**HU Extension Assignment 09 E185 Big Data Analytics**

**Issued on: April 13, 2013 Due by 17:25PM, April 19, 2013**

**An illustrative application of Bayes’ Theorem**

**Suppose we are interested in a rare syntactic construction, parasitic gap, which occur on average once in 100,000 sentences.**

**Peggy the Linguist has developed a complicated pattern matcher that attempts to identify sentences with parasitic gap. It is pretty good, but it is not perfect: if a sentence has a parasitic gap, it will say so with probability 0.95, if it doesn't, it will wrongly say so with probability 0.005.**

**Suppose the test says that a sentence contains a parasitic gap. What is the probability that this is true?**

***Solution***

**According to the problem, we have:**

**p(gap) = 1/100000 – the probability of a gap**

**p(pos | gap) = 0.95 – the probability the system will give the positive result if there is a gap**

**p(pos | ) = 0.005 – the probability the system will give the positive result if there is no gap**

**p(gap | pos) = ? – the probability that a sentence contains a parasitic gap if the system gives the positive result**

**Problem 1. Public officials estimate the incidence of the HIV virus in the general population is about 0.5 percent. A test is introduced which when given to people with HIV correctly identifies the virus 95 percent of the time. The test also gives a false positive 5 percent of the time. That is the test result could be positive even though the person does not have HIV. Of the people who test positive, what percent of them we actually expect to have the virus.**

**Note on problem 1. The purpose of this problem is to illustrate the benefits and sometimes unexpected results produced by the Bayes theorem.**

Applying the Bayes Theorem on the data above, we have the following:

p(sick) = 0.005

p(pos|sick) = 0.95

p(pos|!sick) = 0.05

p(sick|pos) = p(pos|sick) \* p(sick) / p(pos)

p(sick|pos) = p(pos|sick) \* p(sick) / ( p(pos|sick) \* p(sick) + p(pos|!sick) \* (1 - p(sick))

p(sick|pos) = 0.95 \* 0.005 / ((0.95 \* 0.005) + (0.05 \* (1 - 0.005))

p(sick|pos) = 0.00475 / (0.00475 + 0.04975)

**p(sick|pos) = 0.08715 = 8.715%**

**Problem 2. Please reproduce results for Mahout Recommender as described in lecture note on pages 15 to 19. Once you extract recommendations for several users. Go into ratings.csv file and lower ratings for one of the recommended movies. You can use Vi and globally reduce scores for a selected movie to 1, or 3. Examine whether change of score affected the result of the recommender.**

In order to get started, we need to install the Mahout packages in our Virtual Machine. We follow the procedures on the lecture slides.

**[cloudera@centos-e185 ~]$ sudo yum install mahout**

(...)

Installed:

mahout.noarch 0:0.7+12-1.cdh4.2.0.p0.9.el6

Dependency Installed:

hadoop-0.20-mapreduce.x86\_64 0:0.20.2+1341-1.cdh4.2.0.p0.21.el6 hadoop-client.x86\_64 0:2.0.0+922-1.cdh4.2.0.p0.12.el6

Complete!

We also configure the PATH variable for Mahout.

**[cloudera@centos-e185 ~]$ cat /etc/profile.d/java.sh**

## Exporting JAVA\_HOME and an updated PATH to all user on the machine

export JAVA\_HOME=/usr/java/jdk1.6.0\_31

export MAHOUT\_HOME=/usr/lib/mahout

export HADOOP\_HOME=/usr/lib/hadoop

export PATH=$JAVA\_HOME/bin:$PATH:$MAHOUT\_HOME

Now we create a directory for the MovieLens data that we need to download in order to use Mahout against it:

**[cloudera@centos-e185 ~]$ mkdir Assign09**

**[cloudera@centos-e185 ~]$ cd Assign09/**

**[cloudera@centos-e185 Assign09]$ mkdir P2**

**[cloudera@centos-e185 Assign09]$ cd P2**

**[cloudera@centos-e185 P2]$ curl -O http://www.grouplens.org/system/files/ml-1m.zip**

% Total % Received % Xferd Average Speed Time Time Time Current

Dload Upload Total Spent Left Speed

100 5867k 100 5867k 0 0 1448k 0 0:00:04 0:00:04 --:--:-- 1481k

**[cloudera@centos-e185 P2]$ unzip ml-1m.zip**

Archive: ml-1m.zip

creating: ml-1m/

inflating: ml-1m/movies.dat

inflating: ml-1m/ratings.dat

inflating: ml-1m/README

creating: \_\_MACOSX/

creating: \_\_MACOSX/ml-1m/

inflating: \_\_MACOSX/ml-1m/.\_README

inflating: ml-1m/users.dat

**[cloudera@centos-e185 P2]$ cat ml-1m/ratings.dat | wc -l**

1000209

For the next step, we need to convert the file to a CSV format, so that it can be processes by Mahout

**[cloudera@centos-e185 P2]$ head -n 5 ml-1m/ratings.dat**

1::1193::5::978300760

1::661::3::978302109

1::914::3::978301968

1::3408::4::978300275

1::2355::5::978824291

**[cloudera@centos-e185 P2]$ awk -F"::" '{print $1","$2","$3}' ml-1m/ratings.dat > ratings.csv**

**[cloudera@centos-e185 P2]$ head -n 5 ratings.csv**

1,1193,5

1,661,3

1,914,3

1,3408,4

1,2355,5

And we create a User ID file for the first 20 users on our list.

**[cloudera@centos-e185 P2]$ cat user-ids.txt**

1

2

3

(...)

19

20

We upload the files to HDFS and run the Recommender engine against them.

**[cloudera@centos-e185 P2]$ hadoop fs -mkdir Assign09**

**[cloudera@centos-e185 P2]$ hadoop fs -mkdir Assign09/P2**

**[cloudera@centos-e185 P2]$ hadoop fs -put user-ids.txt ratings.csv Assign09/P2**

**[cloudera@centos-e185 P2]$ hadoop fs -ls Assign09/P2**

Found 2 items

-rw-r--r-- 1 cloudera supergroup 11553456 2013-04-16 06:24 Assign09/P2/ratings.csv

-rw-r--r-- 1 cloudera supergroup 51 2013-04-16 06:24 Assign09/P2/user-ids.txt

**[cloudera@centos-e185 P2]$ mahout recommenditembased -Dmapred.reduce.tasks=10 --similarityClassname SIMILARITY\_PEARSON\_CORRELATION --input Assign09/P2/ratings.csv --output Assign09/P2-output --tempDir Assign09/P2-tmp --usersFile Assign09/P2/user-ids.txt**

File Input Format Counters

Bytes Read=2223699

File Output Format Counters

Bytes Written=1867

13/04/16 06:42:54 INFO driver.MahoutDriver: Program took 902548 ms (Minutes: 15.042466666666666)

**[cloudera@centos-e185 P2]$ hadoop fs -ls Assign09/P2-output**

Found 11 items

-rw-r--r-- 1 cloudera supergroup 0 2013-04-16 06:42 Assign09/P2-output/\_SUCCESS

-rw-r--r-- 1 cloudera supergroup 188 2013-04-16 06:42 Assign09/P2-output/part-r-00000

-rw-r--r-- 1 cloudera supergroup 179 2013-04-16 06:42 Assign09/P2-output/part-r-00001

-rw-r--r-- 1 cloudera supergroup 179 2013-04-16 06:42 Assign09/P2-output/part-r-00002

-rw-r--r-- 1 cloudera supergroup 186 2013-04-16 06:42 Assign09/P2-output/part-r-00003

-rw-r--r-- 1 cloudera supergroup 186 2013-04-16 06:42 Assign09/P2-output/part-r-00004

-rw-r--r-- 1 cloudera supergroup 239 2013-04-16 06:42 Assign09/P2-output/part-r-00005

-rw-r--r-- 1 cloudera supergroup 180 2013-04-16 06:42 Assign09/P2-output/part-r-00006

-rw-r--r-- 1 cloudera supergroup 174 2013-04-16 06:42 Assign09/P2-output/part-r-00007

-rw-r--r-- 1 cloudera supergroup 175 2013-04-16 06:42 Assign09/P2-output/part-r-00008

-rw-r--r-- 1 cloudera supergroup 181 2013-04-16 06:42 Assign09/P2-output/part-r-00009

**[cloudera@centos-e185 P2]$ hadoop fs -cat Assign09/P2-output/part-r-0000\***

10 [3952:5.0,3936:5.0,3932:5.0,3929:5.0,3927:5.0,3913:5.0,3912:5.0,3877:5.0,3875:5.0,3873:5.0]

20 [1625:5.0,1653:5.0,1729:5.0,2881:5.0,2447:5.0,1027:5.0,300:5.0,800:5.0,1179:5.0,1909:5.0]

1 [1566:5.0,1036:5.0,1033:5.0,1032:5.0,1031:5.0,1030:5.0,3107:5.0,3114:5.0,1026:5.0,1025:5.0]

11 [1:5.0,3752:5.0,3868:5.0,1902:5.0,2:5.0,11:5.0,1895:5.0,16:5.0,3793:5.0,1885:5.0]

2 [2739:5.0,3811:5.0,3916:5.0,2:5.0,10:5.0,11:5.0,16:5.0,3793:5.0,3791:5.0,3789:5.0]

12 [3255:5.0,1036:5.0,1032:5.0,1080:5.0,480:5.0,1073:5.0,1103:5.0,2640:5.0,1089:5.0,2406:5.0]

3 [1037:5.0,1036:5.0,2402:5.0,3175:5.0,2078:5.0,3108:5.0,10:5.0,1028:5.0,3104:5.0,1025:5.0]

13 [1028:5.0,1293:5.0,2194:5.0,2662:5.0,3147:5.0,3602:5.0,1101:5.0,541:5.0,2762:5.0,1090:5.0]

4 [913:5.0,1356:5.0,1968:5.0,2524:5.0,2951:5.0,1103:5.0,3698:5.0,1101:5.0,1299:5.0,541:5.0]

14 [2908:5.0,1228:5.0,858:5.0,1965:5.0,1931:5.0,1923:5.0,1997:5.0,2064:5.0,3174:5.0,2076:5.0]

5 [1734:5.0,2697:5.0,2076:5.0,3108:5.0,2067:5.0,2065:5.0,2064:5.0,1051:5.0,3095:5.0,3094:5.0]

15 [3594:5.0,3821:4.6831574,3827:4.680762,3554:4.6298866,3555:4.6030307,3879:4.57618,2834:4.5592313,2615:4.5343375,3566:4.525866,3512:4.5152755]

6 [2054:5.0,1036:5.0,5:5.0,1033:5.0,3111:5.0,2:5.0,1030:5.0,3107:5.0,2067:5.0,1042:5.0]

16 [3159:5.0,2054:5.0,1566:5.0,2089:5.0,3219:5.0,480:5.0,1028:5.0,3274:5.0,1018:5.0,48:5.0]

7 [590:5.0,553:5.0,552:5.0,1833:5.0,2641:5.0,548:5.0,3257:5.0,3448:5.0,544:5.0,376:5.0]

17 [3526:5.0,2:5.0,3521:5.0,3507:5.0,21:5.0,3504:5.0,24:5.0,25:5.0,3500:5.0,3499:5.0]

8 [3176:5.0,1921:5.0,3930:5.0,3809:5.0,2:5.0,1898:5.0,10:5.0,3803:5.0,1895:5.0,1892:5.0]

18 [3526:5.0,3525:5.0,6:5.0,3519:5.0,16:5.0,18:5.0,3508:5.0,3507:5.0,21:5.0,3505:5.0]

9 [1580:5.0,1036:5.0,2706:5.0,3755:5.0,2078:5.0,3108:5.0,3107:5.0,11:5.0,3105:5.0,3101:5.0]

19 [3526:5.0,3917:5.0,3519:5.0,3518:5.0,16:5.0,3508:5.0,3507:5.0,21:5.0,3504:5.0,24:5.0]

Given the result, I will change all the ratings of movie 1566 (which is the first result for user-id 1) all to one.

**[cloudera@centos-e185 P2]$ cat ratings.csv | grep ",1566," | wc -l**

469

**[cloudera@centos-e185 P2]$ vi ratings.csv**

**[cloudera@centos-e185 P2]$ cat ratings.csv | grep ",1566,1" | wc -l**

469

And we run the Recommender engine again, but now only for the first 2 user-ids:

**[cloudera@centos-e185 P2]$ cat user-ids.txt**

1

2

**[cloudera@centos-e185 P2]$ hadoop fs -rm Assign09/P2/\***

Deleted Assign09/P2/ratings.csv

Deleted Assign09/P2/user-ids.txt

**[cloudera@centos-e185 P2]$ hadoop fs -put user-ids.txt ratings.csv Assign09/P2**

**[cloudera@centos-e185 P2]$ mahout recommenditembased -Dmapred.reduce.tasks=10 --similarityClassname SIMILARITY\_PEARSON\_CORRELATION --input Assign09/P2/ratings.csv --output Assign09/P2-output2 --tempDir Assign09/P2-tmp2 --usersFile Assign09/P2/user-ids.txt**

(...)

Bytes Written=1869

13/04/16 08:03:07 INFO driver.MahoutDriver: Program took 888696 ms (Minutes: 14.8116)

**[cloudera@centos-e185 P2]$ hadoop fs -ls Assign09/P2-output2**

Found 11 items

-rw-r--r-- 1 cloudera supergroup 0 2013-04-16 08:03 Assign09/P2-output2/\_SUCCESS

-rw-r--r-- 1 cloudera supergroup 188 2013-04-16 08:02 Assign09/P2-output2/part-r-00000

-rw-r--r-- 1 cloudera supergroup 179 2013-04-16 08:02 Assign09/P2-output2/part-r-00001

-rw-r--r-- 1 cloudera supergroup 179 2013-04-16 08:02 Assign09/P2-output2/part-r-00002

-rw-r--r-- 1 cloudera supergroup 186 2013-04-16 08:02 Assign09/P2-output2/part-r-00003

-rw-r--r-- 1 cloudera supergroup 186 2013-04-16 08:02 Assign09/P2-output2/part-r-00004

-rw-r--r-- 1 cloudera supergroup 241 2013-04-16 08:02 Assign09/P2-output2/part-r-00005

-rw-r--r-- 1 cloudera supergroup 180 2013-04-16 08:02 Assign09/P2-output2/part-r-00006

-rw-r--r-- 1 cloudera supergroup 174 2013-04-16 08:03 Assign09/P2-output2/part-r-00007

-rw-r--r-- 1 cloudera supergroup 175 2013-04-16 08:03 Assign09/P2-output2/part-r-00008

-rw-r--r-- 1 cloudera supergroup 181 2013-04-16 08:03 Assign09/P2-output2/part-r-00009

**[cloudera@centos-e185 P2]$ hadoop fs -cat Assign09/P2-output2/part-r-\***

10 [3952:5.0,3936:5.0,3932:5.0,3929:5.0,3927:5.0,3913:5.0,3912:5.0,3877:5.0,3875:5.0,3873:5.0]

20 [1625:5.0,1653:5.0,1729:5.0,2881:5.0,2447:5.0,1027:5.0,300:5.0,800:5.0,1179:5.0,1909:5.0]

**1 [1381:5.0,1036:5.0,1033:5.0,1032:5.0,1031:5.0,1030:5.0,3107:5.0,3114:5.0,1026:5.0,1025:5.0]**

11 [1:5.0,3752:5.0,3868:5.0,1902:5.0,2:5.0,11:5.0,1895:5.0,16:5.0,3793:5.0,1885:5.0]

2 [2739:5.0,3811:5.0,3916:5.0,2:5.0,10:5.0,11:5.0,16:5.0,3793:5.0,3791:5.0,3789:5.0]

12 [3255:5.0,1036:5.0,1032:5.0,1080:5.0,480:5.0,1073:5.0,1103:5.0,2640:5.0,1089:5.0,2406:5.0]

3 [1037:5.0,1036:5.0,2402:5.0,3175:5.0,2078:5.0,3108:5.0,10:5.0,1028:5.0,3104:5.0,1025:5.0]

13 [1028:5.0,1293:5.0,2194:5.0,2662:5.0,3147:5.0,3602:5.0,1101:5.0,541:5.0,2762:5.0,1090:5.0]

4 [913:5.0,1356:5.0,1968:5.0,2524:5.0,2951:5.0,1103:5.0,3698:5.0,1101:5.0,1299:5.0,541:5.0]

14 [2908:5.0,1228:5.0,858:5.0,1965:5.0,1931:5.0,1923:5.0,1997:5.0,2064:5.0,3174:5.0,2076:5.0]

5 [1734:5.0,2697:5.0,2076:5.0,3108:5.0,2067:5.0,2065:5.0,2064:5.0,1051:5.0,3095:5.0,3094:5.0]

15 [3594:5.0,3821:4.6831627,3827:4.6807647,3554:4.6298733,3555:4.6033874,3879:4.576184,2834:4.5592065,2615:4.5343375,3566:4.525874,3512:4.5152683]

6 [2054:5.0,1036:5.0,5:5.0,1033:5.0,3111:5.0,2:5.0,1030:5.0,3107:5.0,2067:5.0,1042:5.0]

16 [3159:5.0,2054:5.0,1018:5.0,2089:5.0,3219:5.0,3274:5.0,1028:5.0,1097:5.0,480:5.0,48:5.0]

7 [590:5.0,553:5.0,552:5.0,1833:5.0,2641:5.0,548:5.0,3257:5.0,3448:5.0,544:5.0,376:5.0]

17 [3526:5.0,2:5.0,3521:5.0,3507:5.0,21:5.0,3504:5.0,24:5.0,25:5.0,3500:5.0,3499:5.0]

8 [3176:5.0,3550:5.0,3828:5.0,3809:5.0,2:5.0,1898:5.0,10:5.0,3803:5.0,1895:5.0,1892:5.0]

18 [3526:5.0,3525:5.0,6:5.0,3519:5.0,16:5.0,18:5.0,3508:5.0,3507:5.0,21:5.0,3505:5.0]

9 [1580:5.0,1036:5.0,2706:5.0,3755:5.0,2078:5.0,3108:5.0,3107:5.0,11:5.0,3105:5.0,3101:5.0]

19 [3526:5.0,3917:5.0,3519:5.0,3518:5.0,16:5.0,3508:5.0,3507:5.0,21:5.0,3504:5.0,24:5.0]

As we can see on the selected item for user-id 1, the movie 1566 has disappeared from the list, as expected.

**Problem 3. Please reproduce results for Mahout Classifier as described on pages 43 to 48 of lecture notes. Select two or three junk emails from your junk folder and examine how will Mahout classifier treat them.**

To start out, we create a directory for this problem and download the corpus from spamassassin:

**[cloudera@centos-e185 Assign09]$ mkdir P3**

**[cloudera@centos-e185 Assign09]$ cd P3**

**[cloudera@centos-e185 P3]$ curl -O http://spamassassin.apache.org/publiccorpus/20021010\_spam.tar.bz2**

% Total % Received % Xferd Average Speed Time Time Time Current

Dload Upload Total Spent Left Speed

100 1164k 100 1164k 0 0 45892 0 0:00:25 0:00:25 --:--:-- 73882

**[cloudera@centos-e185 P3]$ curl -O http://spamassassin.apache.org/publiccorpus/20021010\_easy\_ham.tar.bz2**

% Total % Received % Xferd Average Speed Time Time Time Current

Dload Upload Total Spent Left Speed

100 1637k 100 1637k 0 0 107k 0 0:00:15 0:00:15 --:--:-- 98611

**[cloudera@centos-e185 P3]$ tar -xjf 20021010\_spam.tar.bz2**

**[cloudera@centos-e185 P3]$ tar -xjf 20021010\_easy\_ham.tar.bz2**

**[cloudera@centos-e185 P3]$ ls**

20021010\_easy\_ham.tar.bz2 20021010\_spam.tar.bz2 easy\_ham spam

Then we follow the instructions to parse the data and and separate it into training and test datasets

**[cloudera@centos-e185 P3]$ cp -R easy\_ham/ spam/ 20news-all/**

**[cloudera@centos-e185 P3]$ ls 20news-all/**

easy\_ham spam

**[cloudera@centos-e185 P3]$ hadoop fs -mkdir Assign09/P3**

**[cloudera@centos-e185 P3]$ hadoop fs -put 20news-all/ Assign09/P3**

**[cloudera@centos-e185 P3]$ hadoop fs -ls Assign09/P3/20news-all**

Found 2 items

drwxr-xr-x - cloudera supergroup 0 2013-04-16 10:48 Assign09/P3/20news-all/easy\_ham

drwxr-xr-x - cloudera supergroup 0 2013-04-16 10:48 Assign09/P3/20news-all/spam

We prepare data by sequencing into vectors and splitting it:

**[cloudera@centos-e185 P3]$ mahout seqdirectory -i Assign09/P3/20news-all -o Assign09/P3/20news-seq**

(..)

13/04/16 10:51:08 INFO driver.MahoutDriver: Program took 9022 ms (Minutes: 0.15036666666666668)

**[cloudera@centos-e185 P3]$ mahout seqdirectory -i Assign09/P3/20news-all -o Assign09/P3/20news-seq**

(...)

13/04/16 10:55:00 INFO common.HadoopUtil: Deleting Assign09/P3/20news-vectors/partial-vectors-0

13/04/16 10:55:00 INFO driver.MahoutDriver: Program took 142047 ms (Minutes: 2.36745)

**[cloudera@centos-e185 P3]$ mahout split -i Assign09/P3/20news-vectors/tfidf-vectors --trainingOutput Assign09/P3/20news-train-vectors --testOutput Assign09/P3/20news-test-vectors --randomSelectionPct 20 --overwrite --sequenceFiles -xm sequential**

(...)

13/04/16 10:58:51 INFO utils.SplitInput: part-r-00000 has 48351 lines

13/04/16 10:58:51 INFO utils.SplitInput: part-r-00000 test split size is 9670 based on random selection percentage 20

13/04/16 10:58:51 INFO zlib.ZlibFactory: Successfully loaded & initialized native-zlib library

13/04/16 10:58:51 INFO compress.CodecPool: Got brand-new compressor [.deflate]

13/04/16 10:58:51 INFO compress.CodecPool: Got brand-new compressor [.deflate]

13/04/16 10:58:52 INFO utils.SplitInput: file: part-r-00000, input: 48351 train: 2429, test: 623 starting at 0

13/04/16 10:58:52 INFO driver.MahoutDriver: Program took 3209 ms (Minutes: 0.053483333333333334)

Now, we build out classification model with the data we separated as "train":

**[cloudera@centos-e185 P3]$ mahout trainnb -i Assign09/P3/20news-train-vectors -el -o Assign09/P3/model -li labelindex -ow -c**

File Input Format Counters

Bytes Read=438318

File Output Format Counters

Bytes Written=368124

13/04/16 11:01:05 INFO driver.MahoutDriver: Program took 35229 ms (Minutes: 0.58715)

Finally, we test out our shiny new model against the original training data and the test data.

**[cloudera@centos-e185 P3]$ mahout testnb -i Assign09/P3/20news-train-vectors -m Assign09/P3/model -l labelindex -ow -o Assign09/P3/20news-testing-train -c**

(...)

13/04/16 11:06:22 INFO test.TestNaiveBayesDriver: Complementary Results: =======================================================

Summary

-------------------------------------------------------

Correctly Classified Instances : 2429 100%

Incorrectly Classified Instances : 0 0%

Total Classified Instances : 2429

=======================================================

Confusion Matrix

-------------------------------------------------------

a b <--Classified as

2022 0 | 2022 a = easy\_ham

0 407 | 407 b = spam

13/04/16 11:06:22 INFO driver.MahoutDriver: Program took 23118 ms (Minutes: 0.38531666666666664)

**[cloudera@centos-e185 P3]$ mahout testnb -i Assign09/P3/20news-test-vectors -m Assign09/P3/model -l labelindex -ow -o Assign09/P3/20news-testing-test -c**

(...)

13/04/16 11:08:19 INFO test.TestNaiveBayesDriver: Complementary Results: =======================================================

Summary

-------------------------------------------------------

Correctly Classified Instances : 620 99.5185%

Incorrectly Classified Instances : 3 0.4815%

Total Classified Instances : 623

=======================================================

Confusion Matrix

-------------------------------------------------------

a b <--Classified as

527 2 | 529 a = easy\_ham

1 93 | 94 b = spam

13/04/16 11:08:19 INFO driver.MahoutDriver: Program took 12832 ms (Minutes: 0.21386666666666668)

As expected, it does perfectly on the original training set and does a pretty good job on the test set.

We now create 3 new spam files from my e-mail client and copy them to the server. We then proceed to process them to the right format as it was done previously for the other data files.

**[cloudera@centos-e185 P3]$ hadoop fs -mkdir Assign09/P3/email-alexcp/spam**

**[cloudera@centos-e185 P3]$ hadoop fs -put \*.eml Assign09/P3/email-alexcp/spam**

**[cloudera@centos-e185 P3]$ hadoop fs -ls Assign09/P3/email-alexcp/spam**

Found 3 items

-rw-r--r-- 1 cloudera supergroup 1436 2013-04-16 11:43 Assign09/P3/email-alexcp/spam/001.eml

-rw-r--r-- 1 cloudera supergroup 3615 2013-04-16 11:43 Assign09/P3/email-alexcp/spam/002.eml

-rw-r--r-- 1 cloudera supergroup 1901 2013-04-16 11:43 Assign09/P3/email-alexcp/spam/003.eml

**[cloudera@centos-e185 P3]$ mahout seqdirectory -i Assign09/P3/email-alexcp -o Assign09/P3/email-alexcp-seq**

(...)

13/04/16 11:45:00 INFO driver.MahoutDriver: Program took 1362 ms (Minutes: 0.0227)

13/04/16 11:47:45 INFO common.HadoopUtil: Deleting Assign09/P3/email-alexcp-vectors/partial-vectors-0

13/04/16 11:47:45 INFO driver.MahoutDriver: Program took 103046 ms (Minutes: 1.7174333333333334)

**[cloudera@centos-e185 P3]$ mahout seq2sparse -i Assign09/P3/email-alexcp-seq -o Assign09/P3/email-alexcp-vectors -lnorm -nv -wt tfidf**

(...)

13/04/16 11:47:45 INFO common.HadoopUtil: Deleting Assign09/P3/email-alexcp-vectors/partial-vectors-0

13/04/16 11:47:45 INFO driver.MahoutDriver: Program took 103046 ms (Minutes: 1.7174333333333334)

Finally we test our data:

**[cloudera@centos-e185 P3]$ mahout testnb -i Assign09/P3/email-alexcp-vectors/tfidf-vectors -m Assign09/P3/model -l labelindex -ow -o Assign09/P3/email-alexcp-test -c**

13/04/16 11:51:12 INFO test.TestNaiveBayesDriver: Complementary Results: =======================================================

Summary

-------------------------------------------------------

Correctly Classified Instances : 0 0%

Incorrectly Classified Instances : 3 100%

Total Classified Instances : 3

=======================================================

Confusion Matrix

-------------------------------------------------------

a b <--Classified as

0 0 | 0 a = easy\_ham

3 0 | 3 b = spam

13/04/16 11:51:12 INFO driver.MahoutDriver: Program took 13009 ms (Minutes: 0.21681666666666666)

Unfortunately, it did very badly. I guess it needs more training against the new types of SPAMs we have currently, given that the corpus is from 2002.

**Problem 4. On Mahout wiki there is a small tutorial on synthetic data (** [**https://cwiki.apache.org/confluence/display/MAHOUT/Clustering+of+synthetic+control+data**](https://cwiki.apache.org/confluence/display/MAHOUT/Clustering+of+synthetic+control+data%20) **). Please run the tutorial for K-means and one other clustering algorithm of your choice. Compare the results of those two clustering algorithms.**

We start out by loading the data into our VM

**[cloudera@centos-e185 Assign09]$ mkdir P4**

**[cloudera@centos-e185 Assign09]$ cd P4**

**[cloudera@centos-e185 P4]$ curl -O http://archive.ics.uci.edu/ml/databases/synthetic\_control/synthetic\_control.data**

% Total % Received % Xferd Average Speed Time Time Time Current

Dload Upload Total Spent Left Speed

100 281k 100 281k 0 0 262k 0 0:00:01 0:00:01 --:--:-- 284k

**[cloudera@centos-e185 P4]$ hadoop fs -mkdir testdata**

**[cloudera@centos-e185 P4]$ hadoop fs -put synthetic\_control.data testdata**

**[cloudera@centos-e185 P4]$ hadoop fs -ls testdata**

Found 1 items

-rw-r--r-- 1 cloudera supergroup 288374 2013-04-16 12:07 testdata/synthetic\_control.data

And then we running the K-means clustering first

**[cloudera@centos-e185 P4]$ mahout org.apache.mahout.clustering.syntheticcontrol.kmeans.Job**

(...)

13/04/16 12:22:47 INFO clustering.ClusterDumper: Wrote 12 clusters

13/04/16 12:22:47 INFO driver.MahoutDriver: Program took 190580 ms (Minutes: 3.1763333333333335)

We then copy the output locally and use the mahout clusterdump tool to get a human readable format ou the output.

**[cloudera@centos-e185 P4]$ mkdir kmeans**

**[cloudera@centos-e185 P4]$ hadoop fs -get output kmeans/**

**[cloudera@centos-e185 P4]$ ls kmeans/**

output

**[cloudera@centos-e185 P4]$ mahout clusterdump --input output/clusters-10-final --pointsDir output/clusteredPoints --output kmeans/output/clusteranalyze.txt**

(...)

13/04/16 12:33:40 INFO clustering.ClusterDumper: Wrote 12 clusters

13/04/16 12:33:40 INFO driver.MahoutDriver: Program took 2057 ms (Minutes: 0.03428333333333333)

**[cloudera@centos-e185 P4]$ cat kmeans/output/clusteranalyze.txt**

CL-579{n=15 c=[30.350, 29.798, 30.595, 30.128, 29.711, 30.557, 30.376, 31.248, 31.516, 30.922, 30.765, 30.945, 30.735, 29.992, 28.272, 29.783, 30.143, 29.726, 29.899, 30.281, 30.622, 29.825, 30.697, 29.938, 30.338, 29.717, 29.356, 29.292, 28.392, 27.415, 29.147, 27.915, 25.968, 25.843, 22.653, 19.195, 15.893, 15.073, 14.859, 17.139, 18.528, 18.042, 14.421, 15.809, 16.712, 17.921, 15.378, 15.338, 15.218, 16.648, 17.330, 16.334, 15.999, 15.873, 14.041, 16.100, 17.956, 14.773, 15.128, 16.869] r=[3.364, 3.176, 3.498, 2.720, 2.646, 3.048, 3.667, 3.161, 2.534, 3.572, 3.711, 4.311, 3.625, 3.178, 3.265, 3.118, 3.140, 3.594, 3.767, 3.244, 3.802, 3.351, 3.913, 3.358, 2.982, 3.542, 2.776, 3.519, 3.216, 5.555, 4.984, 6.205, 5.295, 7.336, 7.271, 6.330, 4.018, 3.619, 3.451, 4.399, 2.774, 3.711, 3.308, 3.575, 3.219, 3.444, 3.761, 3.442, 3.147, 3.649, 3.209, 4.526, 4.636, 4.398, 3.159, 3.961, 4.235, 3.535, 3.326, 3.498]}

(...)

We can easily repeat this process for Canopy clustering:

**[cloudera@centos-e185 P4]$ mahout org.apache.mahout.clustering.syntheticcontrol.canopy.Job**

(...)

13/04/16 12:38:05 INFO clustering.ClusterDumper: Wrote 6 clusters

13/04/16 12:38:05 INFO driver.MahoutDriver: Program took 41070 ms (Minutes: 0.6845)

**[cloudera@centos-e185 P4]$ mahout clusterdump --input output/clusters-0-final --pointsDir output/clusteredPoints --output canopy/output/clusteranalyze.txt**

(...)

13/04/16 12:41:27 INFO clustering.ClusterDumper: Wrote 6 clusters

13/04/16 12:41:27 INFO driver.MahoutDriver: Program took 2276 ms (Minutes: 0.03793333333333333)

**[cloudera@centos-e185 P4]$ cat canopy/output/clusteranalyze.txt**

C-0{n=9 c=[30.071, 31.088, 31.885, 31.942, 31.848, 31.503, 30.350, 29.729, 28.543, 28.620, 27.378, 27.514, 27.882, 28.733, 28.990, 29.920, 30.321, 30.484, 30.647, 30.473, 30.092, 27.739, 26.843, 26.265, 25.678, 24.522, 25.253, 25.728, 25.370, 25.716, 25.837, 26.172, 25.384, 25.008, 23.279, 22.892, 21.290, 20.565, 19.599, 19.733, 19.614, 19.605, 20.332, 20.409, 21.045, 21.460, 20.981, 21.349, 21.164, 20.679, 20.158, 19.749, 18.930, 18.254, 18.313, 18.818, 19.311, 19.673, 19.604, 20.898] r=[0.183, 1.462, 2.780, 3.546, 3.481, 3.226, 1.672, 1.132, 1.186, 2.301, 3.296, 2.987, 2.537, 1.837, 1.617, 2.417, 3.044, 4.008, 3.886, 3.688, 3.049, 2.733, 1.988, 2.034, 2.252, 2.901, 2.537, 2.770, 3.181, 4.115, 4.965, 6.178, 6.628, 6.887, 6.646, 5.970, 5.866, 4.883, 4.657, 4.643, 4.652, 5.582, 5.969, 6.598, 7.555, 8.148, 8.941, 8.269, 7.764, 7.117, 6.315, 6.004, 5.574, 5.688, 5.978, 6.526, 6.682, 7.223, 7.878, 8.267]}

(...)

The results are different from each other, where the K-means clustering was able to find 12 clusters in the data, and the Canopy clustering found 6.