

## CS5560- KDM - Lab#3 Assignment

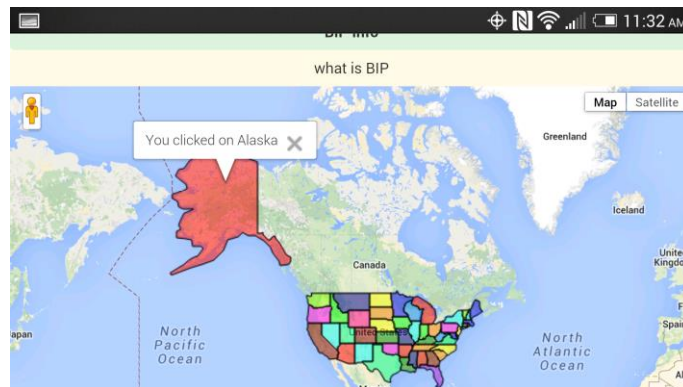
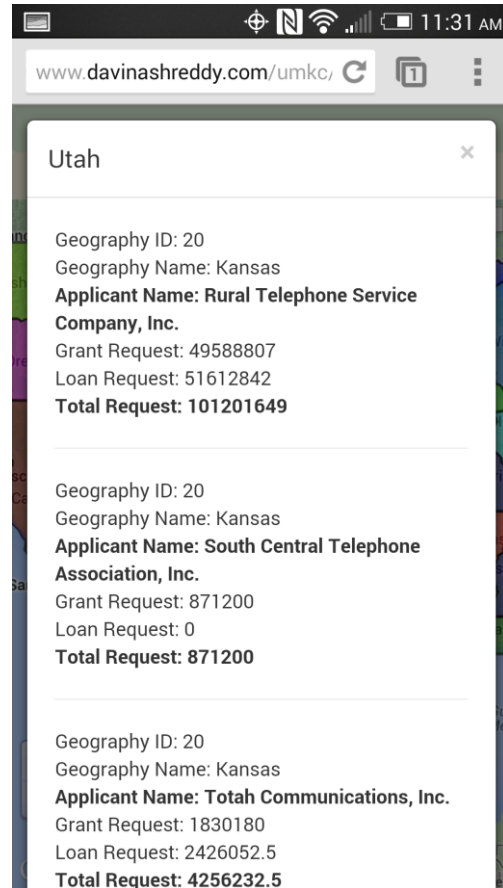
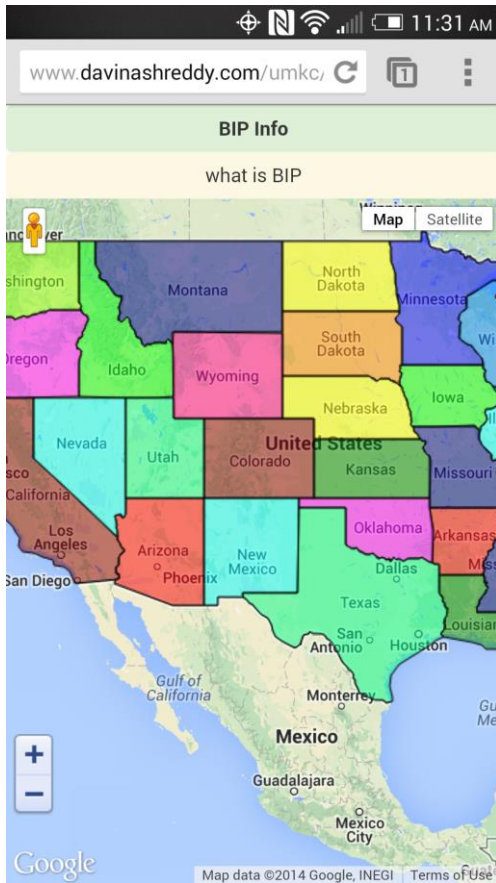
This document consists of following information.

1. Make a **Mashup application** including various services (e.g., Google Map, Google Chart, Google Search, Yahoo, Amazon, Twitter, Facebook) Web Services (e.g., Google Map Services, Weather Services) using either (1) Mobile Web Technology with HMLT5 Local DB (Refer to Tutorial 3).
  2. **Cloudera/MapReduce**: Download the Cloudera Image, implement the WordCount MapReduce and run it. (a bonus point for implementing a new MapReduce algorithm) The code and guidelines will be available in Tutorials/Tutorial 5.
  3. **Cloudera/Mahout**: Configure your Cloudera with Mahout. Run Naive Bayes classifier with the input data (a bonus point for using your own data).
- GitHub Account Screenshots(username: davinashreddy)
  - ScrumDo Account Screenshots(link: <https://www.scrumdo.com/organization/umkc32/dashboard>)

---

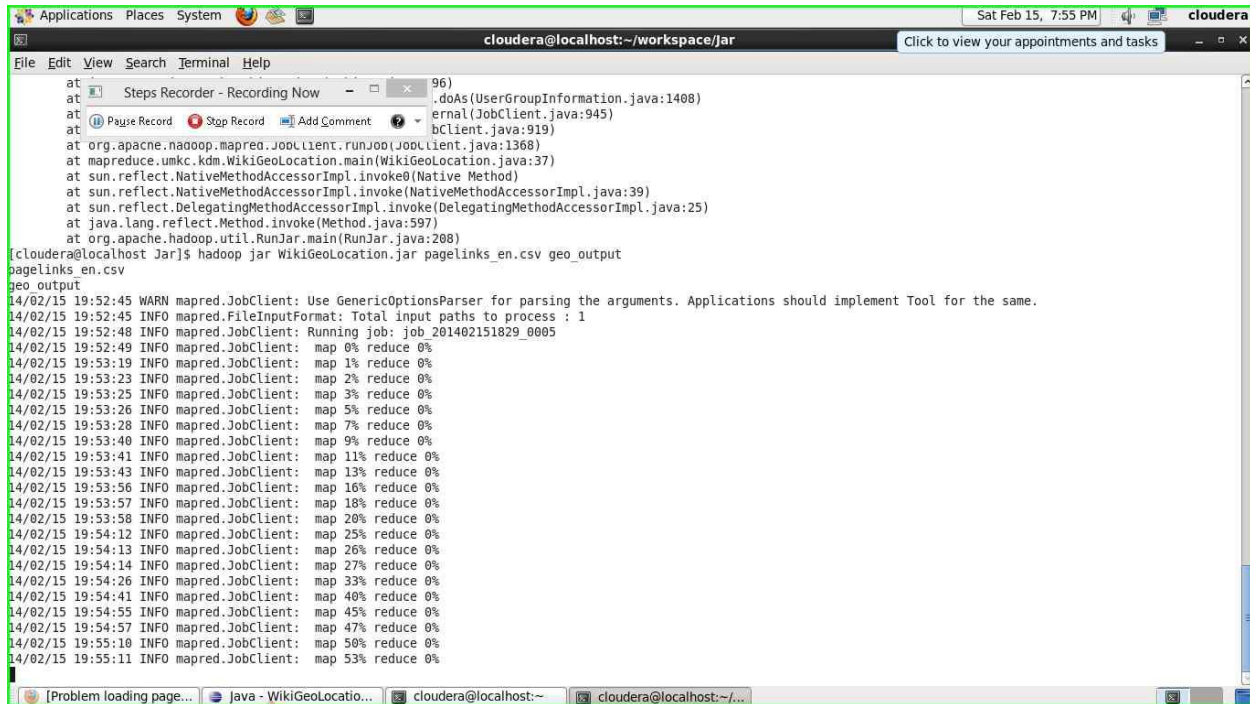
### 1. Mashup Application:

- My application includes Google map service on which States in united states have an overlay with different color and name.
- When you click on a specific state then a service call is made to the following url and the BIP information of the state is retrieved.
  - BIP's goal is to provide financial boost to the nation during the economic crisis [1].
  - <http://www.broadbandmap.gov/broadbandmap/bip/states/kansas?format=json> - Sample Service call, to Kansas statistics.
  - <http://www.responsinator.com/?url=http%3A%2F%2Fwww.davinashreddy.com%2Fumkc%2Fkdm-lab3%2F>
  - <http://www.davinashreddy.com/umkc/kdm-lab3/>
- Following are some screenshots taken in HTC one device in chrome browser.

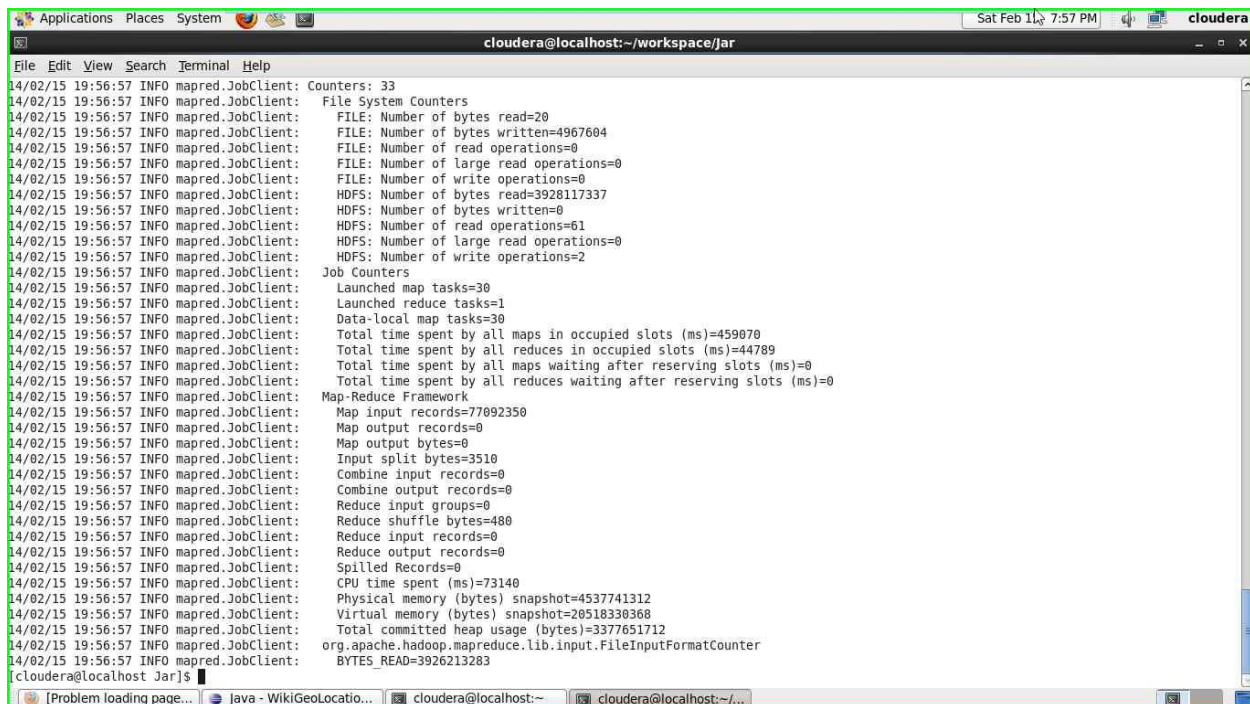


## 2. Cloudera/Mapreduce

- My MapReduce program runs on DB-Pedia GeoLocation data where the goal is to extract geo-location of specific URL [2].
- Source code is available in GitHub
- Screenshots of my work as follows.



```
at org.apache.hadoop.mapred.JobClient.runJob(JobClient.java:1368)
at mapreduce.umkc.kdm.WikiGeoLocation.main(WikiGeoLocation.java:37)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:39)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:25)
at java.lang.reflect.Method.invoke(Method.java:597)
at org.apache.hadoop.util.RunJar.main(RunJar.java:208)
[cloudera@localhost Jar]$ hadoop jar WikiGeoLocation.jar pagelinks_en.csv geo_output
pagelinks_en.csv
geo_output
14/02/15 19:52:45 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.
14/02/15 19:52:45 INFO mapred.FileInputFormat: Total input paths to process : 1
14/02/15 19:52:48 INFO mapred.JobClient: Running job: job_201402151829_0005
14/02/15 19:52:49 INFO mapred.JobClient: map 0% reduce 0%
14/02/15 19:53:19 INFO mapred.JobClient: map 1% reduce 0%
14/02/15 19:53:23 INFO mapred.JobClient: map 2% reduce 0%
14/02/15 19:53:25 INFO mapred.JobClient: map 3% reduce 0%
14/02/15 19:53:26 INFO mapred.JobClient: map 5% reduce 0%
14/02/15 19:53:28 INFO mapred.JobClient: map 7% reduce 0%
14/02/15 19:53:40 INFO mapred.JobClient: map 9% reduce 0%
14/02/15 19:53:41 INFO mapred.JobClient: map 11% reduce 0%
14/02/15 19:53:43 INFO mapred.JobClient: map 13% reduce 0%
14/02/15 19:53:56 INFO mapred.JobClient: map 16% reduce 0%
14/02/15 19:53:57 INFO mapred.JobClient: map 18% reduce 0%
14/02/15 19:53:58 INFO mapred.JobClient: map 20% reduce 0%
14/02/15 19:54:12 INFO mapred.JobClient: map 25% reduce 0%
14/02/15 19:54:13 INFO mapred.JobClient: map 26% reduce 0%
14/02/15 19:54:14 INFO mapred.JobClient: map 27% reduce 0%
14/02/15 19:54:26 INFO mapred.JobClient: map 33% reduce 0%
14/02/15 19:54:41 INFO mapred.JobClient: map 40% reduce 0%
14/02/15 19:54:55 INFO mapred.JobClient: map 45% reduce 0%
14/02/15 19:54:57 INFO mapred.JobClient: map 47% reduce 0%
14/02/15 19:55:10 INFO mapred.JobClient: map 50% reduce 0%
14/02/15 19:55:11 INFO mapred.JobClient: map 53% reduce 0%
```



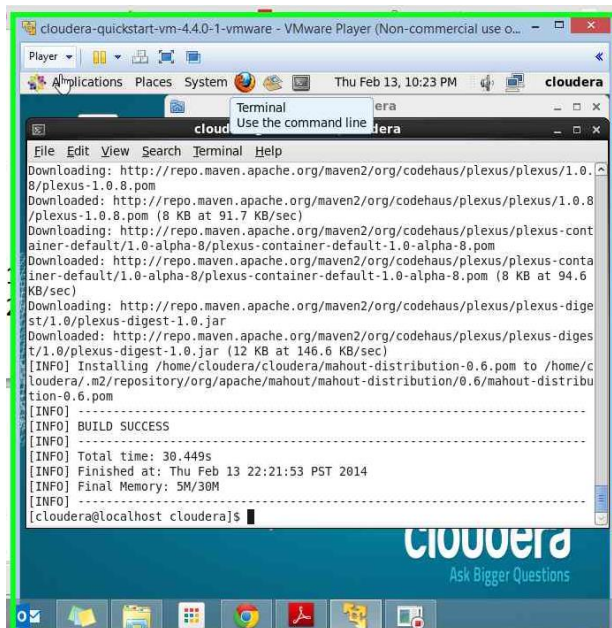
```
14/02/15 19:56:57 INFO mapred.JobClient: Counters: 33
14/02/15 19:56:57 INFO mapred.JobClient: File System Counters
14/02/15 19:56:57 INFO mapred.JobClient: FILE: Number of bytes read=20
14/02/15 19:56:57 INFO mapred.JobClient: FILE: Number of bytes written=4967604
14/02/15 19:56:57 INFO mapred.JobClient: FILE: Number of read operations=0
14/02/15 19:56:57 INFO mapred.JobClient: FILE: Number of large read operations=0
14/02/15 19:56:57 INFO mapred.JobClient: FILE: Number of write operations=0
14/02/15 19:56:57 INFO mapred.JobClient: HDFS: Number of bytes read=3928117337
14/02/15 19:56:57 INFO mapred.JobClient: HDFS: Number of bytes written=0
14/02/15 19:56:57 INFO mapred.JobClient: HDFS: Number of read operations=61
14/02/15 19:56:57 INFO mapred.JobClient: HDFS: Number of large read operations=0
14/02/15 19:56:57 INFO mapred.JobClient: HDFS: Number of write operations=2
14/02/15 19:56:57 INFO mapred.JobClient: Job Counters
14/02/15 19:56:57 INFO mapred.JobClient: Launched map tasks=30
14/02/15 19:56:57 INFO mapred.JobClient: Launched reduce tasks=1
14/02/15 19:56:57 INFO mapred.JobClient: Data-local map tasks=30
14/02/15 19:56:57 INFO mapred.JobClient: Total time spent by all maps in occupied slots (ms)=459070
14/02/15 19:56:57 INFO mapred.JobClient: Total time spent by all reduces in occupied slots (ms)=44789
14/02/15 19:56:57 INFO mapred.JobClient: Total time spent by all maps waiting after reserving slots (ms)=0
14/02/15 19:56:57 INFO mapred.JobClient: Total time spent by all reduces waiting after reserving slots (ms)=0
14/02/15 19:56:57 INFO mapred.JobClient: Map-Reduce Framework
14/02/15 19:56:57 INFO mapred.JobClient: Map input records=77092350
14/02/15 19:56:57 INFO mapred.JobClient: Map output records=0
14/02/15 19:56:57 INFO mapred.JobClient: Map output bytes=0
14/02/15 19:56:57 INFO mapred.JobClient: Input split bytes=3510
14/02/15 19:56:57 INFO mapred.JobClient: Combine input records=0
14/02/15 19:56:57 INFO mapred.JobClient: Combine output records=0
14/02/15 19:56:57 INFO mapred.JobClient: Reduce input groups=0
14/02/15 19:56:57 INFO mapred.JobClient: Reduce shuffle bytes=480
14/02/15 19:56:57 INFO mapred.JobClient: Reduce input records=0
14/02/15 19:56:57 INFO mapred.JobClient: Reduce output records=0
14/02/15 19:56:57 INFO mapred.JobClient: Spilled Records=0
14/02/15 19:56:57 INFO mapred.JobClient: CPU time spent (ms)=73140
14/02/15 19:56:57 INFO mapred.JobClient: Physical memory (bytes) snapshot=4537741312
14/02/15 19:56:57 INFO mapred.JobClient: Virtual memory (bytes) snapshot=20518330368
14/02/15 19:56:57 INFO mapred.JobClient: Total committed heap usage (bytes)=3377651712
14/02/15 19:56:57 INFO mapred.JobClient: org.apache.hadoop.mapreduce.lib.input.FileInputFormatCounter
14/02/15 19:56:57 INFO mapred.JobClient: BYTES_READ=3926213283
[cloudera@localhost Jar]$
```



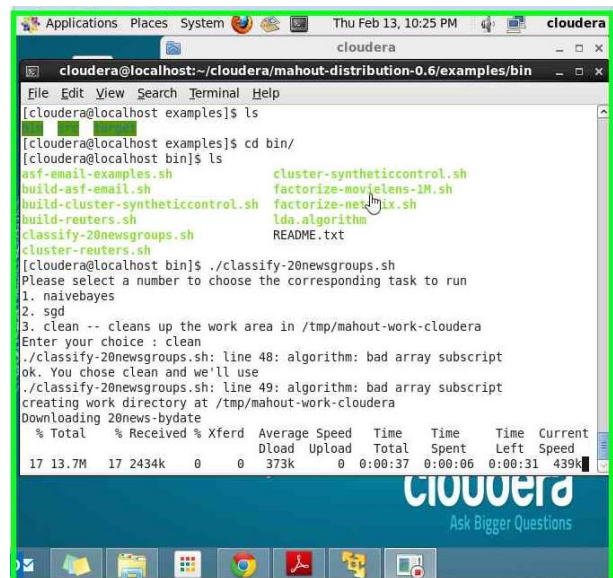
### 3. Cloudera/Mahout:

To run mahout in Cloudera CentOS edition, Maven has to be installed. Steps to install Maven follows:

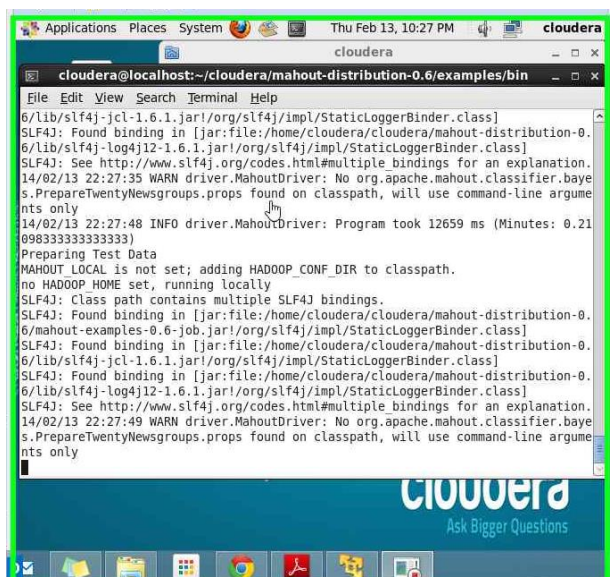
- Download the maven and mahout repos along with .pom file which installs all the dependencies required.
- Using mahout .pom file install mahout using maven by running
  - `mvn -f filename.pom -DskipTests install`
- Set the permissions for Mahout and run Naviebayes.
- Below are the screenshots of my work.



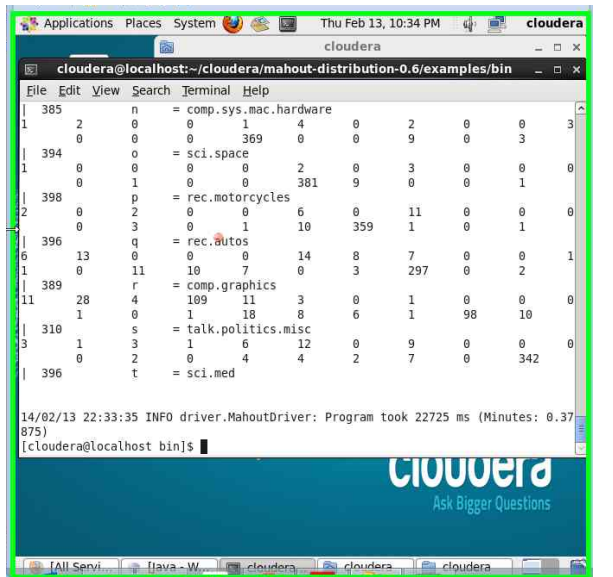
```
cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use o...
Thu Feb 13, 10:23 PM cloudera
Terminal
Use the command line
cloudera@localhost:~/cloudera/mahout-distribution-0.6/examples/bin$ mvn -f /home/cloudera/m2/repository/org/apache/mahout/mahout-distribution-0.6.pom -DskipTests install
Downloading: http://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-1.0.8/plexus-1.0.8.pom
Downloaded: http://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-1.0.8/plexus-1.0.8.pom (8 KB at 91.7 KB/sec)
Downloading: http://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-container-default/1.0-alpha-8/plexus-container-default-1.0-alpha-8.pom
Downloaded: http://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-container-default/1.0-alpha-8/plexus-container-default-1.0-alpha-8.pom (8 KB at 94.6 KB/sec)
Downloading: http://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-digest-1.0/plexus-digest-1.0.jar
Downloaded: http://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-digest-1.0/plexus-digest-1.0.jar (12 KB at 146.6 KB/sec)
[INFO] Installing /home/cloudera/cloudera/mahout-distribution-0.6.pom to /home/cloudera/.m2/repository/org/apache/mahout/mahout-distribution/0.6/mahout-distribution-0.6.pom
[INFO] BUILD SUCCESS
[INFO] Total time: 30.449s
[INFO] Finished at: Thu Feb 13 22:21:53 PST 2014
[INFO] Final Memory: 5M/30M
[INFO] -----
[cloudera@localhost cloudera]$
```



```
cloudera@localhost:~/cloudera/mahout-distribution-0.6/examples/bin$ ls
[cloudera@localhost examples]$ cd bin/
[cloudera@localhost bin]$ ls
asf-email-examples.sh      cluster-syntheticcontrol.sh
build-asf-email.sh         factorize-movie-lens-1M.sh
build-cluster-syntheticcontrol.sh  factorize-netflix.sh
build-reuters.sh           lda.algorithm
classify-20newsgroups.sh   README.txt
cluster-reuters.sh
[cloudera@localhost bin]$ ./classify-20newsgroups.sh
Please select a number to choose the corresponding task to run
1. naivebayes
2. sgd
3. clean -- cleans up the work area in /tmp/mahout-work-cloudera
Enter your choice : clean
./classify-20newsgroups.sh: line 48: algorithm: bad array subscript
ok. You chose clean and we'll use
./classify-20newsgroups.sh: line 49: algorithm: bad array subscript
creating work directory at /tmp/mahout-work-cloudera
Downloading 20news-bydate
% Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
           Dload  Upload   Total   Spent    Left   Speed
17 13.7M  17 2434k    0     0  373k      0  0:00:37  0:00:06  0:00:31 439k
```



```
cloudera@localhost:~/cloudera/mahout-distribution-0.6/examples/bin$ ./cluster-syntheticcontrol.sh
6/lib/slf4j-jcl-1.6.1.jar/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/cloudera/cloudera/mahout-distribution-0.6/lib/slf4j-log4j12-1.6.1.jar/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
14/02/13 22:27:35 WARN driver.MahoutDriver: No org.apache.mahout.classifier.bayes.PrepareTwentyNewsgroups.props found on classpath, will use command-line arguments only
14/02/13 22:27:48 INFO driver.MahoutDriver: Program took 12659 ms (Minutes: 0.21098333333333333)
Preparing Test Data
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:/home/cloudera/cloudera/mahout-distribution-0.6/mahout-examples-0.6-job.jar/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/cloudera/cloudera/mahout-distribution-0.6/lib/slf4j-jcl-1.6.1.jar/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/cloudera/cloudera/mahout-distribution-0.6/lib/slf4j-log4j12-1.6.1.jar/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
14/02/13 22:27:49 WARN driver.MahoutDriver: No org.apache.mahout.classifier.bayes.PrepareTwentyNewsgroups.props found on classpath, will use command-line arguments only
```



```
cloudera@localhost:~/cloudera/mahout-distribution-0.6/examples/bin$ ./cluster-syntheticcontrol.sh
1 385 n = comp.sys.mac.hardware
2 0 0 1 4 0 2 0 0 3
0 0 0 369 0 0 9 0 3
394 o = sci.space
1 0 0 0 2 0 3 0 0 0
0 1 0 0 381 9 0 0 1
398 p = rec.motorcycles
2 0 2 0 6 0 11 0 0 0
0 3 0 1 10 359 1 0 1
396 q = rec.autos
6 13 0 0 14 8 7 0 0 1
1 0 11 10 7 0 3 297 0 2
389 r = comp.graphics
11 28 4 109 11 3 0 1 0 0
1 0 1 18 8 6 1 98 10
310 s = talk.politics.misc
3 1 3 1 6 12 0 9 0 0
0 2 0 4 4 2 7 0 342
396 t = sci.med
14/02/13 22:33:35 INFO driver.MahoutDriver: Program took 22725 ms (Minutes: 0.37875)
[cloudera@localhost bin]$
```

## References

1. [http://www.rurdev.usda.gov/utp\\_bip.html](http://www.rurdev.usda.gov/utp_bip.html)
2. Automatic IO Filtering for Optimizing Cloud Analytics Microsoft Technical Report MSR-TR-2012-3 by *Christos Gkantsidis at Microsoft Research, Cambridge, UK*.