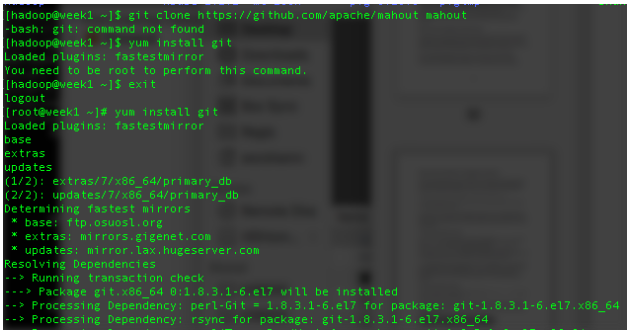
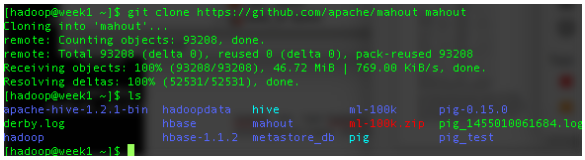
MSDS610 Week 7 Mahout Assignment - Nathan Worsham

Playing with Mahout's Spark Shell

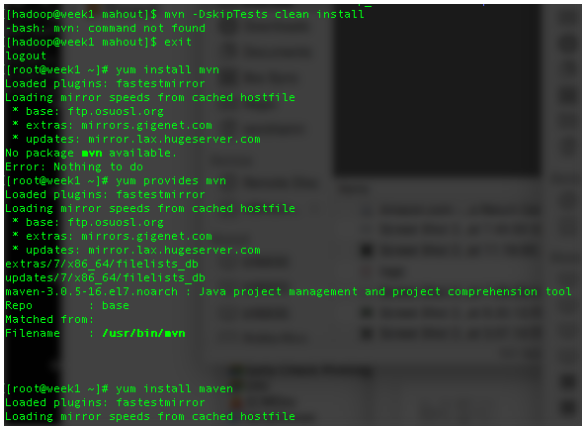
I had trouble navigating the Mahout website for at first I could not find the "Quick Start" but in the "Playing with Mahout's Spark Shell" section it gave instructions for installing both Spark and Mahout. The installation in this section for Mahout is different than previous weeks as it says to clone a github repo. Later I did find the quick start and the normal place to download Mahout, but the installation from git isn't that much different. This meant first using `yum install git` to get the package to use git.



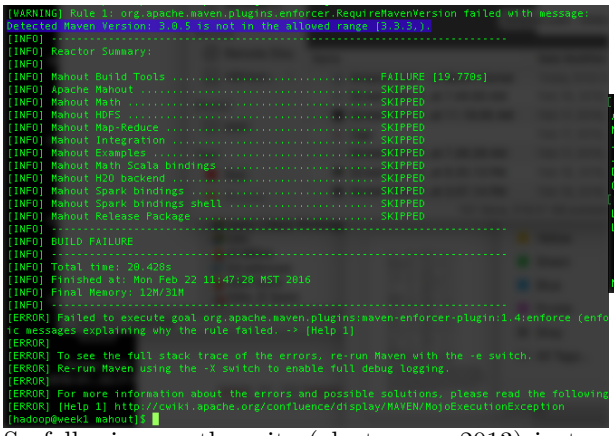
The instructions say to create a directory for Mahout, change to that directory, then clone from git. But the git command it gives creates a folder anyway, so I elected to just run their command from the home directory of my hadoop user.

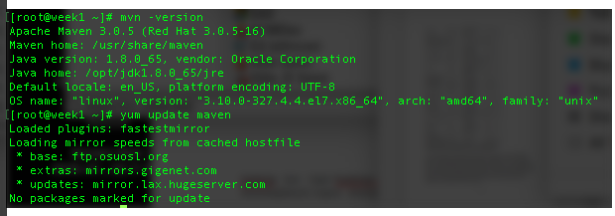


Now that I had the files from git I was ready to build Mahout with the `mvn` command which to no surprise was not installed. So again I switched to root and installed it–I ended up needing to use `yum provides mvn` to find that it was part of the "maven" package.

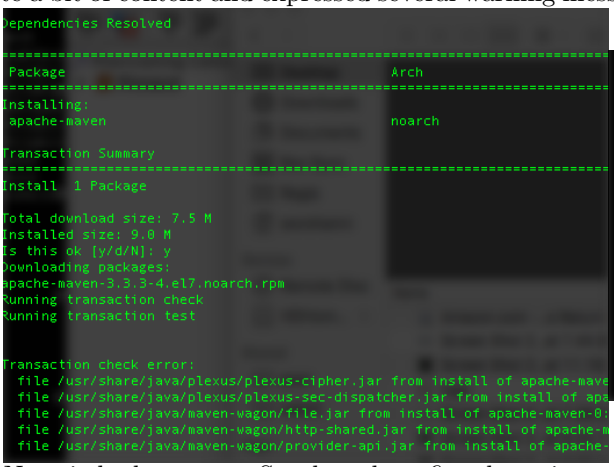


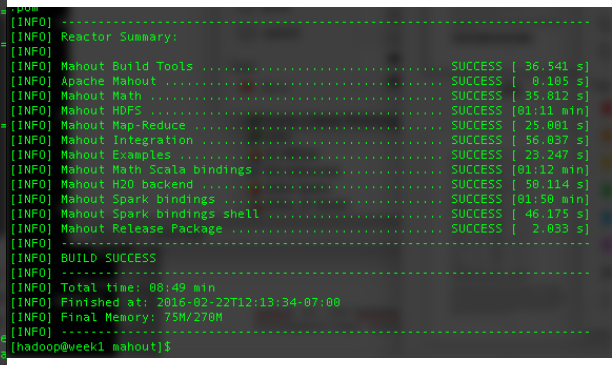
Trying to build using the maven package, I received a build error. After looking at a stackoverflow.com thread (2015), I realized that the maven I installed was an older version and that Mahout has a requirement for a newer version. I went ahead and tried to update through yum but there were no updates for maven through the regular yum repos.



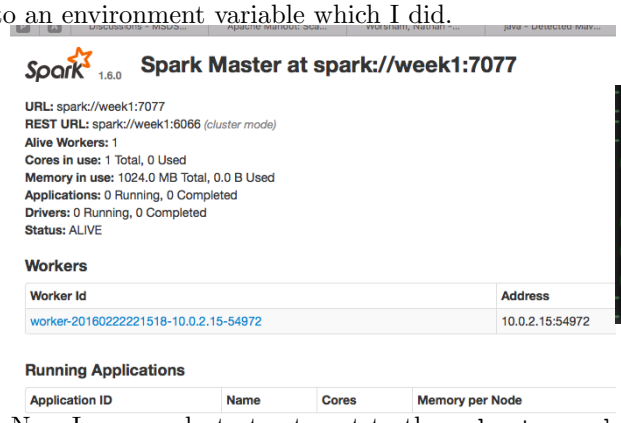


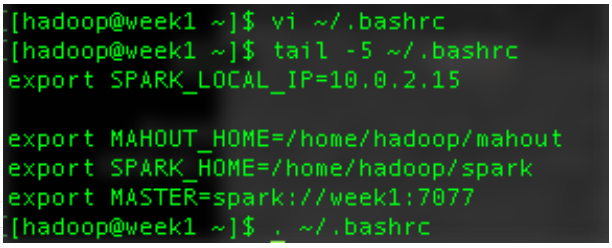
So following another site (gluster.org, 2013) instructions for installing the latest maven through yum, I downloaded a new repo and then used `yum install apache-maven`. But again I received some errors, this time because it was conflicting with some the dependencies that were installed for the previous maven. I went ahead and ran `yum autoremove` to get rid of the old dependencies and then I was able to install apache-maven. The article then said to update symlinks for mvn but did not say how or where to do it. I went ahead and just created a single symbolic link to `/usr/local/bin/mvn` since I knew that was in the PATH. Now I was able to run the mvn command, a command that took a long time to complete, downloaded quite a bit of content and expressed several warning messages, but ultimately did say the build was a success.





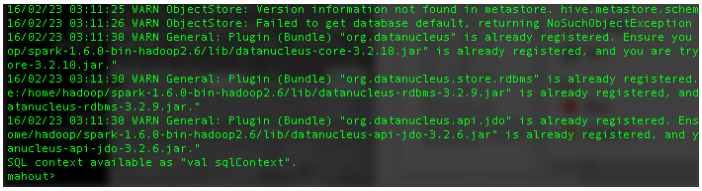
Next it had me start Spark and confirm by going to a local webpage–which is something I did not see last week. After trying for a bit to get to port 8080, I realized the local firewall was in the way and I elected to simply turn it off with `service firewalld stop`. After that I was to copy the URL in order to place it into an environment variable which I did.



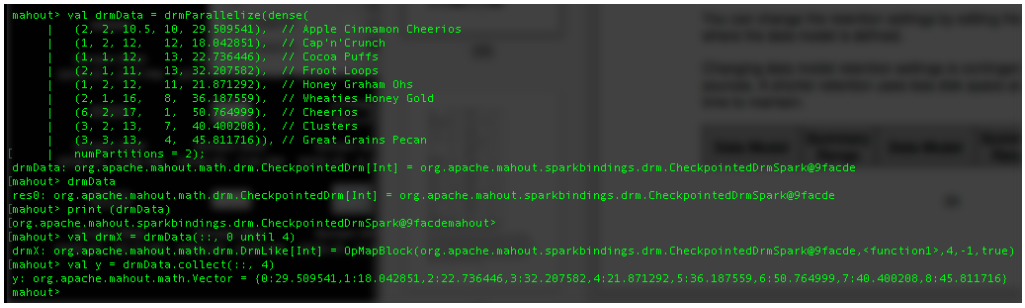


Now I was ready to try to get to the `mahout spark-shell`. The first time I tried it I received an error about trying to connect to localhost on port 9000, which made me realize I had better start up HDFS and yarn (as the instructions did not indicate but was maybe implied). After starting those services and trying

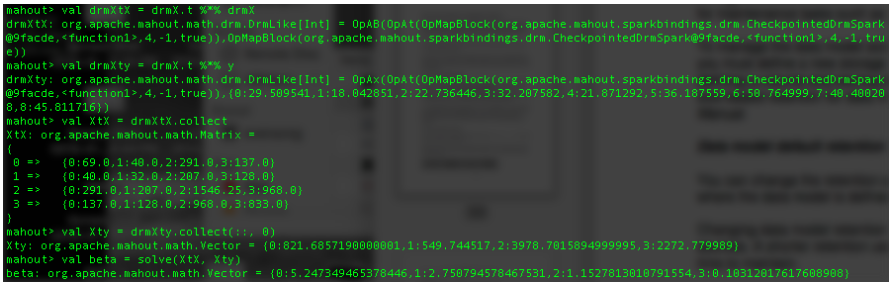
again I received a much better result because I got back the `mahout>` prompt. However I did receive a couple of warnings about plugins already being registered and failing to get a database default.



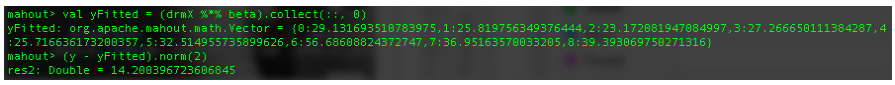
The next step in the exercise was to setup a matrix that represented ingredients (protein, fat, carbohydrate and sugars in milligrams) of various cereals, along with a column of combined user ratings of the cereals. The purpose of the exercise is to fit a linear model which infers the customer rating from the ingredients using a linear regression algorithm. I created the matrix and pulled out the X and Y values into variables.



Using Ordinary Least Squares, I then created variables containing matrix multiplication so that the two variables could then be multiplied to find "beta". What is interesting is after making the variable the equation for each value, I then have to run a "collect" command in order to fetch the values into memory so that the solve function can be run on them.



Now I was to check how well the model fits by first multiplying the ingredient features by the estimated beta (yFitted), and then by looking at the difference between actual y and fitted y using something called L2-norm. I receive an answer of about 14.2. Looking for information on L2-norm brought me to a blog post (Rorasa, 2015) that essentially stated that it is also known as Euclidean distance.



The next part in the exercise was to refactor everything that was just done into two functions and then also add a constant bias term to the model. The result was a goodness of fit of nearly half of what it previously was at approximately 7.6.

```
mahout> def ols(drmX: DrmLike[Int], y: Vector) =
     |   solve(drmX.t %*% drmX, drmX.t %*% y)(::, 0)
ols: (drmX: org.apache.mahout.math.drm.DrmLike[Int], y: org.apache.mahout.math.Vector)org.apache.mahout.math.Vector
mahout> def goodnessOfFit(drmX: DrmLike[Int], beta: Vector, y: Vector) = {
     |   val fittedY = (drmX %*% beta).collect(::, 0)
     |   (y - fittedY).norm(2)
     | }
goodnessOfFit: (drmX: org.apache.mahout.math.drm.DrmLike[Int], beta: org.apache.mahout.math.Vector, y: org.apache.mahout.math.V
ector)Double
mahout> val drmXwithBiasColumn = drmX cbind 1
drmXwithBiasColumn: org.apache.mahout.math.drm.DrmLike[Int] = OpCbindScalar(OpMapBlock(org.apache.mahout.sparkbindings.drm.Chec
kpointedDrmSpark@9facde,<function1>,4,-1,true),1.0,false)
mahout> val betaWithBiasTerm = ols(drmXwithBiasColumn, y)
betaWithBiasTerm: org.apache.mahout.math.Vector = {0:-1.3362653803272289,1:-13.15770132067483,2:-4.152654199020216,3:-5.6799080
94232256,4:163.1793268784127}
mahout> goodnessOfFit(drmXwithBiasColumn, betaWithBiasTerm, y)
res3: Double = 7.623280714561956
```

Last the exercise was to implement a caching functionality to cache the value of the bias column into memory. This was for improved performance as it resulted in the same outcome.

```
mahout> val cachedDrmX = drmXwithBiasColumn.checkpoint()
cachedDrmX: org.apache.mahout.math.drm.CheckpointedDrm[Int] = org.apache.mahout.sparkbindings.drm.CheckpointedDrmSpark@33af522b
mahout> val betaWithBiasTerm = ols(cachedDrmX, y)
betaWithBiasTerm: org.apache.mahout.math.Vector = {0:-1.3362653803272289,1:-13.15770132067483,2:-4.152654199020216,3:-5.6799080
94232256,4:163.1793268784127}
mahout> val goodness = goodnessOfFit(cachedDrmX, betaWithBiasTerm, y)
goodness: Double = 7.623280714561956
mahout> cachedDrmX.uncache()
res4: cachedDrmX.type = org.apache.mahout.sparkbindings.drm.CheckpointedDrmSpark@33af522b
mahout> goodness
res5: Double = 7.623280714561956
```

Twenty Newsgroups Classification Example

After completing the Spark shell exercise, I was really just interested in poking around the other exercises to see if anything looked interesting. I ended up trying the newsgroups classification exercise, which takes a dataset collection of about 20,000 newsgroup documents and classifies them into 20 different newsgroups. The exercise starts off with just having you run one of the examples contained within the Mahout folder. I went ahead and tried running the `./examples/bin/classify-20newsgroups.sh` script but I ended up receiving an error about the "temp/weights" folder not being publicly accessible. I took that to mean that it was not a world writable folder, so I tried `chmod 777` the rights of the folder and ran the example again but I received the same error.

```
        at org.apache.mahout.driver.MahoutDriver.main(MahoutDriver.java:195)
Caused by: java.util.concurrent.ExecutionException: java.io.IOException: Resource file:/home/hadoop/mahout_git/temp/weights is
not publicly accessable and as such cannot be part of the public cache.
        at java.util.concurrent.FutureTask.report(FutureTask.java:122)
        at java.util.concurrent.FutureTask.get(FutureTask.java:192)
        at org.apache.hadoop.mapred.LocalDistributedCacheManager.setup(LocalDistributedCacheManager.java:145)
        ... 21 more
Caused by: java.io.IOException: Resource file:/home/hadoop/mahout_git/temp/weights is not publicly accessable and as such canno
t be part of the public cache.
        at org.apache.hadoop.yarn.util.FSDownload.copy(FSDownload.java:257)
        at org.apache.hadoop.yarn.util.FSDownload.access$000(FSDownload.java:60)
        at org.apache.hadoop.yarn.util.FSDownload$2.run(FSDownload.java:355)
        at org.apache.hadoop.yarn.util.FSDownload$2.run(FSDownload.java:353)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1556)
        at org.apache.hadoop.yarn.util.FSDownload.call(FSDownload.java:352)
        at org.apache.hadoop.yarn.util.FSDownload.call(FSDownload.java:59)
        at java.util.concurrent.FutureTask.run(FutureTask.java:266)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
        at java.lang.Thread.run(Thread.java:745)
[[hadoop@week1 mahout]$ ls
bin            conf         doap_Mahout.rdf  hdfs         math          mr          README.md    src
buildtools     derby.log    examples         integration  math-scala    NOTICE.txt  spark        target
CHANGELOG      distribution h2o              LICENSE.txt  metastore_db  pom.xml     spark-shell  temp
[[hadoop@week1 mahout]$ cd temp/
[[hadoop@week1 temp]$ ls -l
total 0
drwxrwxr-x. 2 hadoop hadoop 84 Feb 23 14:12 summedObservations
drwxrwxr-x. 2 hadoop hadoop 84 Feb 23 14:12 weights
[[hadoop@week1 temp]$ chmod 777 weights/
[[hadoop@week1 temp]$ cd ../
[[hadoop@week1 mahout]$ ./examples/bin/classify-20newsgroups.sh
Discovered Hadoop v2.
Setting dfs command to /home/hadoop/hadoop/bin/hdfs dfs. dfs rm to /home/hadoop/hadoop/bin/hdfs dfs -rm -r -skipTrash.
```

Even after opening the rights on the folder, when I ran the example again the folder's rights went back to the way they were (775), so the script must have a bug I'm guessing. At this point I decided to walk through the example as they gave step-by-step instructions for how the example script works. I first checked to see if I could find the file the script already downloaded but could not find it as it took awhile to download. I then set the environment variables accordingly and made the working directory.

```
[hadoop@week1 mahout]$ find . -iname "20news-bydate*"
[hadoop@week1 mahout]$ vi ~/.bashrc
[hadoop@week1 mahout]$ tail ~/.bashrc

export SPARK_DIST_CLASSPATH=$HADOOP_HOME
export SPARK_LOCAL_IP=10.0.2.15

export MAHOUT_HOME=/home/hadoop/mahout
export SPARK_HOME=/home/hadoop/spark
export MASTER=spark://week1:7077
export MAHOUT_LOCAL=true
export WORK_DIR=/tmp/mahout-work-${USER}
export PATH=$PATH:$MAHOUT_HOME/bin
[hadoop@week1 mahout]$ mkdir -p ${WORK_DIR}
mkdir: missing operand
Try 'mkdir --help' for more information.
[hadoop@week1 mahout]$ . ~/.bashrc
[hadoop@week1 mahout]$ mkdir -p ${WORK_DIR}
```

Next I re-downloaded the data set, unpacked it, and then tried to make another required directory that apparently already existed.

```
[hadoop@week1 mahout]$ curl http://people.csail.mit.edu/jrennie/20Newsgroups/20news-bydate.tar.gz -o ${WORK_DIR}/20news-bydate.
tar.gz
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 13.7M  100 13.7M    0     0  67563      0  0:03:34  0:03:34 --:--:--  129k
[hadoop@week1 mahout]$ mkdir -p ${WORK_DIR}/20news-bydate
[hadoop@week1 mahout]$ cd ${WORK_DIR}/20news-bydate && tar xzf ../20news-bydate.tar.gz && cd .. && cd ..
[hadoop@week1 tmp]$ mkdir ${WORK_DIR}/20news-all
mkdir: cannot create directory '/tmp/mahout-work-hadoop/20news-all': File exists
```

The instructions had an optional section about trying to "put" a folder and contents into the HDFS, but it seems the command is deprecated and it complained of "No such file or directory". Given that it was optional, I decided to move on without it.

```
[hadoop@week1 tmp]$ cp -R ${WORK_DIR}/20news-bydate/*/* ${WORK_DIR}/20news-all
[hadoop@week1 tmp]$ hadoop dfs -put ${WORK_DIR}/20news-all ${WORK_DIR}/20news-all
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

put: `/tmp/mahout-work-hadoop/20news-all': No such file or directory
```

I then successfully ran the commands to convert the dataset into a sequence file and then a sequence file with term frequencies for each document.

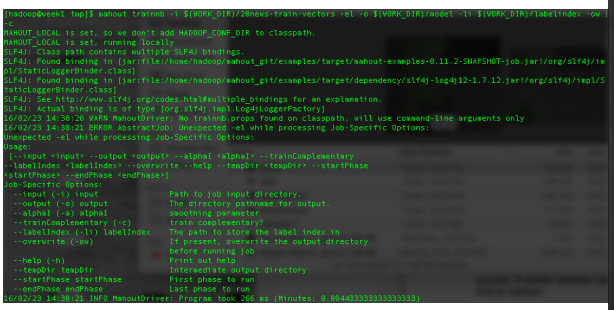
```
16/02/23 14:32:17 INFO Job:  map 100% reduce 0%
16/02/23 14:32:17 INFO Job: Job job_local1617541775_0001 completed successfully
16/02/23 14:32:17 INFO Job: Counters: 18
        File System Counters
                FILE: Number of bytes read=95444499
                FILE: Number of bytes written=79622251
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=18846
                Map output records=18846
                Input split bytes=1484488
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=234
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=49455104
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=19352079
16/02/23 14:32:17 INFO MahoutDriver: Program took 67801 ms (Minutes: 1.13135)
```

```
16/02/23 14:34:07 INFO Job: Job job_local523752949_0089 completed successfully
16/02/23 14:34:07 INFO Job: Counters: 33
        File System Counters
                FILE: Number of bytes read=2137567366
                FILE: Number of bytes written=2050102067
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=18846
                Map output records=18846
                Map output bytes=28362585
                Map output materialized bytes=28437750
                Input split bytes=139
                Combine input records=0
                Combine output records=0
                Reduce input groups=18846
                Reduce shuffle bytes=28437750
                Reduce input records=18846
                Reduce output records=18846
                Spilled Records=37692
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=127
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=198451200
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=28913427
        File Output Format Counters
                Bytes Written=28913427
16/02/23 14:34:07 INFO HadoopUtil: Deleting /tmp/mahout-work-hadoop/20news-vectors/partial-vectors-0
16/02/23 14:34:07 INFO MahoutDriver: Program took 62070 ms (Minutes: 1.0345)
[hadoop@week1 tmp]$
```

Split the dataset into training and testing sets.

```
[hadoop@week1 tmp]$ 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
MAHOUT_LOCAL is set, so we don't add HADOOP_CONF_DIR to classpath.
MAHOUT_LOCAL is set, running locally
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hadoop/mahout_git/examples/target/mahout-examples-0.11.2-SNAPSHOT-job.jar!/org/slf4j/im
pl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hadoop/mahout_git/examples/target/dependency/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/S
taticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
16/02/23 14:35:54 VARN MahoutDriver: No split.props found on classpath, will use command-line arguments only
16/02/23 14:35:54 INFO AbstractJob: Command line arguments: {--endPhase=[2147483647], --input=[/tmp/mahout-work-hadoop/20news-v
ectors/tfidf-vectors], --method=[sequential], --overwrite=null, --randomSelectionPct=[40], --sequenceFiles=null, --startPhase=[
0], --tempDir=[temp], --testOutput=[/tmp/mahout-work-hadoop/20news-test-vectors], --trainingOutput=[/tmp/mahout-work-hadoop/20n
ews-train-vectors]}
16/02/23 14:35:54 INFO HadoopUtil: Deleting /tmp/mahout-work-hadoop/20news-train-vectors
16/02/23 14:35:54 INFO HadoopUtil: Deleting /tmp/mahout-work-hadoop/20news-test-vectors
16/02/23 14:35:57 INFO SplitInput: part-r-00000 has 162419 lines
16/02/23 14:35:57 INFO SplitInput: part-r-00000 test split size is 64968 based on random selection percentage 40
16/02/23 14:35:57 INFO ZlibFactory: Successfully loaded & initialized native-zlib library
16/02/23 14:35:57 INFO CodecPool: Got brand-new compressor [.deflate]
16/02/23 14:35:57 INFO CodecPool: Got brand-new compressor [.deflate]
16/02/23 14:36:04 INFO SplitInput: file: part-r-00000, input: 162419 train: 11322, test: 7524 starting at 0
16/02/23 14:36:04 INFO MahoutDriver: Program took 18078 ms (Minutes: 0.16796666666666665)
[hadoop@week1 tmp]$
```

When I tried to run the command to train the classifier I received an error complaining about the `-el` option, so not really knowing what to do, I decided to simply remove the option which luckily still worked.



```
16/02/23 14:38:45 INFO Job: Counters: 33
        File System Counters
                FILE: Number of bytes read=422901078
                FILE: Number of bytes written=369750038
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=20
                Map output records=1
                Map output bytes=167
                Map output materialized bytes=185
                Input split bytes=111
                Combine input records=1
                Combine output records=1
                Reduce input groups=1
                Reduce shuffle bytes=185
                Reduce input records=1
                Reduce output records=1
                Spilled Records=2
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=44
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=338029312
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=2784969
        File Output Format Counters
                Bytes Written=277
16/02/23 14:38:46 INFO MahoutDriver: Program took 13241 ms (Minutes: 0.22068333333333334)
```

Finally I was able to test the classifier and received the output that the example script did not supply.

```
16/02/23 14:40:36 INFO TestNaiveBayesDriver: Complementary Results:
=======================================================
Summary
-------------------------------------------------------
Correctly Classified Instances          :       6686       88.8623%
Incorrectly Classified Instances        :        838       11.1377%
Total Classified Instances              :       7524

=======================================================
Confusion Matrix
-------------------------------------------------------
a       b       c       d       e       f       g       h       i       j       k       l       m       n       o       p       q       r       s       t       <--Classified as
294     1       0       1       1       0       0       0       0       0       0       1       0       2       2       5       0       0       1       11      |  319       a     = alt.atheism
2       320     2       12      6       14      7       4       0       1       2       10      2       1       6       1       4       1       3       1       |  399       b     = comp.graphics
4       17      235     52      5       13      11      3       3       3       3       7       10      3       7       3       1       2       4       3       |  389       c     = comp.os.ms-windows.misc
1       9       4       331     8       5       17      3       1       2       4       2       12      2       6       5       4       1       1       3       |  421       d     = comp.sys.ibm.pc.hardware
0       1       3       3       335     3       3       3       1       2       1       4       2       2       1       1       0       1       2       2       |  370       e     = comp.sys.mac.hardware
1       11      1       5       3       343     1       3       3       3       0       4       1       2       3       0       1       2       0       1       |  388       f     = comp.windows.x
0       3       1       23      7       1       303     13      2       3       3       3       14      3       1       0       4       1       5       3       |  393       g     = misc.forsale
0       1       1       3       1       0       0       347     4       3       3       0       1       3       0       0       2       1       3       0       |  373       h     = rec.autos
0       1       0       0       0       0       1       4       397     0       1       0       0       0       0       0       2       0       0       0       |  406       i     = rec.motorcycles
0       0       0       0       1       0       0       1       1       390     1       1       3       2       0       1       0       0       0       1       |  402       j     = rec.sport.baseball
0       1       0       0       0       0       1       1       0       4       372     0       0       0       0       0       0       0       0       0       |  379       k     = rec.sport.hockey
1       0       0       0       0       0       0       1       0       0       0       397     2       0       0       0       0       0       3       1       |  405       l     = sci.crypt
0       2       0       9       6       1       10      5       5       3       1       3       329     1       5       2       1       1       2       3       |  389       m     = sci.electronics
2       1       2       0       0       1       1       2       0       1       1       1       4       367     3       2       3       0       3       2       |  396       n     = sci.med
1       2       0       0       0       0       0       1       0       0       0       1       3       2       399     0       3       3       3       3       |  421       o     = sci.space
3       0       1       3       0       0       0       2       0       2       0       2       0       1       1       371     0       1       2       3       |  392       p     = soc.religion.christian
0       0       0       0       0       0       0       0       0       1       0       4       0       2       1       0       358     2       10      0       |  378       q     = talk.politics.guns
3       0       0       0       0       0       0       0       0       1       0       0       0       0       0       3       0       358     0       0       |  365       r     = talk.politics.mideast
2       0       0       0       0       1       0       1       0       1       3       1       0       0       0       0       12      6       261     2       |  290       s     = talk.politics.misc
33      2       0       0       0       0       0       0       0       0       0       0       0       1       4       23      2       0       5       179     |  249       t     = talk.religion.misc

=======================================================
Statistics
-------------------------------------------------------
Kappa                                       0.8533
Accuracy                                   88.8623%
Reliability                                84.4187%
Reliability (standard deviation)            0.2176
Weighted precision                          0.8891
Weighted recall                             0.8886
Weighted F1 score                           0.8863

16/02/23 14:40:36 INFO MahoutDriver: Program took 10642 ms (Minutes: 0.17736666666666667)
```

References

stackoverflow.com, 2015. Retrieved from http://stackoverflow.com/questions/31968515/detected-maven-version-3-0-5-is-not-in-the-allowed-range-3-2
rorasa.wordpress.com, 2015. Retrieved from https://rorasa.wordpress.com/2012/05/13/l0-norm-l1-norm-l2-norm-l-infinity-norm/
mahout.apache.org, 2016. Retrieved from http://mahout.apache.org/users/sparkbindings/play-with-shell.html and http://mahout.apache.org/users/classification/twenty-newsgroups.html