanlou supergroup     529051 2015-06-10 01:56 /class9/20news/bayes-test-input/talk.politics.guns.txt
-rw-r--r--   1 shiyanlou supergroup     929163 2015-06-10 01:56 /class9/20news/bayes-test-input/talk.politics.mideast.txt
-rw-r--r--   1 shiyanlou supergroup     626966 2015-06-10 01:56 /class9/20news/bayes-test-input/talk.politics.misc.txt
-rw-r--r--   1 shiyanlou supergroup     438974 2015-06-10 01:56 /class9/20news/bayes-test-input/talk.religion.misc.txt
```

4.2.5 训练贝叶斯分类器

使用 trainclassifier 类训练在 HDFS 中/class9/20news/bayes-train-input 的数据，生成的模型放到/class9/ 20news/newsmodel 目录中：

*mahout trainclassifier *

*-i /class9/20news/bayes-train-input *

*-o /class9/20news/newsmodel *

*-type cbayes *

*-ng 2 *

-source hdfs

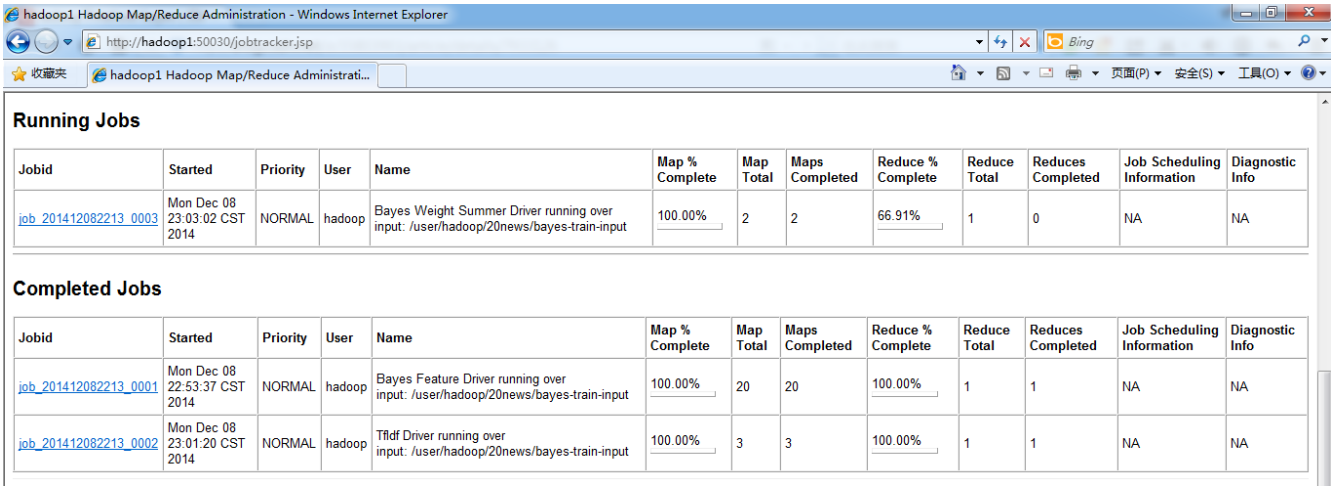
```
[shiyanlou@b393a04554e1 mahout-0.6]$ mahout trainclassifier \
> -i /class9/20news/bayes-train-input \
> -o /class9/20news/newsmodel \
> -type cbayes \
> -ng 2 \
> -source hdfs
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using HADOOP_HOME=/app/hadoop-1.1.2
No HADOOP_CONF_DIR set, using /app/hadoop-1.1.2/conf
MAHOUT-JOB: /app/mahout-0.6/mahout-examples-0.6-job.jar
Warning: $HADOOP_HOME is deprecated.

15/06/10 02:16:27 INFO bayes.TrainClassifier: Training Complementary Bayes Classifier
15/06/10 02:16:28 INFO cbayes.CBayesDriver: Reading features...
15/06/10 02:16:29 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Ap
15/06/10 02:16:30 INFO util.NativeCodeLoader: Loaded the native-hadoop library
15/06/10 02:16:30 WARN snappy.LoadSnappy: Snappy native library not loaded
15/06/10 02:16:30 INFO mapred.FileInputFormat: Total input paths to process : 20
15/06/10 02:16:30 INFO mapred.JobClient: Running job: job_201506090901_0001
15/06/10 02:16:31 INFO mapred.JobClient:  map 0% reduce 0%
15/06/10 02:16:51 INFO mapred.JobClient:  map 2% reduce 0%
15/06/10 02:17:03 INFO mapred.JobClient:  map 4% reduce 0%
15/06/10 02:17:12 INFO mapred.JobClient:  map 6% reduce 0%
15/06/10 02:17:21 INFO mapred.JobClient:  map 7% reduce 0%
15/06/10 02:17:24 INFO mapred.JobClient:  map 9% reduce 0%
```

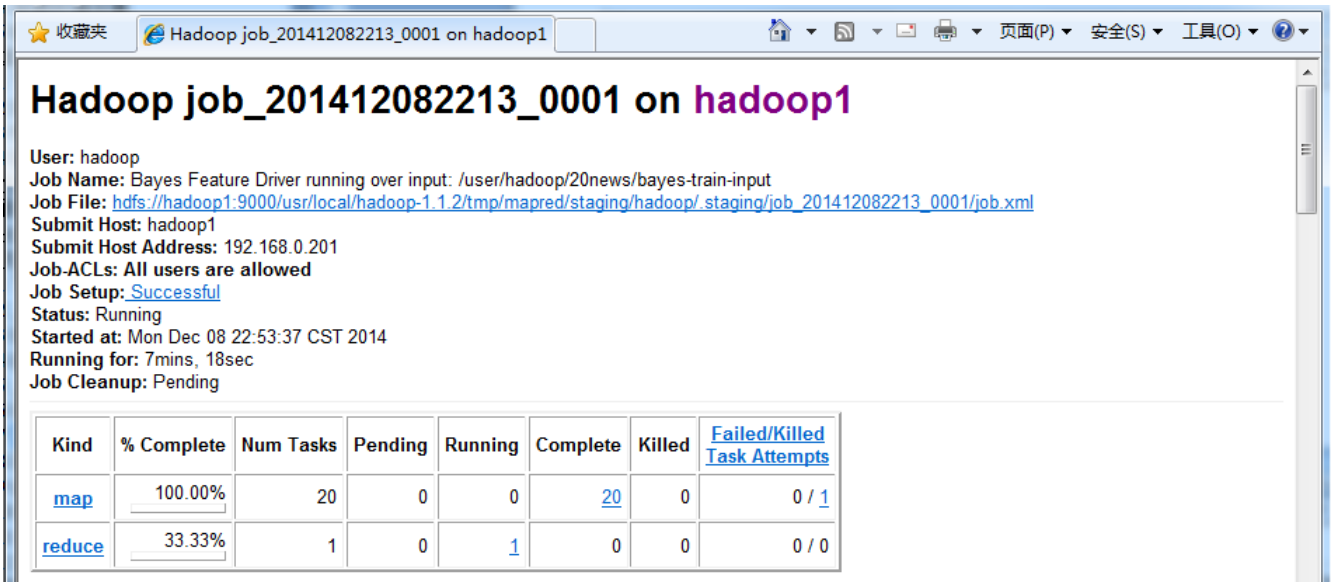
4.2.6 观察训练作业运行过程

注：实验楼为命令行界面，无法观测到该步骤界面，以下描述仅做参考

在训练过程中在 JobTracker 页面观察运行情况，链接地址为 http://**.***.**.***:50030/jobtracker.jsp，训练任务四个作业，大概运行了 15 分钟左右：



点击查看具体作业信息

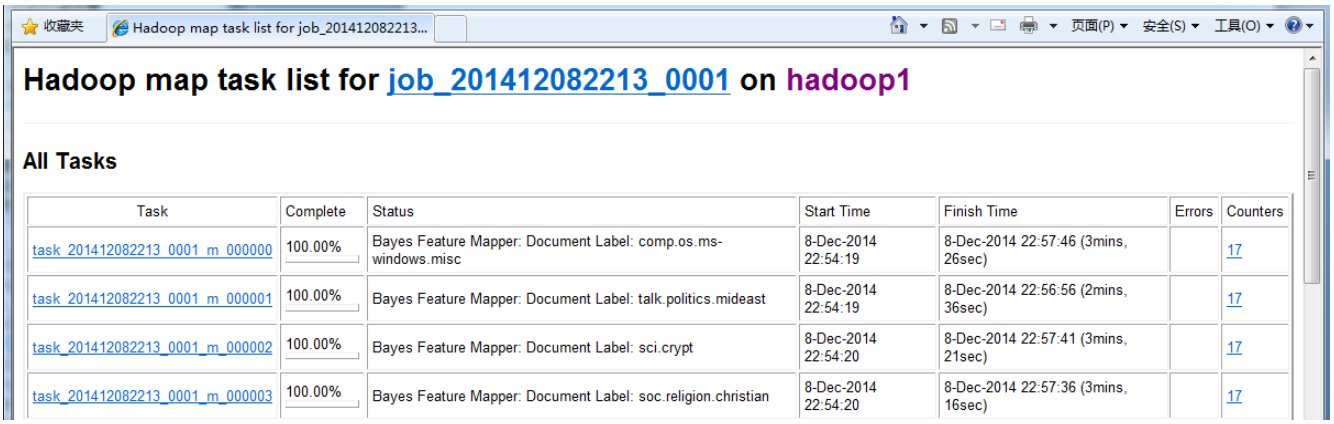


Hadoop job_201412082213_0001 on hadoop1

User: hadoop
Job Name: Bayes Feature Driver running over input: /user/hadoop/20news/bayes-train-input
Job File: hdfs://hadoop1:9000/usr/local/hadoop-1.1.2/tmp/mapred/staging/hadoop/.staging/job_201412082213_0001/job.xml
Submit Host: hadoop1
Submit Host Address: 192.168.0.201
Job-ACLs: All users are allowed
Job Setup: Successful
Status: Running
Started at: Mon Dec 08 22:53:37 CST 2014
Running for: 7mins, 18sec
Job Cleanup: Pending

Kind	% Complete	Num Tasks	Pending	Running	Complete	Killed	Failed/Killed Task Attempts
map	100.00%	20	0	0	20	0	0 / 1
reduce	33.33%	1	0	1	0	0	0 / 0

map 运行情况



Hadoop map task list for job_201412082213_0001 on hadoop1

All Tasks

Task	Complete	Status	Start Time	Finish Time	Errors	Counters
task_201412082213_0001_m_000000	100.00%	Bayes Feature Mapper: Document Label: comp.os.ms-windows.misc	8-Dec-2014 22:54:19	8-Dec-2014 22:57:46 (3mins, 26sec)		17
task_201412082213_0001_m_000001	100.00%	Bayes Feature Mapper: Document Label: talk.politics.mideast	8-Dec-2014 22:54:19	8-Dec-2014 22:56:56 (2mins, 36sec)		17
task_201412082213_0001_m_000002	100.00%	Bayes Feature Mapper: Document Label: sci.crypt	8-Dec-2014 22:54:20	8-Dec-2014 22:57:41 (3mins, 21sec)		17
task_201412082213_0001_m_000003	100.00%	Bayes Feature Mapper: Document Label: soc.religion.christian	8-Dec-2014 22:54:20	8-Dec-2014 22:57:36 (3mins, 16sec)		17

作业运行情况

Hadoop reduce task list for job_201412082213_0001 on hadoop1

All Tasks

Task	Complete	Status	Start Time	Finish Time	Errors	Counters
task_201412082213_0001_r_000000	100.00%	Bayes Feature Reducer: [__WT, rec.sport.baseball, zzzzzzt] => 0.05041903845493945 > reduce	8-Dec-2014 22:56:56	8-Dec-2014 23:01:11 (4mins, 15sec)		15

4.2.7 查看生成模型

通过如下命令查看模型内容：

hadoop fs -ls /class9/20news

hadoop fs -ls /class9/20news/newsmodel

hadoop fs -ls /class9/20news/newsmodel/trainer-tfIdf

```
[shiyanlou@b393a04554e1 mahout-0.6]$ hadoop fs -ls /class9/20news
Warning: $HADOOP_HOME is deprecated.

Found 3 items
drwxr-xr-x   - shiyanlou supergroup          0 2015-06-10 01:56 /class9/20news/bayes-test-input
drwxr-xr-x   - shiyanlou supergroup          0 2015-06-10 01:56 /class9/20news/bayes-train-input
drwxr-xr-x   - shiyanlou supergroup          0 2015-06-10 02:36 /class9/20news/newsmodel
[shiyanlou@b393a04554e1 mahout-0.6]$ hadoop fs -ls /class9/20news/newsmodel
Warning: $HADOOP_HOME is deprecated.

Found 5 items
-rw-r--r--   1 shiyanlou supergroup          0 2015-06-10 02:26 /class9/20news/newsmodel/_SUCCESS
drwxr-xr-x   - shiyanlou supergroup          0 2015-06-10 02:16 /class9/20news/newsmodel/_logs
drwxr-xr-x   - shiyanlou supergroup          0 2015-06-10 02:36 /class9/20news/newsmodel/trainer-tfIdf
drwxr-xr-x   - shiyanlou supergroup          0 2015-06-10 02:36 /class9/20news/newsmodel/trainer-thetaNormalizer
drwxr-xr-x   - shiyanlou supergroup          0 2015-06-10 02:31 /class9/20news/newsmodel/trainer-weights
[shiyanlou@b393a04554e1 mahout-0.6]$ hadoop fs -ls /class9/20news/newsmodel/trainer-tfIdf
Warning: $HADOOP_HOME is deprecated.

Found 3 items
-rw-r--r--   1 shiyanlou supergroup          0 2015-06-10 02:29 /class9/20news/newsmodel/trainer-tfIdf/_SUCCESS
drwxr-xr-x   - shiyanlou supergroup          0 2015-06-10 02:26 /class9/20news/newsmodel/trainer-tfIdf/_logs
drwxr-xr-x   - shiyanlou supergroup          0 2015-06-10 02:29 /class9/20news/newsmodel/trainer-tfIdf/trainer-tfIdf
[shiyanlou@b393a04554e1 mahout-0.6]$
```

4.2.8 测试贝叶斯分类器

使用 testclassifier 类训练在 HDFS 中./20news/bayestest-input 的数据，使用的模型路径为./20news/newsmodel：

```
mahout testclassifier \
-m /class9/20news/newsmodel \
-d /class9/20news/bayes-test-input \
-type cbayes \
-ng 2 \
-source hdfs \
-method mapreduce
```

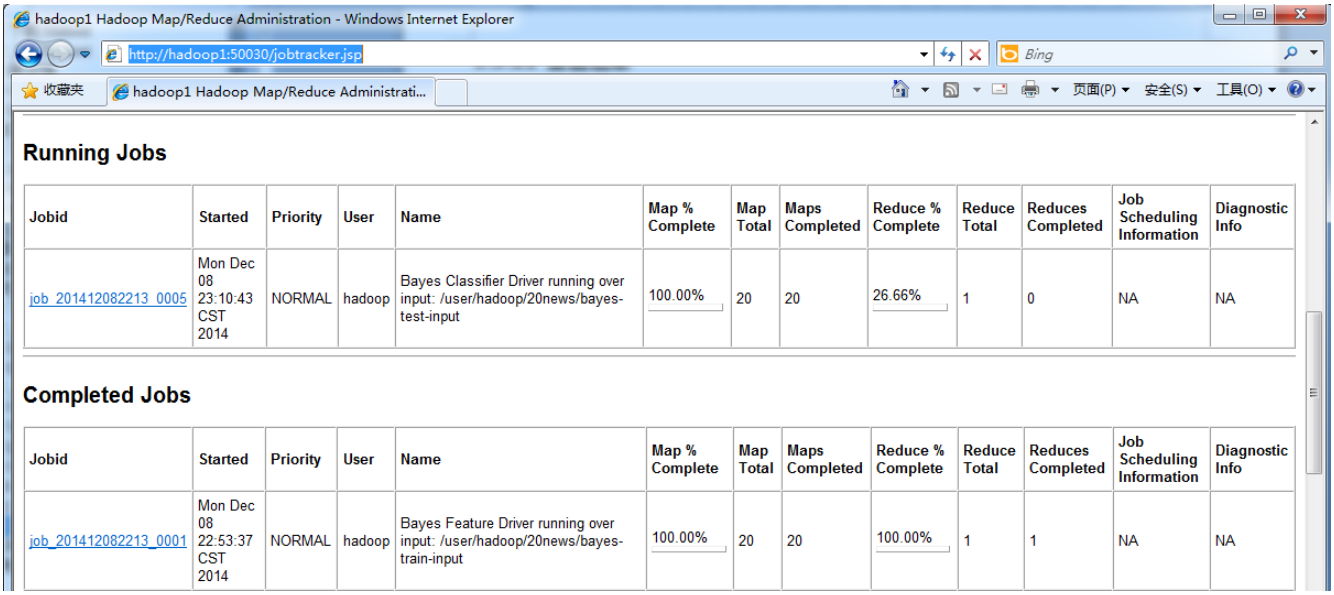
```
[shiyanlou@b393a04554e1 ~]$ mahout testclassifier \
> -m /class9/20news/newsmodel \
> -d /class9/20news/bayes-test-input \
> -type cbayes \
> -ng 2 \
> -source hdfs \
> -method mapreduce
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using HADOOP_HOME=/app/hadoop-1.1.2
No HADOOP_CONF_DIR set, using /app/hadoop-1.1.2/conf
MAHOUT-JOB: /app/mahout-0.6/mahout-examples-0.6-job.jar
Warning: $HADOOP_HOME is deprecated.

15/06/10 02:56:47 INFO common.HadoopUtil: Deleting /class9/20news/bayes-test-input-output
15/06/10 02:56:48 WARN mapred.JobClient: Use GenericOptionsParser for parsing the argumen
ts. Applications should implement Tool for the same.
15/06/10 02:56:49 INFO util.NativeCodeLoader: Loaded the native-hadoop library
15/06/10 02:56:49 WARN snappy.LoadSnappy: Snappy native library not loaded
15/06/10 02:56:49 INFO mapred.FileInputFormat: Total input paths to process : 20
15/06/10 02:56:50 INFO mapred.JobClient: Running job: job_201506100247_0003
15/06/10 02:56:51 INFO mapred.JobClient:  map 0% reduce 0%
```

4.2.9 观察训练作业运行过程

注：实验楼为命令行界面，无法观测到该步骤界面，以下描述仅做参考

在执行过程中在 JobTracker 页面观察运行情况，链接地址为 http://hadoop:50030/jobtracker.jsp，训练任务 1 个作业，大概运行了 5 分钟左右：


hadoop1 Hadoop Map/Reduce Administration - Windows Internet Explorer

http://hadoop1:50030/jobtracker.jsp

hadoop1 Hadoop Map/Reduce Administrati...

Running Jobs

Jobid	Started	Priority	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total	Reduces Completed	Job Scheduling Information	Diagnostic Info
job_201412082213_0005	Mon Dec 08 23:10:43 CST 2014	NORMAL	hadoop	Bayes Classifier Driver running over input: /user/hadoop/20news/bayes-test-input	100.00%	20	20	26.66%	1	0	NA	NA

Completed Jobs

Jobid	Started	Priority	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total	Reduces Completed	Job Scheduling Information	Diagnostic Info
job_201412082213_0001	Mon Dec 08 22:53:37 CST 2014	NORMAL	hadoop	Bayes Feature Driver running over input: /user/hadoop/20news/bayes-train-input	100.00%	20	20	100.00%	1	1	NA	NA

作业的基本信息

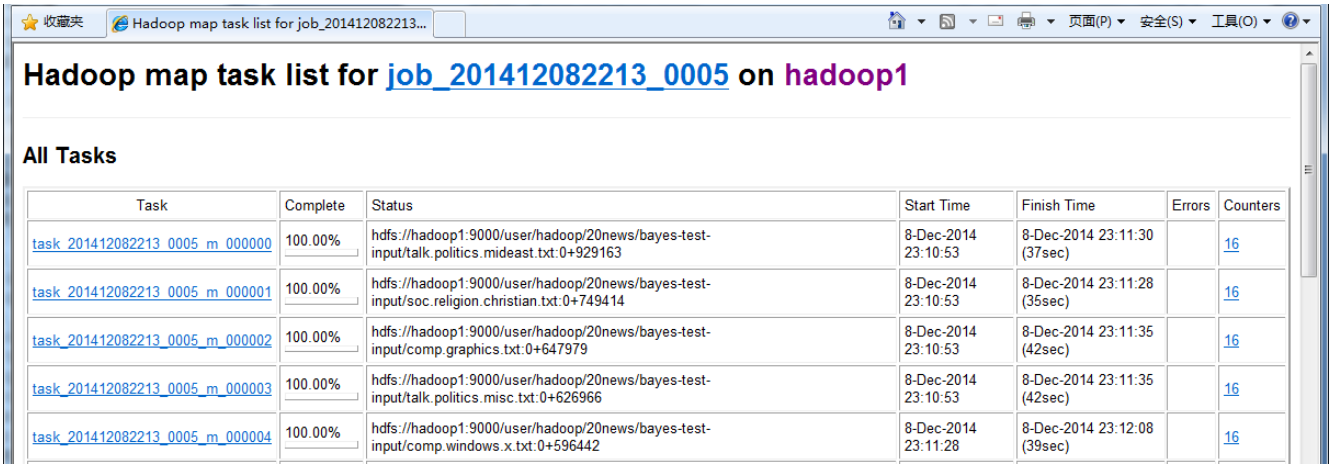

Hadoop job_201412082213_0005 on hadoop1

Hadoop job_201412082213_0005 on hadoop1

User: hadoop
Job Name: Bayes Classifier Driver running over input: /user/hadoop/20news/bayes-test-input
Job File: hdfs://hadoop1:9000/usr/local/hadoop-1.1.2/tmp/mapred/staging/hadoop/.staging/job_201412082213_0005/job.xml
Submit Host: hadoop1
Submit Host Address: 192.168.0.201
Job-ACLs: All users are allowed
Job Setup: Successful
Status: Succeeded
Started at: Mon Dec 08 23:10:43 CST 2014
Finished at: Mon Dec 08 23:14:00 CST 2014
Finished in: 3mins, 17sec
Job Cleanup: Successful

Kind	% Complete	Num Tasks	Pending	Running	Complete	Killed	Failed/Killed Task Attempts
map	100.00%	20	0	0	20	0	0 / 1
reduce	100.00%	1	0	0	1	0	0 / 0

map 运行情况


Hadoop map task list for job_201412082213...

Hadoop map task list for job_201412082213_0005 on hadoop1

All Tasks

Task	Complete	Status	Start Time	Finish Time	Errors	Counters
task_201412082213_0005_m_000000	100.00%	hdfs://hadoop1:9000/user/hadoop/20news/bayes-test-input/talk.politics.mideast.txt:0+929163	8-Dec-2014 23:10:53	8-Dec-2014 23:11:30 (37sec)		16
task_201412082213_0005_m_000001	100.00%	hdfs://hadoop1:9000/user/hadoop/20news/bayes-test-input/soc.religion.christian.txt:0+749414	8-Dec-2014 23:10:53	8-Dec-2014 23:11:28 (35sec)		16
task_201412082213_0005_m_000002	100.00%	hdfs://hadoop1:9000/user/hadoop/20news/bayes-test-input/comp.graphics.txt:0+647979	8-Dec-2014 23:10:53	8-Dec-2014 23:11:35 (42sec)		16
task_201412082213_0005_m_000003	100.00%	hdfs://hadoop1:9000/user/hadoop/20news/bayes-test-input/talk.politics.misc.txt:0+626966	8-Dec-2014 23:10:53	8-Dec-2014 23:11:35 (42sec)		16
task_201412082213_0005_m_000004	100.00%	hdfs://hadoop1:9000/user/hadoop/20news/bayes-test-input/comp.windows.x.txt:0+596442	8-Dec-2014 23:11:28	8-Dec-2014 23:12:08 (39sec)		16

reduce 运行情况

Hadoop reduce task list for job_201412082213_0005 on hadoop1

All Tasks

Task	Complete	Status	Start Time	Finish Time	Errors	Counters
task_201412082213_0005_r_000000	100.00%	Bayes Classifier Reducer: [__CT, talk.religion.misc, talk.religion.misc] => 95.0 > reduce	8-Dec-2014 23:11:28	8-Dec-2014 23:13:56 (2mins, 27sec)		15

4.2.10 查看结果

这个混合矩阵的意思说明：上述 a 到 u 分别是代表了有 20 类别，这就是我们之前给的 20 个输入文件，列中的数据说明每个类别中被分配到的字节个数，classified 说明应该被分配到的总数 381 0 0 0 0 9 1 0 0 0 1 0 0 2 0 1 0 0 3 0 0 | 398 a = rec.motorcycles 意思为 rec.motorcycles 本来是属于 a，有 381 篇文档被划为了 a 类，这个是正确的数据，其它的分别表示划到 b~u 类中的数目。我们可以看到其正确率为 381/398=0.9573 ,可见其正确率还是很高的了。

```
Confusion Matrix
-------------------------------------------------------
a       b       c       d       e       f       g       h       i       j       k       l       m       n       o       p
        q       r       s       t       <--Classified as
393     0       0       0       0       2       1       0       0       0       1       0       1       0       0       0
        0       0       0       0       |  398          a     = rec.motorcycles
0       335     0       0       0       0       0       0       1       6       6       0       2       28      0       1
        3       1       4       8       |  395          b     = comp.windows.x
0       0       366     0       0       0       1       1       0       0       0       0       1       3       1
        0       3       0       0       |  376          c     = talk.politics.mideast
1       1       1       340     1       1       2       0       0       2       0       3       0       1       0       3
        6       1       1       0       |  364          d     = talk.politics.guns
0       0       5       22      95      1       5       0       0       6       0       4       0       0       64      4
        0       44      0       1       |  251          e     = talk.religion.misc
13      0       0       2       0       356     1       0       1       1       6       0       7       1       0       1
        0       0       6       1       |  396          f     = rec.autos
1       0       0       0       0       0       382     11      0       1       0       0       0       0       1       0
        0       0       1       0       |  397          g     = rec.sport.baseball
0       0       0       0       0       0       4       394     1       0       0       0       0       0       0       0
        0       0       0       0       |  399          h     = rec.sport.hockey
1       3       0       0       0       4       4       1       329     3       15      1       9       2       0       0
        1       0       8       4       |  385          i     = comp.sys.mac.hardware
0       3       1       1       0       2       1       0       1       369     0       1       3       3       0       6
        2       0       1       0       |  394          j     = sci.space
1       0       0       0       0       1       0       0       18      5       296     0       21      10      0       0
        3       0       16      21      |  392          k     = comp.sys.ibm.pc.hardware
0       0       9       99      2       0       1       4       3       7       0       166     0       1       3       4
        8       2       1       0       |  310          l     = talk.politics.misc
9       1       2       0       0       2       3       0       6       9       16      0       301     6       2       5
        14      1       9       7       |  393          m     = sci.electronics
0       21      2       0       0       1       6       1       7       5       12      0       8       294     2       2
        9       0       6       13      |  389          n     = comp.graphics
1       0       0       0       2       0       2       1       0       2       1       0       0       1       381     2
        0       3       0       2       |  398          o     = soc.religion.christian
0       0       3       1       1       1       4       2       0       1       1       2       9       3       7       3
56      1       1       3       0       |  396          p     = sci.med
0       1       0       1       0       2       3       0       3       0       0       0       0       0       0       2
        382     0       2       0       |  396          q     = sci.crypt
1       0       9       4       3       0       2       0       2       6       2       1       0       0       41      1
0       3       235     0       0       |  319          r     = alt.atheism
5       1       0       0       0       11      3       4       11      2       20      0       8       2       0       6
        0       0       313     4       |  390          s     = misc.forsale
4       14      0       0       0       3       6       3       4       7       43      0       1       21      3       2
        6       1       6       270     |  394          t     = comp.os.ms-windows.misc
```

4.3 实现过程（mahout 0.7+版本）

在 0.7 版本的安装目录下 $MAHOUT_HOME/examples/bin 下有个脚本文件 classifu-20newsgroups.sh，这个脚本中执行过程是和前面分布执行结果是一致的，只不过将各个 API 用 shell 脚本封装到一起了。从 0.7 版本开始，Mahout 移除了命令行调用的 API：prepare20newsgroups、trainclassifier 和 testclassifier，只能通过 shell 脚本执行。

执行 $MAHOUT_HOME/examples/bin/classify-20newsgroups.sh 四个选项中选择第

一个选项，

```
[hadoop@hadoop1 bin]$
[hadoop@hadoop1 bin]$ ls
classify-20newsgroups.sh  cluster-syntheticcontrol.sh  factorize-netflix.sh  README.txt
cluster-reuters.sh        factorize-movielens-1M.sh    lda.algorithm
[hadoop@hadoop1 bin]$ ./classify-20newsgroups.sh
Please select a number to choose the corresponding task to run
1. cnaivebayes
2. naivebayes
3. sgd
4. clean -- cleans up the work area in /tmp/mahout-work-hadoop
Enter your choice : 1
ok. You chose 1 and we'll use cnaivebayes
creating work directory at /tmp/mahout-work-hadoop
Downloading 20news-bydate
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
  5 13.7M    5  757k    0     0  25177      0  0:09:34  0:00:30  0:09:04 61782
```

执行结果如下：

```
=======================================================
Confusion Matrix
-------------------------------------------------------
a    b    c    d    e    f    g    h    i    j    k    l    m    n    o    p    q    r    s    t    u    <--Classified as
381  0    0    0    0    9    1    0    0    0    1    0    0    2    0    1    0    0    3    0    0     |  398  a     = rec.motorcycles
1    284  0    0    0    0    1    0    6    3    11   0    66   3    0    1    6    0    4    9    0     |  395  b     = comp.windows.x
2    0    339  2    0    3    5    1    0    0    0    0    1    1    12   1    7    0    2    0    0     |  376  c     = talk.politics.mideast
4    0    1    327  0    2    2    0    0    2    1    1    0    5    1    4    12   0    2    0    0     |  364  d     = talk.politics.guns
7    0    4    32   27   7    7    2    0    12   0    0    6    0    100  9    7    31   0    0    0     |  251  e     = talk.religion.misc
10   0    0    0    0    359  2    2    0    1    3    0    1    6    0    1    0    0    11   0    0     |  396  f     = rec.autos
0    0    0    0    0    1    383  9    1    0    0    0    0    0    0    0    0    0    3    0    0     |  397  g     = rec.sport.baseball
1    0    0    0    0    0    9    382  0    0    0    0    1    1    1    0    2    0    2    0    0     |  399  h     = rec.sport.hockey
2    0    0    0    0    4    3    0    330  4    4    0    5    12   0    0    2    0    12   7    0     |  385  i     = comp.sys.mac.hardware
0    3    0    0    0    0    1    0    0    368  0    0    10   4    1    3    2    0    2    0    0     |  394  j     = sci.space
0    0    0    0    0    3    1    0    27   2    291  0    11   25   0    0    1    0    13   18   0     |  392  k     = comp.sys.ibm.pc.hardware
8    0    1    109  0    6    11   4    1    18   0    98   1    3    11   10   27   1    1    0    0     |  310  l     = talk.politics.misc
0    11   0    0    0    3    6    0    10   6    11   0    299  13   0    2    13   0    7    8    0     |  389  m     = comp.graphics
6    0    1    0    0    4    2    0    5    2    12   0    8    321  0    4    14   0    8    6    0     |  393  n     = sci.electronics
2    0    0    0    0    0    4    1    0    3    1    0    3    1    372  6    0    2    1    2    0     |  398  o     = soc.religion.christian
4    0    0    1    0    2    3    3    0    4    2    0    7    12   6    342  1    0    9    0    0     |  396  p     = sci.med
0    1    0    1    0    1    4    0    3    0    1    0    8    4    0    2    369  0    1    1    0     |  396  q     = sci.crypt
10   0    4    10   1    5    6    2    2    6    2    0    2    1    86   15   14   152  0    1    0     |  319  r     = alt.atheism
4    0    0    0    0    9    1    1    8    1    12   0    3    6    0    2    0    0    341  2    0     |  390  s     = misc.forsale
8    5    0    0    0    1    6    0    8    5    50   0    40   2    1    0    9    0    3    256  0     |  394  t     = comp.os.ms-windows.misc
0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0     |  0    u     = unknown
```

5 问题解决

5.1 使用 mahout0.7+版本对 20Newsgroup 数据建立训练集时出错

使用如下命令对 20Newsgroupt 数据建立训练集时：

```
mahout org.apache.mahout.classifier.bayes.PrepareTwentyNewsgroups \
-p /app/mahout-0.9/data/20news-bydate-train \
-o /app/mahout-0.9/data/bayes-train-input \
-a org.apache.mahout.vectorizer.DefaultAnalyzer\
-c UTF-8
```

出现如下错误，原因在于从 0.7 版本开始，Mahout 移除了命令行调用的 prepare20newsgroups、trainclassifier 和 testclassifier API，只能通过 shell 脚本执行 $MAHOUT_HOME/examples/bin/classify-20newsgroups.sh 进行

```
14/12/7 21:31:35 WARN driver.MahoutDriver: Unable to add class:
    org.apache.mahout.classifier.bayes.PrepareTwentyNewsgroups
```

14/12/7 21:31:35 WARN driver.MahoutDriver: No org.apache.mahout.classifier.bayes.PrepareTwentyNewsgroups.props found on classpath, will use command-line arguments only

Unknown program 'org.apache.mahout.classifier.bayes.PrepareTwentyNewsgroups' chosen.

Valid program names are:

arff.vector: : Generate Vectors from an ARFF file or directory

baumwelch: : Baum-Welch algorithm for unsupervised HMM training

.....