Homework 2

Alejandro J. Rojas
ale@ischool.berkeley.edu
W261: Machine Learning at Scale
Week: 02
Jan 26, 2016, 1:15 PM</li>

HW2.0.

What is a race condition in the context of parallel computation? Give an example. What is MapReduce? How does it differ from Hadoop? Which programming paradigm is Hadoop based on? Explain and give a simple example in code and show the code running.

Race conditions

Race conditions essentially refers to situations where the results of one process is dependant on the order of when other processes are executed. For instance any processes that involve transactions are usually in this category and they're managed by establishing locks in databases so that transactions are updated against the locked database and executed in order and not simultaneously. In the context of paralalization when many processes occur simultaneosly, it is important to understand which processes can be embarassingly parallel and which can't so that we design data architecture to deal accordingly.

Map Reduce

Method essentially developed by Google in which processes run embarassingly parallel following two major steps. In the map stage, data is broken into chunks and mapped to a specific output. In the reduce stage, data from chunks is consolidated and a summary of that data is output.

Hadoop

Hadoop is an open source project that was developed to not only implement the Map Reduce programming model but also to serve as a distributed file system of storage under a paradigm of parallel computing running on a great number of 'cheap' servers that are prone to fail. This distributed files system provides fault tolerance against those failures by implementing replicates of all data. Data is broken into chunks that are replicated usually 3 times across a cluster of computer nodes.

A simplified example of Hadoop can be as follows:

Imagine a very viral video that is published on Facebook. This is a 1280 x 720 high resolution video that lasts about 5 minutes and it's a file of about 350 MB in size. Facebook stores the video as an object that is broken into 64MB chunks resulting roughly in about one chunk for everey minute of video. Each chunk is replicated by Facebook at least 3 times across its clusters of computer nodes.Facebook also serves this combined file on multiple servers depending on where the user wants to watch the video. If a user "likes" the video, then this information is attached to a log that is stored in the server where the like was collected. A "Map" job then assigns likes, comments, and shares to that video using a json format on each server where the video is served. A "Reduce" job consolidates all jsons from all the computer nodes were the video was served to then totalize and come up with the number of "Likes" that the video shows.

In pseudocode we can see the following:

User "likes" video object v

Map job

```
inputs:
    object_v = {object_id: video, publisher_id: publisher, date_of_publishing: date, ....}
    user = {user_id: unique identifier, name: user name, lastname: user lastname, ...}
function get_like(object_id, user_id):
object_v[like]=[user]

    return object_v
```

Reduce job

```
function sum_likes(object_v):
    for node in nodes:
        for like in object_v:
            sum_likes = sum_likes +1

     return sum_likes
```

HW2.1. Sort in Hadoop MapReduce

Given as input: Records of the form '<'integer, “NA”'>', where integer is any integer, and “NA” is just the empty string. Output: sorted key value pairs of the form '<'integer, “NA”'>' in decreasing order; what happens if you have multiple reducers? Do you need additional steps? Explain.

Write code to generate N random records of the form '<'integer, “NA”'>'. Let N = 10,000. Write the python Hadoop streaming map-reduce job to perform this sort. Display the top 10 biggest numbers. Display the 10 smallest numbers

Data

In [224]:
```
import random

N = 10000                           ### for a sample size of N
random.seed(0)                      ### pick a random seed to replicate results


input_file = open("numcount.txt", "w") # writing file

for i in range(N):
    a = random.randint(0, 100)       ### Select a random integer from 0 to 100
    b = ''
    input_file.write(str(a))
    input_file.write(b)
    input_file.write('\n')



input_file.close()
```

Mapper

In [225]:
```
%%writefile mapper.py
#!/usr/bin/python
import sys

for line in sys.stdin:                        ### input comes from STDIN (standard input)

    number = line.strip()                     ### remove leading and trailing whitespace
    print ('%s\t%s' % (number, 1))            ### mapper out looks like 'number' \t 1
```

```
Overwriting mapper.py
```

In [298]:
```
!chmod +x mapper.py
```

Reducer

In [227]:
```
%%writefile reducer.py
#!/usr/bin/python
from operator import itemgetter
import sys

current_number = None
current_count = 0
number = None
numlist = []

# input comes from STDIN
for line in sys.stdin:

    line = line.strip()                        ### remove leading and trailing whitespace
    line = line.split('\t')                    ### parse the input we got from mapper.py
    number = line[0]                           ### integer generated randomly we got from m
apper.py



    try:
        count = line[1]
        count = int(count)                      ### convert count (currently a string) to i
nt
    except ValueError:                          ### if count was not a number then silently
        continue                                ### ignore/discard this line



    if current_number == number:                ### this IF-switch only works because Hadoo
p sorts map output
        current_count += count                  ### by key (here: number) before it is pass
ed to the reducer
    else:
        if current_number:

            numlist.append((current_number,current_count))  ### store tuple in a list once
totalize count per number

        current_count = count                   ### set current count
```

```
        current_number = number                 ### set current number


if current_number == number:                    ### do not forget to output the last word i
f needed!
    numlist.append((current_number,current_count))


toplist = sorted(numlist,key=lambda record: record[1], reverse=True) ### sort list from lar
gest count to smallest
bottomlist = sorted(numlist,key=lambda record: record[1])  ### sort list from smalles to la
rgest

print '%25s' %'TOP 10', '%25s' % '', '%28s' %'BOTTOM 10'
print '%20s' %'Number', '%10s' %'Count', '%20s' % '', '%20s' %'Number','%10s' %'Count'
for i in range (10):
    print '%20s%10s' % (toplist[i][0], toplist[i][1]),'%20s' % '', '%20s%10s' % (bottomlis
t[i][0], bottomlist[i][1])
```

```
Overwriting reducer.py
```

In [228]:
```
!chmod +x reducer.py
```

Run numcount in Hadoop

start yarn and hdfs

In [201]:
```
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-yarn.sh       ### start up yarn
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-dfs.sh        ### start up dfs
```

```
starting yarn daemons
starting resourcemanager, logging to /usr/local/Cellar/hadoop/2.7.1/libexec/logs/yarn-vena
max-resourcemanager-VENAMAX.local.out
localhost: starting nodemanager, logging to /usr/local/Cellar/hadoop/2.7.1/libexec/logs/ya
rn-venamax-nodemanager-VENAMAX.local.out
16/01/24 11:33:49 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/Cellar/hadoop/2.7.1/libexec/logs/hadoo
p-venamax-namenode-VENAMAX.local.out
localhost: starting datanode, logging to /usr/local/Cellar/hadoop/2.7.1/libexec/logs/hadoo
p-venamax-datanode-VENAMAX.local.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/Cellar/hadoop/2.7.1/libexec/log
s/hadoop-venamax-secondarynamenode-VENAMAX.local.out
16/01/24 11:34:13 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
```

remove files from prior runs

In [342]:
```
!hdfs dfs -rm -r /user/venamax                           ### remove prior files
```

```
16/01/25 21:51:00 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
16/01/25 21:51:01 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion inter
val = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/venamax
```

create folder

In [343]:
```
!hdfs dfs -mkdir -p /user/venamax                          ### create hdfs folder
```

```
16/01/25 21:51:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
```

upload numcount.txt to hdfs

In [232]:
```
!hdfs dfs -put numcount.txt /user/venamax                  #### save source data file to hdf
s
```

```
16/01/24 11:42:43 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
```

Hadoop streaming command

hadoop jar hadoopstreamingjarfile \

```
-D stream.num.map.output.key.fields=n \
-mapper mapperfile \
-reducer reducerfile \
-input inputfile \
-output outputfile
```

In [233]:
```
!hadoop jar hadoop-*streaming*.jar -mapper mapper.py -reducer reducer.py -input numcount.tx
t -output numcountOutput
```

```
16/01/24 11:42:50 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
16/01/24 11:42:51 INFO Configuration.deprecation: session.id is deprecated. Instead, use d
fs.metrics.session-id
16/01/24 11:42:51 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracke
r, sessionId=
16/01/24 11:42:51 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobT
racker, sessionId= - already initialized
16/01/24 11:42:51 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/24 11:42:52 INFO mapreduce.JobSubmitter: number of splits:1
16/01/24 11:42:53 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1147326
746_0001
16/01/24 11:42:53 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/24 11:42:53 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/24 11:42:53 INFO mapreduce.Job: Running job: job_local1147326746_0001
16/01/24 11:42:53 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapre
d.FileOutputCommitter
16/01/24 11:42:53 INFO output.FileOutputCommitter: File Output Committer Algorithm version
is 1
16/01/24 11:42:53 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/24 11:42:53 INFO mapred.LocalJobRunner: Starting task: attempt_local1147326746_000
1_m_000000_0
16/01/24 11:42:53 INFO output.FileOutputCommitter: File Output Committer Algorithm version
is 1
16/01/24 11:42:53 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is su
pported only on Linux.
16/01/24 11:42:53 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/24 11:42:53 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/venama
x/numcount.txt:0+29121
16/01/24 11:42:53 INFO mapred.MapTask: numReduceTasks: 1
16/01/24 11:42:53 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/24 11:42:53 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/24 11:42:53 INFO mapred.MapTask: soft limit at 83886080
16/01/24 11:42:53 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/24 11:42:53 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/24 11:42:53 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapr
ed.MapTask$MapOutputBuffer
16/01/24 11:42:53 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/venamax/ucb-261/hw
2/./mapper.py]
16/01/24 11:42:53 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, us
```

```
e mapreduce.task.id
16/01/24 11:42:53 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead,
use mapreduce.cluster.local.dir
16/01/24 11:42:53 INFO Configuration.deprecation: map.input.file is deprecated. Instead, u
se mapreduce.map.input.file
16/01/24 11:42:53 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, u
se mapreduce.job.skiprecords
16/01/24 11:42:53 INFO Configuration.deprecation: map.input.length is deprecated. Instead,
use mapreduce.map.input.length
16/01/24 11:42:53 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. In
stead, use mapreduce.task.output.dir
16/01/24 11:42:53 INFO Configuration.deprecation: map.input.start is deprecated. Instead,
use mapreduce.map.input.start
16/01/24 11:42:53 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, us
e mapreduce.job.id
16/01/24 11:42:53 INFO Configuration.deprecation: user.name is deprecated. Instead, use ma
preduce.job.user.name
16/01/24 11:42:53 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instea
d, use mapreduce.task.ismap
16/01/24 11:42:53 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, u
se mapreduce.task.attempt.id
16/01/24 11:42:53 INFO Configuration.deprecation: mapred.task.partition is deprecated. Ins
tead, use mapreduce.task.partition
16/01/24 11:42:53 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/24 11:42:53 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/24 11:42:53 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/24 11:42:53 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/24 11:42:53 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/24 11:42:54 INFO mapreduce.Job: Job job_local1147326746_0001 running in uber mode :
false
16/01/24 11:42:54 INFO mapreduce.Job:  map 0% reduce 0%
16/01/24 11:42:55 INFO streaming.PipeMapRed: Records R/W=10000/1
16/01/24 11:42:55 INFO streaming.PipeMapRed: MRErrorThread done
16/01/24 11:42:55 INFO streaming.PipeMapRed: mapRedFinished
16/01/24 11:42:55 INFO mapred.LocalJobRunner:
16/01/24 11:42:55 INFO mapred.MapTask: Starting flush of map output
16/01/24 11:42:55 INFO mapred.MapTask: Spilling map output
16/01/24 11:42:55 INFO mapred.MapTask: bufstart = 0; bufend = 49121; bufvoid = 104857600
16/01/24 11:42:55 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26174400(104
697600); length = 39997/6553600
```

```
16/01/24 11:42:55 INFO mapred.MapTask: Finished spill 0
16/01/24 11:42:55 INFO mapred.Task: Task:attempt_local1147326746_0001_m_000000_0 is done.
And is in the process of committing
16/01/24 11:42:55 INFO mapred.LocalJobRunner: Records R/W=10000/1
16/01/24 11:42:55 INFO mapred.Task: Task 'attempt_local1147326746_0001_m_000000_0' done.
16/01/24 11:42:55 INFO mapred.LocalJobRunner: Finishing task: attempt_local1147326746_000
1_m_000000_0
16/01/24 11:42:55 INFO mapred.LocalJobRunner: map task executor complete.
16/01/24 11:42:55 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/24 11:42:55 INFO mapred.LocalJobRunner: Starting task: attempt_local1147326746_000
1_r_000000_0
16/01/24 11:42:55 INFO output.FileOutputCommitter: File Output Committer Algorithm version
is 1
16/01/24 11:42:55 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is su
pported only on Linux.
16/01/24 11:42:55 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/24 11:42:55 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.m
apreduce.task.reduce.Shuffle@32812234
16/01/24 11:42:55 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxS
ingleShuffleLimit=83584616, mergeThreshold=220663392, ioSortFactor=10, memToMemMergeOutput
sThreshold=10
16/01/24 11:42:55 INFO reduce.EventFetcher: attempt_local1147326746_0001_r_000000_0 Thread
started: EventFetcher for fetching Map Completion Events
16/01/24 11:42:55 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map
attempt_local1147326746_0001_m_000000_0 decomp: 69123 len: 69127 to MEMORY
16/01/24 11:42:55 INFO reduce.InMemoryMapOutput: Read 69123 bytes from map-output for atte
mpt_local1147326746_0001_m_000000_0
16/01/24 11:42:55 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 6
9123, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->69123
16/01/24 11:42:55 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/24 11:42:55 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/24 11:42:55 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-out
puts and 0 on-disk map-outputs
16/01/24 11:42:55 INFO mapreduce.Job:  map 100% reduce 0%
16/01/24 11:42:55 INFO mapred.Merger: Merging 1 sorted segments
16/01/24 11:42:55 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of
total size: 69119 bytes
16/01/24 11:42:55 INFO reduce.MergeManagerImpl: Merged 1 segments, 69123 bytes to disk to
satisfy reduce memory limit
16/01/24 11:42:55 INFO reduce.MergeManagerImpl: Merging 1 files, 69127 bytes from disk
```

```
16/01/24 11:42:55 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory in
to reduce
16/01/24 11:42:55 INFO mapred.Merger: Merging 1 sorted segments
16/01/24 11:42:55 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of
total size: 69119 bytes
16/01/24 11:42:55 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/24 11:42:55 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/venamax/ucb-261/hw
2/./reducer.py]
16/01/24 11:42:55 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instea
d, use mapreduce.jobtracker.address
16/01/24 11:42:55 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead,
use mapreduce.job.maps
16/01/24 11:42:55 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/24 11:42:55 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/24 11:42:55 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/24 11:42:55 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/24 11:42:55 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/24 11:42:55 INFO streaming.PipeMapRed: MRErrorThread done
16/01/24 11:42:55 INFO streaming.PipeMapRed: Records R/W=10000/1
16/01/24 11:42:55 INFO streaming.PipeMapRed: mapRedFinished
16/01/24 11:42:55 INFO mapred.Task: Task:attempt_local1147326746_0001_r_000000_0 is done.
And is in the process of committing
16/01/24 11:42:55 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/24 11:42:55 INFO mapred.Task: Task attempt_local1147326746_0001_r_000000_0 is allowe
d to commit now
16/01/24 11:42:55 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1147
326746_0001_r_000000_0' to hdfs://localhost:9000/user/venamax/numcountOutput/_temporar
y/0/task_local1147326746_0001_r_000000
16/01/24 11:42:55 INFO mapred.LocalJobRunner: Records R/W=10000/1 > reduce
16/01/24 11:42:55 INFO mapred.Task: Task 'attempt_local1147326746_0001_r_000000_0' done.
16/01/24 11:42:55 INFO mapred.LocalJobRunner: Finishing task: attempt_local1147326746_000
1_r_000000_0
16/01/24 11:42:55 INFO mapred.LocalJobRunner: reduce task executor complete.
16/01/24 11:42:56 INFO mapreduce.Job:  map 100% reduce 100%
16/01/24 11:42:56 INFO mapreduce.Job: Job job_local1147326746_0001 completed successfully
16/01/24 11:42:56 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=350072
                FILE: Number of bytes written=1007367
                FILE: Number of read operations=0
```

```
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=58242
                HDFS: Number of bytes written=1008
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=10000
                Map output records=10000
                Map output bytes=49121
                Map output materialized bytes=69127
                Input split bytes=99
                Combine input records=0
                Combine output records=0
                Reduce input groups=101
                Reduce shuffle bytes=69127
                Reduce input records=10000
                Reduce output records=12
                Spilled Records=20000
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=4
                Total committed heap usage (bytes)=551550976
        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=29121
        File Output Format Counters
                Bytes Written=1008
16/01/24 11:42:56 INFO streaming.StreamJob: Output directory: numcountOutput
```

show the results

In [234]:
```
!hdfs dfs -cat numcountOutput/part-00000
```

```
16/01/24 11:43:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
                  TOP 10                                                  BOTTOM 10
              Number      Count                                       Number      Count
                  77       125                                           9        76
                  42       122                                          88        77
                  90       117                                          70        83
                  26       116                                          15        85
                  80       115                                          35        85
                  14       114                                          41        85
                  86       113                                          59        85
                  45       112                                          73        85
                  53       112                                           7        86
                  13       111                                          23        87
```

stop yarn and hdfs

In [169]:
```
!/usr/local/Cellar/hadoop/2.7.1/sbin/stop-yarn.sh
!/usr/local/Cellar/hadoop/2.7.1/sbin/stop-dfs.sh
```

```
stopping yarn daemons
stopping resourcemanager
localhost: stopping nodemanager
no proxyserver to stop
16/01/23 21:13:33 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
Stopping namenodes on [localhost]
localhost: stopping namenode
localhost: stopping datanode
Stopping secondary namenodes [0.0.0.0]
0.0.0.0: stopping secondarynamenode
16/01/23 21:13:54 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
```

========== END OF HW 2.1

HW2.2. WORDCOUNT

Using the Enron data from HW1 and Hadoop MapReduce streaming, write the mapper/reducer job that will determine the word count (number of occurrences) of each white-space delimitted token (assume spaces, fullstops, comma as delimiters). Examine the word “assistance” and report its word count results.

CROSSCHECK: >grep assistance enronemail_1h.txt|cut -d$'\t' -f4| grep assistance|wc -l
8

```
#NOTE  "assistance" occurs on 8 lines but how many times does the token occur? 10 times! Thi
s is the number we are looking for!
```

Mapper

In [406]:
```
%%writefile mapper.py
#!/usr/bin/python
## mapper.py
## Author: Alejandro J. Rojas
## Description: mapper code for HW2.2

import sys
import re



########## Collect user input  ###############
filename = 'enronemail_1h.txt'
findwords = ['assistance']
counts = {}
##filename = sys.argv[1]
##findwords = re.split(" ",sys.argv[2].lower())


for keyword in findwords:  ### Initialize to zero all keywords to find
    counts[keyword] = 0

for line in sys.stdin:                          ### input comes from STDIN (standard input)
    record = re.split(r'\t+', line)
    if len(record) == 4:                            ### Take only complete records
        for keyword in findwords:                   ### For each word to find
            for i in range (2,len(record)):         ### Compare it to the words from each
email
                bagofwords = re.split(" ",record[i]) ### Break each email records into word
s
                for word in bagofwords:
                    neword = word.strip(',')        ### eliminate comas
                    if keyword in neword:
                        counts[keyword] += 1


for keyword in findwords:                                ### output results in the form:
    print ('%s\t%s'% (keyword, str(counts[keyword])))     ### word to find, count
```

```
Overwriting mapper.py
```

In [407]:
```
!chmod +x mapper.py
```

Reducer

In [408]:
```
%%writefile reducer.py
#!/usr/bin/python
from operator import itemgetter
import sys

current_word = None
current_count = 0
word = None
words = {}

# input comes from STDIN


# input comes from STDIN
for line in sys.stdin:

    line = line.strip()                        ### remove leading and trailing whitespace
    line = line.split('\t')                    ### parse the input we got from mappe
    word = line[0]

    try:

        count = line[1]
        count = int(count)                      ### convert count (currently a string) to i
nt
    except ValueError:                          ### if count was not a number then silently
        continue                                ### ignore/discard this line



    if current_word == word:                     ### this IF-switch only works because Hado
op sorts map output
        current_count += count                  ### by key (here: number) before it is pass
ed to the reducer
    else:
        if current_word:

            words[current_word] = current_count  ### store tuple in a list once totalize co
unt per number
```

```
        current_count = count                    ### set current count
        current_word = word                      ### set current word


if current_word == word:                         ### do not forget to output the last word
if needed!
    words[current_word] = current_count

for word in words:
    print 'We found "%s"' %word, ' on %s'%words[word] , 'occassions.'
```

```
Overwriting reducer.py
```

In [409]:
```
!chmod +x reducer.py
```

Run numcount in Hadoop

start yarn and hdfs

In [114]:
```
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-yarn.sh       ### start up yarn
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-dfs.sh        ### start up dfs
```

```
starting yarn daemons
resourcemanager running as process 12434. Stop it first.
localhost: nodemanager running as process 13130. Stop it first.
16/01/25 16:46:12 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: namenode running as process 13875. Stop it first.
localhost: datanode running as process 14563. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 15278. Stop it first.
16/01/25 16:46:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
```

remove files from prior runs

In [410]:
```
!hdfs dfs -rm -r /user/venamax                          ### remove prior files
```

```
16/01/25 22:58:09 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
16/01/25 22:58:10 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion inter
val = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/venamax
```

create folder

In [411]:
```
!hdfs dfs -mkdir -p /user/venamax                        ### create hdfs folder
```

```
16/01/25 22:58:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
```

upload enronemail_1h.txt to hdfs

In [412]:
```
!hdfs dfs -put enronemail_1h.txt /user/venamax           #### save source data file to hdf
s
```

```
16/01/25 22:58:23 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
```

Hadoop streaming command

In [413]:
```
!hadoop jar hadoop-*streaming*.jar -mapper mapper.py -reducer reducer.py -input enronemai
l_1h.txt -output wordcountOutput
```

```
16/01/25 22:58:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
16/01/25 22:58:45 INFO Configuration.deprecation: session.id is deprecated. Instead, use d
fs.metrics.session-id
16/01/25 22:58:45 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracke
r, sessionId=
16/01/25 22:58:45 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobT
racker, sessionId= - already initialized
16/01/25 22:58:45 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/25 22:58:45 INFO mapreduce.JobSubmitter: number of splits:1
16/01/25 22:58:45 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1108989
492_0001
16/01/25 22:58:45 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/25 22:58:45 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/25 22:58:45 INFO mapreduce.Job: Running job: job_local1108989492_0001
16/01/25 22:58:45 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapre
d.FileOutputCommitter
16/01/25 22:58:45 INFO output.FileOutputCommitter: File Output Committer Algorithm version
is 1
16/01/25 22:58:45 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/25 22:58:45 INFO mapred.LocalJobRunner: Starting task: attempt_local1108989492_000
1_m_000000_0
16/01/25 22:58:45 INFO output.FileOutputCommitter: File Output Committer Algorithm version
is 1
16/01/25 22:58:45 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is su
pported only on Linux.
16/01/25 22:58:45 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/25 22:58:45 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/venama
x/enronemail_1h.txt:0+203981
16/01/25 22:58:45 INFO mapred.MapTask: numReduceTasks: 1
16/01/25 22:58:45 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/25 22:58:45 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/25 22:58:45 INFO mapred.MapTask: soft limit at 83886080
16/01/25 22:58:45 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/25 22:58:45 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/25 22:58:45 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapr
ed.MapTask$MapOutputBuffer
16/01/25 22:58:46 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/venamax/ucb-261/hw
2/./mapper.py]
16/01/25 22:58:46 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, us
```

```
e mapreduce.task.id
16/01/25 22:58:46 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead,
use mapreduce.cluster.local.dir
16/01/25 22:58:46 INFO Configuration.deprecation: map.input.file is deprecated. Instead, u
se mapreduce.map.input.file
16/01/25 22:58:46 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, u
se mapreduce.job.skiprecords
16/01/25 22:58:46 INFO Configuration.deprecation: map.input.length is deprecated. Instead,
use mapreduce.map.input.length
16/01/25 22:58:46 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. In
stead, use mapreduce.task.output.dir
16/01/25 22:58:46 INFO Configuration.deprecation: map.input.start is deprecated. Instead,
use mapreduce.map.input.start
16/01/25 22:58:46 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, us
e mapreduce.job.id
16/01/25 22:58:46 INFO Configuration.deprecation: user.name is deprecated. Instead, use ma
preduce.job.user.name
16/01/25 22:58:46 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instea
d, use mapreduce.task.ismap
16/01/25 22:58:46 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, u
se mapreduce.task.attempt.id
16/01/25 22:58:46 INFO Configuration.deprecation: mapred.task.partition is deprecated. Ins
tead, use mapreduce.task.partition
16/01/25 22:58:46 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/25 22:58:46 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/25 22:58:46 INFO mapreduce.Job: Job job_local1108989492_0001 running in uber mode :
false
16/01/25 22:58:46 INFO mapreduce.Job:  map 0% reduce 0%
16/01/25 22:58:47 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/25 22:58:47 INFO streaming.PipeMapRed: MRErrorThread done
16/01/25 22:58:47 INFO streaming.PipeMapRed: Records R/W=101/1
16/01/25 22:58:47 INFO streaming.PipeMapRed: mapRedFinished
16/01/25 22:58:47 INFO mapred.LocalJobRunner:
16/01/25 22:58:47 INFO mapred.MapTask: Starting flush of map output
16/01/25 22:58:47 INFO mapred.MapTask: Spilling map output
16/01/25 22:58:47 INFO mapred.MapTask: bufstart = 0; bufend = 13; bufvoid = 104857600
16/01/25 22:58:47 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214396(104
857584); length = 1/6553600
16/01/25 22:58:47 INFO mapred.MapTask: Finished spill 0
16/01/25 22:58:47 INFO mapred.Task: Task:attempt_local1108989492_0001_m_000000_0 is done.
```

```
And is in the process of committing
16/01/25 22:58:47 INFO mapred.LocalJobRunner: Records R/W=101/1
16/01/25 22:58:47 INFO mapred.Task: Task 'attempt_local1108989492_0001_m_000000_0' done.
16/01/25 22:58:47 INFO mapred.LocalJobRunner: Finishing task: attempt_local1108989492_000
1_m_000000_0
16/01/25 22:58:47 INFO mapred.LocalJobRunner: map task executor complete.
16/01/25 22:58:47 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/25 22:58:47 INFO mapred.LocalJobRunner: Starting task: attempt_local1108989492_000
1_r_000000_0
16/01/25 22:58:47 INFO output.FileOutputCommitter: File Output Committer Algorithm version
is 1
16/01/25 22:58:47 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is su
pported only on Linux.
16/01/25 22:58:47 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/25 22:58:47 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.m
apreduce.task.reduce.Shuffle@45e6ddb1
16/01/25 22:58:47 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxS
ingleShuffleLimit=83584616, mergeThreshold=220663392, ioSortFactor=10, memToMemMergeOutput
sThreshold=10
16/01/25 22:58:47 INFO reduce.EventFetcher: attempt_local1108989492_0001_r_000000_0 Thread
started: EventFetcher for fetching Map Completion Events
16/01/25 22:58:47 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map
attempt_local1108989492_0001_m_000000_0 decomp: 17 len: 21 to MEMORY
16/01/25 22:58:47 INFO reduce.InMemoryMapOutput: Read 17 bytes from map-output for attemp
t_local1108989492_0001_m_000000_0
16/01/25 22:58:47 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1
7, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->17
16/01/25 22:58:47 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/25 22:58:47 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 22:58:47 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-out
puts and 0 on-disk map-outputs
16/01/25 22:58:47 INFO mapred.Merger: Merging 1 sorted segments
16/01/25 22:58:47 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of
total size: 4 bytes
16/01/25 22:58:47 INFO reduce.MergeManagerImpl: Merged 1 segments, 17 bytes to disk to sat
isfy reduce memory limit
16/01/25 22:58:47 INFO reduce.MergeManagerImpl: Merging 1 files, 21 bytes from disk
16/01/25 22:58:47 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory in
to reduce
16/01/25 22:58:47 INFO mapred.Merger: Merging 1 sorted segments
```

```
16/01/25 22:58:47 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of
total size: 4 bytes
16/01/25 22:58:47 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 22:58:47 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/venamax/ucb-261/hw
2/./reducer.py]
16/01/25 22:58:47 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instea
d, use mapreduce.jobtracker.address
16/01/25 22:58:47 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead,
use mapreduce.job.maps
16/01/25 22:58:47 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/25 22:58:47 INFO streaming.PipeMapRed: MRErrorThread done
16/01/25 22:58:47 INFO streaming.PipeMapRed: Records R/W=1/1
16/01/25 22:58:47 INFO streaming.PipeMapRed: mapRedFinished
16/01/25 22:58:47 INFO mapred.Task: Task:attempt_local1108989492_0001_r_000000_0 is done.
And is in the process of committing
16/01/25 22:58:47 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 22:58:47 INFO mapred.Task: Task attempt_local1108989492_0001_r_000000_0 is allowe
d to commit now
16/01/25 22:58:47 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1108
989492_0001_r_000000_0' to hdfs://localhost:9000/user/venamax/wordcountOutput/_temporar
y/0/task_local1108989492_0001_r_000000
16/01/25 22:58:47 INFO mapred.LocalJobRunner: Records R/W=1/1 > reduce
16/01/25 22:58:47 INFO mapred.Task: Task 'attempt_local1108989492_0001_r_000000_0' done.
16/01/25 22:58:47 INFO mapred.LocalJobRunner: Finishing task: attempt_local1108989492_000
1_r_000000_0
16/01/25 22:58:47 INFO mapred.LocalJobRunner: reduce task executor complete.
16/01/25 22:58:47 INFO mapreduce.Job:  map 100% reduce 100%
16/01/25 22:58:47 INFO mapreduce.Job: Job job_local1108989492_0001 completed successfully
16/01/25 22:58:47 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=211878
                FILE: Number of bytes written=800091
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=407962
                HDFS: Number of bytes written=41
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
```

```
        Map-Reduce Framework
                Map input records=101
                Map output records=1
                Map output bytes=13
                Map output materialized bytes=21
                Input split bytes=104
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=21
                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)=4
                Total committed heap usage (bytes)=549453824
        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=203981
        File Output Format Counters
                Bytes Written=41
16/01/25 22:58:47 INFO streaming.StreamJob: Output directory: wordcountOutput
```

show the results

In [414]:
```
!hdfs dfs -cat wordcountOutput/part-00000
```

```
16/01/25 22:59:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
We found "assistance"  on 9 occassions.
```

stop yarn and hdfs

In [16]:
```
!/usr/local/Cellar/hadoop/2.7.1/sbin/stop-yarn.sh
!/usr/local/Cellar/hadoop/2.7.1/sbin/stop-dfs.sh
```

```
stopping yarn daemons
stopping resourcemanager
localhost: stopping nodemanager
no proxyserver to stop
16/01/25 15:11:02 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
Stopping namenodes on [localhost]
localhost: stopping namenode
localhost: stopping datanode
Stopping secondary namenodes [0.0.0.0]
0.0.0.0: stopping secondarynamenode
16/01/25 15:11:24 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
```

========== END OF HW 2.2

HW2.2.1

Using Hadoop MapReduce and your wordcount job (from HW2.2) determine the top-10 occurring tokens (most frequent tokens)

Mapper

In [548]:
```
%%writefile mapper.py
#!/usr/bin/python
## mapper.py
## Author: Alejandro J. Rojas
## Description: mapper code for HW2.2

import sys
import re
from operator import itemgetter


counts = {}

for line in sys.stdin:                                ### input comes from STDIN (standard i
nput)
    record = re.split(r'\t+', line)
    if len(record) == 4:                              ### Take only complete records
        for i in range (2,len(record)):               ###  words in subject and content from
each email
            bagofwords = re.split(" ",record[i])      ### Break each email records into word
s
            for word in bagofwords:
                neword = word.strip(',')              ### eliminate comas and other clean up
                neword = re.sub('^\'+','',word)
                neword = re.sub('^\-+','',neword)
                neword = re.sub('\-+$','',neword)
                neword = re.sub('\'+','\'',neword)
                neword = re.sub('\-+','-',neword)
                if neword in counts.keys():            ### check if word has a prior count
                    counts[neword] +=1                 ### add one to the count
                else:                                  ### if not
                    counts[neword] = 1                 ### initialize word to one

for word in counts.keys():
    print ('%s\t%s'% (word, str(counts[word])))         ### word, count
```

```
Overwriting mapper.py
```

In [549]:
```
!chmod +x mapper.py
```

Reducer

In [550]:
```
%%writefile reducer.py
#!/usr/bin/python
from operator import itemgetter
import sys

current_word = None
current_count = 0
word = None
words = {}




# input comes from STDIN
for line in sys.stdin:

    line = line.strip()                         ### remove leading and trailing whitespace
    line = line.split('\t')                     ### parse the input we got from mappe
    if len(line) == 2:
        word = line[0]

        try:

            count = line[1]
            count = int(count)                      ### convert count (currently a string)
to int
        except ValueError:                          ### if count was not a number then sile
ntly
            continue                                ### ignore/discard this line



        if current_word == word:                     ### this IF-switch only works because
Hadoop sorts map output
            current_count += count                  ### by key (here: number) before it is
passed to the reducer
        else:
            if current_word:

                words[current_word] = current_count  ### store tuple in a list once totaliz
```

```
e count per number

            current_count = count                    ### set current count
            current_word = word                      ### set current word


if current_word == word:                        ### do not forget to output the last word
if needed!
    words[current_word] = current_count



topwords = sorted(words, key=words.__getitem__, reverse = True) ## sort
topwordcount = [value for (key, value) in sorted(words.items(), reverse=True)]

for i in range(10):
    word = topwords[i]
    print 'We found "%s"' %word, ' on %s'%words[word] , 'occassions.'
```

```
Overwriting reducer.py
```

In [551]:
```
!chmod +x reducer.py
```

remove files from prior runs

In [552]:
```
!hdfs dfs -rm -r /user/venamax                          ### remove prior files
```

```
16/01/26 00:57:29 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
16/01/26 00:57:29 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion inter
val = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/venamax
```

create folder

In [553]:
```
!hdfs dfs -mkdir -p /user/venamax                          ### create hdfs folder
```

```
16/01/26 00:57:32 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
```

upload enronemail_1h.txt to hdfs

In [554]:
```
!hdfs dfs -put enronemail_1h.txt /user/venamax            #### save source data file to hdf
s
```

```
16/01/26 00:57:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
```

Hadoop streaming command

In [555]:
```
!hadoop jar hadoop-*streaming*.jar -mapper mapper.py -reducer reducer.py -input enronemai
l_1h.txt -output wordtopcountOutput
```

```
16/01/26 00:57:40 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
16/01/26 00:57:42 INFO Configuration.deprecation: session.id is deprecated. Instead, use d
fs.metrics.session-id
16/01/26 00:57:42 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracke
r, sessionId=
16/01/26 00:57:42 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobT
racker, sessionId= - already initialized
16/01/26 00:57:42 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/26 00:57:42 INFO mapreduce.JobSubmitter: number of splits:1
16/01/26 00:57:42 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1864178
331_0001
16/01/26 00:57:42 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/26 00:57:42 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/26 00:57:42 INFO mapreduce.Job: Running job: job_local1864178331_0001
16/01/26 00:57:42 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapre
d.FileOutputCommitter
16/01/26 00:57:42 INFO output.FileOutputCommitter: File Output Committer Algorithm version
is 1
16/01/26 00:57:42 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/26 00:57:42 INFO mapred.LocalJobRunner: Starting task: attempt_local1864178331_000
1_m_000000_0
16/01/26 00:57:42 INFO output.FileOutputCommitter: File Output Committer Algorithm version
is 1
16/01/26 00:57:42 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is su
pported only on Linux.
16/01/26 00:57:42 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 00:57:42 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/venama
x/enronemail_1h.txt:0+203981
16/01/26 00:57:42 INFO mapred.MapTask: numReduceTasks: 1
16/01/26 00:57:43 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/26 00:57:43 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/26 00:57:43 INFO mapred.MapTask: soft limit at 83886080
16/01/26 00:57:43 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/26 00:57:43 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/26 00:57:43 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapr
ed.MapTask$MapOutputBuffer
16/01/26 00:57:43 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/venamax/ucb-261/hw
2/./mapper.py]
16/01/26 00:57:43 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, us
```

```
e mapreduce.task.id
16/01/26 00:57:43 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead,
use mapreduce.cluster.local.dir
16/01/26 00:57:43 INFO Configuration.deprecation: map.input.file is deprecated. Instead, u
se mapreduce.map.input.file
16/01/26 00:57:43 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, u
se mapreduce.job.skiprecords
16/01/26 00:57:43 INFO Configuration.deprecation: map.input.length is deprecated. Instead,
use mapreduce.map.input.length
16/01/26 00:57:43 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. In
stead, use mapreduce.task.output.dir
16/01/26 00:57:43 INFO Configuration.deprecation: map.input.start is deprecated. Instead,
use mapreduce.map.input.start
16/01/26 00:57:43 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, us
e mapreduce.job.id
16/01/26 00:57:43 INFO Configuration.deprecation: user.name is deprecated. Instead, use ma
preduce.job.user.name
16/01/26 00:57:43 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instea
d, use mapreduce.task.ismap
16/01/26 00:57:43 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, u
se mapreduce.task.attempt.id
16/01/26 00:57:43 INFO Configuration.deprecation: mapred.task.partition is deprecated. Ins
tead, use mapreduce.task.partition
16/01/26 00:57:43 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 00:57:43 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 00:57:43 INFO mapreduce.Job: Job job_local1864178331_0001 running in uber mode :
false
16/01/26 00:57:43 INFO mapreduce.Job:  map 0% reduce 0%
16/01/26 00:57:44 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:100=100/1 [rec/s] out:0=0/1
[rec/s]
16/01/26 00:57:48 INFO streaming.PipeMapRed: Records R/W=101/1
16/01/26 00:57:48 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 00:57:48 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 00:57:48 INFO mapred.LocalJobRunner:
16/01/26 00:57:48 INFO mapred.MapTask: Starting flush of map output
16/01/26 00:57:48 INFO mapred.MapTask: Spilling map output
16/01/26 00:57:48 INFO mapred.MapTask: bufstart = 0; bufend = 75391; bufvoid = 104857600
16/01/26 00:57:48 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26184528(104
738112); length = 29869/6553600
16/01/26 00:57:48 INFO mapred.MapTask: Finished spill 0
```

```
16/01/26 00:57:48 INFO mapred.Task: Task:attempt_local1864178331_0001_m_000000_0 is done.
And is in the process of committing
16/01/26 00:57:48 INFO mapred.LocalJobRunner: Records R/W=101/1
16/01/26 00:57:48 INFO mapred.Task: Task 'attempt_local1864178331_0001_m_000000_0' done.
16/01/26 00:57:48 INFO mapred.LocalJobRunner: Finishing task: attempt_local1864178331_000
1_m_000000_0
16/01/26 00:57:48 INFO mapred.LocalJobRunner: map task executor complete.
16/01/26 00:57:48 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/26 00:57:48 INFO mapred.LocalJobRunner: Starting task: attempt_local1864178331_000
1_r_000000_0
16/01/26 00:57:48 INFO output.FileOutputCommitter: File Output Committer Algorithm version
is 1
16/01/26 00:57:48 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is su
pported only on Linux.
16/01/26 00:57:48 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 00:57:48 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.m
apreduce.task.reduce.Shuffle@5150e88d
16/01/26 00:57:48 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxS
ingleShuffleLimit=83584616, mergeThreshold=220663392, ioSortFactor=10, memToMemMergeOutput
sThreshold=10
16/01/26 00:57:48 INFO reduce.EventFetcher: attempt_local1864178331_0001_r_000000_0 Thread
started: EventFetcher for fetching Map Completion Events
16/01/26 00:57:48 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map
attempt_local1864178331_0001_m_000000_0 decomp: 90329 len: 90333 to MEMORY
16/01/26 00:57:48 INFO reduce.InMemoryMapOutput: Read 90329 bytes from map-output for atte
mpt_local1864178331_0001_m_000000_0
16/01/26 00:57:48 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 9
0329, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->90329
16/01/26 00:57:48 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/26 00:57:48 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 00:57:48 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-out
puts and 0 on-disk map-outputs
16/01/26 00:57:48 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 00:57:48 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of
total size: 90326 bytes
16/01/26 00:57:48 INFO reduce.MergeManagerImpl: Merged 1 segments, 90329 bytes to disk to
satisfy reduce memory limit
16/01/26 00:57:48 INFO reduce.MergeManagerImpl: Merging 1 files, 90333 bytes from disk
16/01/26 00:57:48 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory in
to reduce
```

```
16/01/26 00:57:48 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 00:57:48 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of
total size: 90326 bytes
16/01/26 00:57:48 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 00:57:48 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/venamax/ucb-261/hw
2/./reducer.py]
16/01/26 00:57:48 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instea
d, use mapreduce.jobtracker.address
16/01/26 00:57:48 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead,
use mapreduce.job.maps
16/01/26 00:57:48 INFO mapreduce.Job:  map 100% reduce 0%
16/01/26 00:57:48 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 00:57:48 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 00:57:48 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 00:57:48 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 00:57:48 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 00:57:48 INFO streaming.PipeMapRed: Records R/W=7468/1
16/01/26 00:57:48 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 00:57:49 INFO mapred.Task: Task:attempt_local1864178331_0001_r_000000_0 is done.
And is in the process of committing
16/01/26 00:57:49 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 00:57:49 INFO mapred.Task: Task attempt_local1864178331_0001_r_000000_0 is allowe
d to commit now
16/01/26 00:57:49 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1864
178331_0001_r_000000_0' to hdfs://localhost:9000/user/venamax/wordtopcountOutput/_temporar
y/0/task_local1864178331_0001_r_000000
16/01/26 00:57:49 INFO mapred.LocalJobRunner: Records R/W=7468/1 > reduce
16/01/26 00:57:49 INFO mapred.Task: Task 'attempt_local1864178331_0001_r_000000_0' done.
16/01/26 00:57:49 INFO mapred.LocalJobRunner: Finishing task: attempt_local1864178331_000
1_r_000000_0
16/01/26 00:57:49 INFO mapred.LocalJobRunner: reduce task executor complete.
16/01/26 00:57:49 INFO mapreduce.Job:  map 100% reduce 100%
16/01/26 00:57:49 INFO mapreduce.Job: Job job_local1864178331_0001 completed successfully
16/01/26 00:57:49 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=392502
                FILE: Number of bytes written=1071039
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
```

```
                HDFS: Number of bytes read=407962
                HDFS: Number of bytes written=355
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=101
                Map output records=7468
                Map output bytes=75391
                Map output materialized bytes=90333
                Input split bytes=104
                Combine input records=0
                Combine output records=0
                Reduce input groups=7378
                Reduce shuffle bytes=90333
                Reduce input records=7468
                Reduce output records=10
                Spilled Records=14936
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=6
                Total committed heap usage (bytes)=547356672
        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=203981
        File Output Format Counters
                Bytes Written=355
16/01/26 00:57:49 INFO streaming.StreamJob: Output directory: wordtopcountOutput
```

show the results

In [556]:
```
!hdfs dfs -cat wordtopcountOutput/part-00000
```

```
16/01/26 00:58:04 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
We found "the"  on 1201 occassions.
We found "to"  on 888 occassions.
We found "and"  on 607 occassions.
We found "of"  on 534 occassions.
We found "a"  on 506 occassions.
We found "in"  on 399 occassions.
We found "your"  on 379 occassions.
We found "you"  on 375 occassions.
We found "@"  on 360 occassions.
We found "for"  on 353 occassions.
```

========== END OF HW 2.2.1

HW2.3. Multinomial NAIVE BAYES with NO Smoothing

Using the Enron data from HW1 and Hadoop MapReduce, write a mapper/reducer job(s) that will both learn Naive Bayes classifier and classify the Enron email messages using the learnt Naive Bayes classifier. Use all white-space delimitted tokens as independent input variables (assume spaces, fullstops, commas as delimiters). Note: for multinomial Naive Bayes, the Pr(X="assistance"|Y=SPAM) is calculated as follows:

the number of times "assistance" occurs in SPAM labeled documents / the number of words in documents labeled SPAM

E.g., "assistance" occurs 5 times in all of the documents Labeled SPAM, and the length in terms of the number of words in all documents labeled as SPAM (when concatenated) is 1,000. Then Pr(X="assistance"|Y=SPAM) = 5/1000. Note this is a multinomial estimation of the class conditional for a Naive Bayes Classifier. No smoothing is needed in this HW. Multiplying lots of probabilities, which are between 0 and 1, can result in floating-point underflow. Since log(xy) = log(x) + log(y), it is better to perform all computations by summing logs of probabilities rather than multiplying probabilities. Please pay attention to probabilites that are zero! They will need special attention. Count up how many times you need to process a zero probabilty for each class and report.

Report the performance of your learnt classifier in terms of misclassifcation error rate of your multinomial Naive Bayes Classifier. Plot a histogram of the log posterior probabilities (i.e., Pr(Class|Doc))) for each class over the training set. Summarize what you see.

Error Rate = misclassification rate with respect to a provided set (say training set in this case). It is more formally defined here:

Let DF represent the evalution set in the following: Err(Model, DF) = |{(X, c(X)) ∈ DF : c(X) != Model(x)}| / |DF|

Where || denotes set cardinality; c(X) denotes the class of the tuple X in DF; and Model(X) denotes the class inferred by the Model "Model"

Mapper

In [630]:
```
%%writefile mapper.py
#!/usr/bin/python
## mapper.py
## Author: Alejandro J. Rojas
## Description: mapper code for HW2.2

import sys
import re
from operator import itemgetter


counts = {}

for line in sys.stdin:                                  ### input comes from STDIN (standard i
nput)
    record = re.split(r'\t+', line)
    if len(record) == 4:                                ### Take only complete records
        for i in range (2,len(record)):                 ###  words in subject and content from
each email
            bagofwords = re.split(" ",record[i])        ### Break each email records into word
s
            for word in bagofwords:
                neword = word.strip(',')                ### eliminate comas and other clean up
                neword = re.sub('^\'+','',word)
                neword = re.sub('^\-+','',neword)
                neword = re.sub('\-+$','',neword)
                neword = re.sub('\!+$','',neword)
                neword = re.sub('^\!+','',neword)
                neword = re.sub('\'+','\'',neword)
                neword = re.sub('\-+','-',neword)
                print '%s\t%s\t%s\t%s' % (neword, 1,record[0], record[1])
                                                        ### output: word, 1, id, spam truth
```

```
Overwriting mapper.py
```

In [631]:
```
!chmod +x mapper.py
```

Reducer

In [909]:
```
%%writefile reducer.py
#!/usr/bin/python

import numpy as np
from operator import itemgetter
import sys
from itertools import groupby
import math


current_word, word = None, None
current_email, email = None, None
current_y_true, y_true = None, None


sum_records, sum_spamrecords, sum_hamrecords = 0,0,0
sum_spamwords, sum_hamwords = 0,0
spam_zeroes, ham_zeroes = 0,0


emails={} #Associative array to hold email data
words={} #Associative array for word data
conf_matrix = np.zeros((2,2))

# input comes from STDIN
for line in sys.stdin:

    line = line.strip()                        ### remove leading and trailing whitespace
    line = line.split('\t')                    ### parse the input we got from mapper.py
    if len(line) == 4:

        try:
            word = line[0]                          ### word we get from mapper.py
            count = line[1]                         ### corresponding count
            count = int(count)                      ### convert count (currently a string)
to int
            email = line[2]                         ### id that identifies each email
            y_true = line[3]                        ### spam truth
            y_true = int(y_true)                    ### spam truth as an integer
        except ValueError:                          ### any error then silently
```

```
            continue                                ### ignore/discard this line



        if current_word == word:                    ### this IF-switch only works because H
adoop sorts map output
            current_count += count                  ### by key (here: word) before it is pa
ssed to the reducer

            if current_word not in words.keys():     ### initialize word count of new words
                words[current_word]={'ham_count':0,'spam_count':0}

            if current_email not in emails.keys():          ### initialize word counts and
label of new email
                emails[current_email]={'y_true':current_y_true,'word_count':0,'words':[]}
                sum_records +=1.0
                if current_y_true == 1:                     ### identified new email as eit
her spam or ham
                    sum_spamrecords +=1.0

            elif current_email in emails.keys():
                emails[current_email]['word_count'] += 1
                emails[current_email]['words'].append(current_word)### store words in email

            if current_y_true == 1:                         ### if record where word is loc
ated is a spam
                current_spamcount += count         ### add to spam count of that word
                sum_spamwords += 1.0
            else:
                current_hamcount += count          ### if not add to ham count of thet word
                sum_hamwords +=1.0

        else:
            if current_word:
                try:
                    if current_word in words.keys():
                        words[current_word]['spam_count'] += current_spamcount ### update s
pam count for current word
                        words[current_word]['ham_count'] += current_hamcount ### update ham
count for current word
```

```
                except ValueError:                          ### if count was not a number t
hen silently
                    continue                                ### ignore/discard this line


            current_count = count                   ### set current count
            current_word = word                     ### set current number
            current_email = email                   ### set current id of email
            current_y_true = y_true                 ### set current spam truth
            current_spamcount, current_hamcount = 0,0



if current_word == word:                        ### do not forget to output the last word if
needed!
    emails[current_email]['word_count'] += 1
    emails[current_email]['words'].append(current_word)### store words in email
    words[current_word]['spam_count'] += current_spamcount ### update spam count for curren
t word
    words[current_word]['ham_count'] += current_hamcount ### update ham count for current w
ord



#Calculate stats for entire corpus

prior_spam= sum_spamrecords/sum_records
prior_ham= 1 - prior_spam
vocab_count=len(words)#number of unique words in the total vocabulary
prob_spamwords = sum_spamwords/(sum_spamwords+sum_hamwords)
prob_hamwords = 1-prob_spamwords



for k,word in words.iteritems():
    #These versions calculate conditional probabilities WITH Laplace smoothing.
    #word['p_spam']=(word['spam_count']+1)/(spam_word_count+vocab_count)
    #word['p_ham']=(word['ham_count']+1)/(ham_word_count+vocab_count)

    #Compute conditional probabilities WITHOUT Laplace smoothing
```

```
    word['p_spam']=(word['spam_count'])/(sum_spamwords)
    word['p_ham']=(word['ham_count'])/(sum_hamwords)

    if word['p_spam'] == 0 and word['p_ham'] == 0:
        spam_zeroes +=1
        ham_zeroes +=1
        word['p_spam']= 1
        word['p_ham'] = 1

    elif word['p_spam'] == 0:
        spam_zeroes +=1
        word['p_ham'] = min(0.5 ** (1/word['ham_count']),0.99999999) ### back-calculates pr
obability
        word['p_spam']= 1-word['p_ham']             ### of the sequence of events assuming
                                                    ### sequence is 50% likely to happen
    elif word['p_ham'] == 0:
        ham_zeroes +=1
        word['p_spam'] = min(0.5 ** (1/word['spam_count']),0.99999999) ### Adjust probabili
ty
        word['p_ham']= 1-word['p_spam']              ### just as above

    word['p_spam'] = math.log(word['p_spam'])             ### convert to log scale
    word['p_ham'] = math.log(word['p_ham'])

#At this point the model is now trained, and we can use it to make our predictions

print 'Analyzed %s' % sum_records, 'using a vocabulary of %s' % vocab_count , 'terms'
print 'On %s'%spam_zeroes , 'times we found spam probabilities that were zero'
print 'On %s'%ham_zeroes, 'times we found not spam probabilities that were zero'
print '%30s' %'ID', '%10s' %'TRUTH', '%10s' %'CLASS', '%20s' %'CUMULATIVE ACCURACY', '%10
s'%'P-SPAM', '%10s'%'P-HAM'
miss, sample_size = 0,0
prob_spam_list, prob_ham_list = [], []
for j,email in emails.iteritems():

    #Log versions - no longer used
    p_spam= math.log(prior_spam)
    p_ham= math.log(prior_ham)

    #p_spam=prior_spam
```

```
    #p_ham=prior_ham

    for word in email['words']:

        try:
            #p_spam+=log(words[word]['p_spam']) #Log version - no longer used
            p_spam+=words[word]['p_spam']
        except ValueError:
            pass #This means that words that do not appear in a class will use the class pr
ior
        try:
            #p_ham+=log(words[word]['p_ham']) #Log version - no longer used
            p_ham+=words[word]['p_ham']
        except ValueError:
            pass


    if p_spam>p_ham:
        y_pred=1

    else:
        y_pred=0
    y_true = email['y_true']
    if y_true == 1:
        prob_spam_list.append(math.exp(p_spam))       ### update probability list
        if y_pred == 1:
            conf_matrix[0][0] +=1                     ### update confusion matrix
        else:
            conf_matrix[0][1] +=1

    else:
        prob_ham_list.append(math.exp(p_spam))
        if y_pred == 0:
            conf_matrix[1][1] +=1
        else:
            conf_matrix[1][0] +=1


    if y_pred != y_true:
        miss+= 1.0
```

```
    sample_size += 1.0
    accuracy = ((sample_size-miss)/sample_size)*100
    prob_spam = math.exp(p_spam)
    prob_ham = math.exp(p_ham)
    print  '%30s' %j, '%10d' %y_true, '%10d' %y_pred, '%18.2f %%' % accuracy, '%10.6f'%pro
b_spam, '%10.6f'%prob_ham

print 'See confusion matrix. Top Left is predicted correctly as spam. Bottom right predicte
d correctly as ham:'
print conf_matrix
print 'Error rate: %4.2f'%(100-accuracy),'%%'
```

```
Overwriting reducer.py
```

In [910]:
```
!chmod +x reducer.py
```

Run Hadoop MapReduce Streaming

start up yarn and dfs

In [ ]:
```
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-yarn.sh       ### start up yarn
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-dfs.sh        ### start up dfs
```

remove files from prior runs

In [916]:
```
!hdfs dfs -rm -r /user/venamax                          ### remove prior files
```

```
16/01/26 18:13:49 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
16/01/26 18:13:49 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion inter
val = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/venamax
```

create folder

In [917]:
```
!hdfs dfs -mkdir -p /user/venamax                       ### create hdfs folder
```

```
16/01/26 18:13:53 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
```

upload enronmail_1h.txt file

In [918]:
```
!hdfs dfs -put enronemail_1h.txt /user/venamax                #### save source data file t
o hdfs
```

```
16/01/26 18:13:56 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
```

Hadoop streaming

In [919]:
```
!hadoop jar hadoop-*streaming*.jar -mapper mapper.py -reducer reducer.py -input enronemai
l_1h.txt -output NBNoSmoothtOutput
```

```
16/01/26 18:14:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
16/01/26 18:14:06 INFO Configuration.deprecation: session.id is deprecated. Instead, use d
fs.metrics.session-id
16/01/26 18:14:06 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracke
r, sessionId=
16/01/26 18:14:06 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobT
racker, sessionId= - already initialized
16/01/26 18:14:06 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/26 18:14:06 INFO mapreduce.JobSubmitter: number of splits:1
16/01/26 18:14:07 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local2264993
16_0001
16/01/26 18:14:07 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/26 18:14:07 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/26 18:14:07 INFO mapreduce.Job: Running job: job_local226499316_0001
16/01/26 18:14:07 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapre
d.FileOutputCommitter
16/01/26 18:14:07 INFO output.FileOutputCommitter: File Output Committer Algorithm version
is 1
16/01/26 18:14:07 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/26 18:14:07 INFO mapred.LocalJobRunner: Starting task: attempt_local226499316_000
1_m_000000_0
16/01/26 18:14:07 INFO output.FileOutputCommitter: File Output Committer Algorithm version
is 1
16/01/26 18:14:07 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is su
pported only on Linux.
16/01/26 18:14:07 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 18:14:07 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/venama
x/enronemail_1h.txt:0+203981
16/01/26 18:14:07 INFO mapred.MapTask: numReduceTasks: 1
16/01/26 18:14:07 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/26 18:14:07 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/26 18:14:07 INFO mapred.MapTask: soft limit at 83886080
16/01/26 18:14:07 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/26 18:14:07 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/26 18:14:07 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapr
ed.MapTask$MapOutputBuffer
16/01/26 18:14:07 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/venamax/ucb-261/hw
2/./mapper.py]
16/01/26 18:14:07 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, us
```

```
e mapreduce.task.id
16/01/26 18:14:07 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead,
use mapreduce.cluster.local.dir
16/01/26 18:14:07 INFO Configuration.deprecation: map.input.file is deprecated. Instead, u
se mapreduce.map.input.file
16/01/26 18:14:07 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, u
se mapreduce.job.skiprecords
16/01/26 18:14:07 INFO Configuration.deprecation: map.input.length is deprecated. Instead,
use mapreduce.map.input.length
16/01/26 18:14:07 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. In
stead, use mapreduce.task.output.dir
16/01/26 18:14:07 INFO Configuration.deprecation: map.input.start is deprecated. Instead,
use mapreduce.map.input.start
16/01/26 18:14:07 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, us
e mapreduce.job.id
16/01/26 18:14:07 INFO Configuration.deprecation: user.name is deprecated. Instead, use ma
preduce.job.user.name
16/01/26 18:14:07 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instea
d, use mapreduce.task.ismap
16/01/26 18:14:07 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, u
se mapreduce.task.attempt.id
16/01/26 18:14:07 INFO Configuration.deprecation: mapred.task.partition is deprecated. Ins
tead, use mapreduce.task.partition
16/01/26 18:14:07 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:14:07 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:14:07 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:14:07 INFO streaming.PipeMapRed: Records R/W=101/1
16/01/26 18:14:08 INFO mapreduce.Job: Job job_local226499316_0001 running in uber mode : f
alse
16/01/26 18:14:08 INFO mapreduce.Job:  map 0% reduce 0%
16/01/26 18:14:08 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 18:14:08 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 18:14:08 INFO mapred.LocalJobRunner:
16/01/26 18:14:08 INFO mapred.MapTask: Starting flush of map output
16/01/26 18:14:08 INFO mapred.MapTask: Spilling map output
16/01/26 18:14:08 INFO mapred.MapTask: bufstart = 0; bufend = 1172828; bufvoid = 104857600
16/01/26 18:14:08 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26072600(104
290400); length = 141797/6553600
16/01/26 18:14:08 INFO mapred.MapTask: Finished spill 0
16/01/26 18:14:08 INFO mapred.Task: Task:attempt_local226499316_0001_m_000000_0 is done. A
```

```
nd is in the process of committing
16/01/26 18:14:08 INFO mapred.LocalJobRunner: Records R/W=101/1
16/01/26 18:14:08 INFO mapred.Task: Task 'attempt_local226499316_0001_m_000000_0' done.
16/01/26 18:14:08 INFO mapred.LocalJobRunner: Finishing task: attempt_local226499316_000
1_m_000000_0
16/01/26 18:14:08 INFO mapred.LocalJobRunner: map task executor complete.
16/01/26 18:14:08 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/26 18:14:08 INFO mapred.LocalJobRunner: Starting task: attempt_local226499316_000
1_r_000000_0
16/01/26 18:14:08 INFO output.FileOutputCommitter: File Output Committer Algorithm version
is 1
16/01/26 18:14:08 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is su
pported only on Linux.
16/01/26 18:14:08 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 18:14:08 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.m
apreduce.task.reduce.Shuffle@1b9cd035
16/01/26 18:14:08 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxS
ingleShuffleLimit=83584616, mergeThreshold=220663392, ioSortFactor=10, memToMemMergeOutput
sThreshold=10
16/01/26 18:14:08 INFO reduce.EventFetcher: attempt_local226499316_0001_r_000000_0 Thread
started: EventFetcher for fetching Map Completion Events
16/01/26 18:14:08 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map
attempt_local226499316_0001_m_000000_0 decomp: 1243730 len: 1243734 to MEMORY
16/01/26 18:14:08 INFO reduce.InMemoryMapOutput: Read 1243730 bytes from map-output for at
tempt_local226499316_0001_m_000000_0
16/01/26 18:14:08 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1
243730, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->1243730
16/01/26 18:14:08 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/26 18:14:08 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 18:14:08 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-out
puts and 0 on-disk map-outputs
16/01/26 18:14:08 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 18:14:08 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of
total size: 1243727 bytes
16/01/26 18:14:08 INFO reduce.MergeManagerImpl: Merged 1 segments, 1243730 bytes to disk t
o satisfy reduce memory limit
16/01/26 18:14:08 INFO reduce.MergeManagerImpl: Merging 1 files, 1243734 bytes from disk
16/01/26 18:14:08 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory in
to reduce
16/01/26 18:14:08 INFO mapred.Merger: Merging 1 sorted segments
```

```
16/01/26 18:14:08 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of
total size: 1243727 bytes
16/01/26 18:14:08 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 18:14:08 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/venamax/ucb-261/hw
2/./reducer.py]
16/01/26 18:14:08 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instea
d, use mapreduce.jobtracker.address
16/01/26 18:14:08 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead,
use mapreduce.job.maps
16/01/26 18:14:08 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:14:08 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:14:08 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:14:08 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:14:09 INFO mapreduce.Job:  map 100% reduce 0%
16/01/26 18:14:09 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:14:10 INFO streaming.PipeMapRed: Records R/W=35450/1
16/01/26 18:14:10 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 18:14:10 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 18:14:10 INFO mapred.Task: Task:attempt_local226499316_0001_r_000000_0 is done. A
nd is in the process of committing
16/01/26 18:14:10 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 18:14:10 INFO mapred.Task: Task attempt_local226499316_0001_r_000000_0 is allowed
to commit now
16/01/26 18:14:10 INFO output.FileOutputCommitter: Saved output of task 'attempt_local2264
99316_0001_r_000000_0' to hdfs://localhost:9000/user/venamax/NBNoSmoothtOutput/_temporar
y/0/task_local226499316_0001_r_000000
16/01/26 18:14:10 INFO mapred.LocalJobRunner: Records R/W=35450/1 > reduce
16/01/26 18:14:10 INFO mapred.Task: Task 'attempt_local226499316_0001_r_000000_0' done.
16/01/26 18:14:10 INFO mapred.LocalJobRunner: Finishing task: attempt_local226499316_000
1_r_000000_0
16/01/26 18:14:10 INFO mapred.LocalJobRunner: reduce task executor complete.
16/01/26 18:14:11 INFO mapreduce.Job:  map 100% reduce 100%
16/01/26 18:14:11 INFO mapreduce.Job: Job job_local226499316_0001 completed successfully
16/01/26 18:14:11 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=2699304
                FILE: Number of bytes written=4528242
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
```

```
                HDFS: Number of bytes read=407962
                HDFS: Number of bytes written=9732
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=101
                Map output records=35450
                Map output bytes=1172828
                Map output materialized bytes=1243734
                Input split bytes=104
                Combine input records=0
                Combine output records=0
                Reduce input groups=7275
                Reduce shuffle bytes=1243734
                Reduce input records=35450
                Reduce output records=104
                Spilled Records=70900
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=4
                Total committed heap usage (bytes)=545259520
        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=203981
        File Output Format Counters
                Bytes Written=9732
16/01/26 18:14:11 INFO streaming.StreamJob: Output directory: NBNoSmoothtOutput
```

show the results

In [920]:
```
!hdfs dfs -cat NBNoSmoothtOutput/part-00000
```

```
16/01/26 18:14:17 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
Analyzed 96.0 using a vocabulary of 3077 terms
On 1330 times we found spam probabilities that were zero
On 1747 times we found not spam probabilities that were zero
                           ID      TRUTH      CLASS  CUMULATIVE ACCURACY     P-SPAM
P-HAM
            0010.2003-12-18.GP          1          0               0.00 %   0.447917   0.5
52083
     0010.2001-06-28.SA_and_HP          1          1              50.00 %   0.000000   0.0
00000
          0001.2000-01-17.beck          0          0              66.67 %   0.000000   0.0
00017
      0018.1999-12-14.kaminski          0          0              75.00 %   0.000000   0.2
76041
      0005.1999-12-12.kaminski          0          0              80.00 %   0.000000   0.1
38021
     0011.2001-06-29.SA_and_HP          1          1              83.33 %   0.000000   0.0
00000
            0008.2004-08-01.BG          1          1              85.71 %   0.000000   0.0
00000
        0009.1999-12-14.farmer          0          0              87.50 %   0.000000   0.0
34505
            0017.2003-12-18.GP          1          1              88.89 %   0.447917   0.0
00000
     0011.2001-06-28.SA_and_HP          1          1              90.00 %   0.000000   0.0
00000
     0015.2001-07-05.SA_and_HP          1          1              90.91 %   0.001750   0.0
00000
       0015.2001-02-12.kitchen          0          0              91.67 %   0.000000   0.0
00000
      0017.1999-12-14.kaminski          0          0              92.31 %   0.000000   0.2
76042
          0012.2000-01-17.beck          0          0              92.86 %   0.000000   0.5
52081
          0003.2000-01-17.beck          0          0              93.33 %   0.000000   0.0
08626
     0004.2001-06-12.SA_and_HP          1          1              93.75 %   0.013997   0.0
00000
     0008.2001-06-12.SA_and_HP          1          1              94.12 %   0.006999   0.0
```

```
00000
       0007.2001-02-09.kitchen          0          0               94.44 %   0.000000   0.0
02157
            0016.2004-08-01.BG          1          1               94.74 %   0.055990   0.0
00000
        0016.1999-12-15.farmer          0          0               95.00 %   0.000000   0.0
34505
            0013.2004-08-01.BG          1          1               95.24 %   0.003499   0.0
00000
            0005.2003-12-18.GP          1          1               95.45 %   0.000000   0.0
00000
       0012.2001-02-09.kitchen          0          0               95.65 %   0.000000   0.1
38021
        0011.1999-12-14.farmer          0          0               95.83 %   0.000000   0.0
04313
       0009.2001-02-09.kitchen          0          0               96.00 %   0.000000   0.0
00000
       0006.2001-02-08.kitchen          0          0               96.15 %   0.000000   0.0
00000
            0014.2003-12-19.GP          1          0               92.59 %   0.223958   0.2
76042
        0010.1999-12-14.farmer          0          0               92.86 %   0.000000   0.0
04313
            0010.2004-08-01.BG          1          1               93.10 %   0.000014   0.0
00000
      0014.1999-12-14.kaminski          0          0               93.33 %   0.000000   0.0
04313
      0006.1999-12-13.kaminski          0          0               93.55 %   0.000000   0.0
34505
        0005.1999-12-14.farmer          0          0               93.75 %   0.000000   0.0
08626
       0003.2001-02-08.kitchen          0          0               93.94 %   0.000000   0.0
02157
       0001.2001-02-07.kitchen          0          0               94.12 %   0.000000   0.0
08626
       0008.2001-02-09.kitchen          0          0               94.29 %   0.000000   0.0
00000
            0017.2004-08-02.BG          1          1               94.44 %   0.000001   0.0
00000
            0007.2003-12-18.GP          1          1               94.59 %   0.001750   0.0
```

```
00000
            0014.2004-08-01.BG          1          1               94.74 %   0.027995   0.0
00000
            0006.2003-12-18.GP          1          1               94.87 %   0.003499   0.0
00000
     0016.2001-07-05.SA_and_HP          1          1               95.00 %   0.001750   0.0
00000
         0005.2000-06-06.lokay          0          0               95.12 %   0.000000   0.1
38021
     0014.2001-07-04.SA_and_HP          1          1               95.24 %   0.000014   0.0
00000
      0001.2001-04-02.williams          0          0               95.35 %   0.000000   0.0
01078
         0012.2000-06-08.lokay          0          0               95.45 %   0.000000   0.0
69010
        0014.1999-12-15.farmer          0          0               95.56 %   0.000000   0.0
02157
         0009.2000-06-07.lokay          0          0               95.65 %   0.000000   0.0
00270
        0001.1999-12-10.farmer          0          0               95.74 %   0.447917   0.5
52083
     0008.2001-06-25.SA_and_HP          1          1               95.83 %   0.000219   0.0
00000
      0017.2001-04-03.williams          0          0               95.92 %   0.000000   0.2
76042
       0014.2001-02-12.kitchen          0          0               96.00 %   0.000000   0.0
01078
     0016.2001-07-06.SA_and_HP          1          1               96.08 %   0.000000   0.0
00000
        0015.1999-12-15.farmer          0          0               96.15 %   0.000000   0.0
00017
      0004.1999-12-10.kaminski          0          0               96.23 %   0.000000   0.2
76042
            0011.2004-08-01.BG          1          1               96.30 %   0.055990   0.0
00000
            0004.2004-08-01.BG          1          1               96.36 %   0.111979   0.0
00000
            0018.2003-12-18.GP          1          1               96.43 %   0.000001   0.0
00000
        0007.1999-12-14.farmer          0          0               96.49 %   0.000000   0.0
```

```
01078
            0016.2003-12-19.GP          1          1               96.55 %   0.000000   0.0
00000
        0004.1999-12-14.farmer          0          0               96.61 %   0.000000   0.0
17253
            0015.2003-12-19.GP          1          1               96.67 %   0.001750   0.0
00000
            0006.2004-08-01.BG          1          1               96.72 %   0.013997   0.0
00000
            0009.2003-12-18.GP          1          1               96.77 %   0.000000   0.0
00000
        0002.1999-12-13.farmer          0          0               96.83 %   0.000000   0.0
00270
            0008.2003-12-18.GP          1          1               96.88 %   0.027995   0.0
00000
      0010.1999-12-14.kaminski          0          0               96.92 %   0.447917   0.5
52083
          0007.2000-01-17.beck          0          0               96.97 %   0.000000   0.5
52082
        0003.1999-12-14.farmer          0          0               97.01 %   0.000000   0.5
52083
            0003.2004-08-01.BG          1          1               97.06 %   0.006999   0.0
00000
            0017.2004-08-01.BG          1          1               97.10 %   0.223958   0.0
00000
     0013.2001-06-30.SA_and_HP          1          1               97.14 %   0.000000   0.0
00000
      0003.1999-12-10.kaminski          0          0               97.18 %   0.000000   0.2
76042
        0012.1999-12-14.farmer          0          0               97.22 %   0.000000   0.0
00000
      0009.1999-12-13.kaminski          0          0               97.26 %   0.000000   0.0
00135
     0018.2001-07-13.SA_and_HP          1          1               97.30 %   0.000109   0.0
00000
       0002.2001-02-07.kitchen          0          0               97.33 %   0.000000   0.1
38021
            0007.2004-08-01.BG          1          1               97.37 %   0.003499   0.0
00000
      0012.1999-12-14.kaminski          0          0               97.40 %   0.000000   0.0
```

```
69010
     0005.2001-06-23.SA_and_HP          1          1              97.44 %   0.447917   0.0
00000
      0013.1999-12-14.kaminski          0          0              97.47 %   0.000000   0.0
69010
      0007.1999-12-13.kaminski          0          0              97.50 %   0.000000   0.0
34505
          0017.2000-01-17.beck          0          0              97.53 %   0.000000   0.5
52082
     0006.2001-06-25.SA_and_HP          1          1              97.56 %   0.223958   0.0
00000
      0006.2001-04-03.williams          0          0              97.59 %   0.000000   0.5
52083
       0005.2001-02-08.kitchen          0          0              97.62 %   0.000000   0.0
17253
            0002.2003-12-18.GP          1          1              97.65 %   0.055989   0.0
00000
            0003.2003-12-18.GP          1          1              97.67 %   0.111979   0.0
00000
      0013.2001-04-03.williams          0          0              97.70 %   0.000000   0.1
38021
      0004.2001-04-02.williams          0          0              97.73 %   0.000000   0.1
38021
       0010.2001-02-09.kitchen          0          0              97.75 %   0.000000   0.0
00000
        0013.1999-12-14.farmer          0          0              97.78 %   0.000000   0.0
00000
      0015.1999-12-14.kaminski          0          0              97.80 %   0.000000   0.0
69010
            0012.2003-12-19.GP          1          1              97.83 %   0.223958   0.0
00000
       0016.2001-02-12.kitchen          0          0              97.85 %   0.000000   0.0
00135
            0002.2004-08-01.BG          1          1              97.87 %   0.055990   0.0
00000
     0002.2001-05-25.SA_and_HP          1          1              97.89 %   0.111979   0.0
00000
            0011.2003-12-18.GP          1          1              97.92 %   0.055990   0.0
00000
See confusion matrix. Top Left is predicted correctly as spam. Bottom right predicted corr
```

```
ectly as ham:
[[ 41.   2.]
 [  0.  53.]]
Error rate: 2.08 %%
```

Report

Handling each zero probability with a special treatment we reach an accuracy of 98%. Our special treatment was as follows: if a word shows a zero probability for any class, then count the number of times it appears in the other class. We can then back calculate the probability of each single event assuming that the probability of getting the sequence we observe is 50%.

In [ ]:
```
# This tells matplotlib not to try opening a new window for each plot.
%matplotlib inline
import matplotlib.pyplot as plt
# Create a new figure and set the figsize argument so we get square-ish plots of the 4 feat
ures.
plt.figure(figsize=(15, 3))


for prob in prob_spam_list:
    plt.subplot(1, 4, 1)
    plt.hist(prob, 10)
    plt.title('Probabilities of Spam Class')

for prob in prob_ham_list:
    plt.subplot(1, 4, 2)
    plt.hist(prob, 10)
    plt.title('Probabilities of Ham Class')
```

========== END OF HW 2.3

HW2.4 Repeat HW2.3 with the following modification: use Laplace plus-one smoothing.

Compare the misclassifcation error rates for 2.3 versus 2.4 and explain the differences.

For a quick reference on the construction of the Multinomial NAIVE BAYES classifier that you will code, please consult the "Document Classification" section of the following wikipedia page:

https://en.wikipedia.org/wiki/Naive_Bayes_classifier#Document_classification (https://en.wikipedia.org/wiki/Naive_Bayes_classifier#Document_classification)

OR the original paper by the curators of the Enron email data:

http://www.aueb.gr/users/ion/docs/ceas2006_paper.pdf (http://www.aueb.gr/users/ion/docs/ceas2006_paper.pdf)

Modified Reducer

In [921]:
```
%%writefile reducer.py
#!/usr/bin/python

import numpy as np
from operator import itemgetter
import sys
from itertools import groupby
import math


current_word, word = None, None
current_email, email = None, None
current_y_true, y_true = None, None


sum_records, sum_spamrecords, sum_hamrecords = 0,0,0
sum_spamwords, sum_hamwords = 0,0
spam_zeroes, ham_zeroes = 0,0


emails={} #Associative array to hold email data
words={} #Associative array for word data
conf_matrix = np.zeros((2,2))

# input comes from STDIN
for line in sys.stdin:

    line = line.strip()                         ### remove leading and trailing whitespace
    line = line.split('\t')                     ### parse the input we got from mapper.py
    if len(line) == 4:

        try:
            word = line[0]                          ### word we get from mapper.py
            count = line[1]                         ### corresponding count
            count = int(count)                      ### convert count (currently a string)
to int
            email = line[2]                         ### id that identifies each email
            y_true = line[3]                        ### spam truth
            y_true = int(y_true)                    ### spam truth as an integer
        except ValueError:                          ### any error then silently
```

```
            continue                                ### ignore/discard this line



        if current_word == word:                    ### this IF-switch only works because H
adoop sorts map output
            current_count += count                  ### by key (here: word) before it is pa
ssed to the reducer

            if current_word not in words.keys():     ### initialize word count of new words
                words[current_word]={'ham_count':0,'spam_count':0}

            if current_email not in emails.keys():          ### initialize word counts and
label of new email
                emails[current_email]={'y_true':current_y_true,'word_count':0,'words':[]}
                sum_records +=1.0
                if current_y_true == 1:                     ### identified new email as eit
her spam or ham
                    sum_spamrecords +=1.0

            elif current_email in emails.keys():
                emails[current_email]['word_count'] += 1
                emails[current_email]['words'].append(current_word)### store words in email

            if current_y_true == 1:                         ### if record where word is loc
ated is a spam
                current_spamcount += count         ### add to spam count of that word
                sum_spamwords += 1.0
            else:
                current_hamcount += count          ### if not add to ham count of thet word
                sum_hamwords +=1.0

        else:
            if current_word:
                try:
                    if current_word in words.keys():
                        words[current_word]['spam_count'] += current_spamcount ### update s
pam count for current word
                        words[current_word]['ham_count'] += current_hamcount ### update ham
count for current word
```

```
                except ValueError:                          ### if count was not a number t
hen silently
                    continue                                ### ignore/discard this line


            current_count = count                   ### set current count
            current_word = word                     ### set current number
            current_email = email                   ### set current id of email
            current_y_true = y_true                 ### set current spam truth
            current_spamcount, current_hamcount = 0,0



if current_word == word:                        ### do not forget to output the last word if
needed!
    emails[current_email]['word_count'] += 1
    emails[current_email]['words'].append(current_word)### store words in email
    words[current_word]['spam_count'] += current_spamcount ### update spam count for curren
t word
    words[current_word]['ham_count'] += current_hamcount ### update ham count for current w
ord



#Calculate stats for entire corpus

prior_spam= sum_spamrecords/sum_records
prior_ham= 1 - prior_spam
vocab_count=len(words)#number of unique words in the total vocabulary
prob_spamwords = sum_spamwords/(sum_spamwords+sum_hamwords)
prob_hamwords = 1-prob_spamwords



for k,word in words.iteritems():
    ##These versions calculate conditional probabilities WITH Laplace smoothing.
    word['p_spam']=(word['spam_count']+1)/(sum_spamwords+vocab_count)
    word['p_ham']=(word['ham_count']+1)/(sum_hamwords+vocab_count)

    #Compute conditional probabilities WITHOUT Laplace smoothing
```

```
    #word['p_spam']=(word['spam_count'])/(sum_spamwords)
    #word['p_ham']=(word['ham_count'])/(sum_hamwords)

    if word['p_spam'] == 0 and word['p_ham'] == 0:
        spam_zeroes +=1
        ham_zeroes +=1
        word['p_spam']= 1
        word['p_ham'] = 1

    elif word['p_spam'] == 0:
        spam_zeroes +=1
        word['p_ham'] = min(0.5 ** (1/word['ham_count']),0.99999999) ### back-calculates pr
obability
        word['p_spam']= 1-word['p_ham']              ### of the sequence of events assuming
                                                     ### sequence is 50% likely to happen
    elif word['p_ham'] == 0:
        ham_zeroes +=1
        word['p_spam'] = min(0.5 ** (1/word['spam_count']),0.99999999) ### Adjust probabili
ty
        word['p_ham']= 1-word['p_spam']               ### just as above

    word['p_spam'] = math.log(word['p_spam'])              ### convert to log scale
    word['p_ham'] = math.log(word['p_ham'])

#At this point the model is now trained, and we can use it to make our predictions

print 'Analyzed %s' % sum_records, 'using a vocabulary of %s' % vocab_count , 'terms'
print 'On %s'%spam_zeroes , 'times we found spam probabilities that were zero'
print 'On %s'%ham_zeroes, 'times we found not spam probabilities that were zero'
print '%30s' %'ID', '%10s' %'TRUTH', '%10s' %'CLASS', '%20s' %'CUMULATIVE ACCURACY', '%10
s'%'P-SPAM', '%10s'%'P-HAM'
miss, sample_size = 0,0
prob_spam_list, prob_ham_list = [], []
for j,email in emails.iteritems():

    #Log versions - no longer used
    p_spam= math.log(prior_spam)
    p_ham= math.log(prior_ham)

    #p_spam=prior_spam
```

```
    #p_ham=prior_ham

    for word in email['words']:

        try:
            #p_spam+=log(words[word]['p_spam']) #Log version - no longer used
            p_spam+=words[word]['p_spam']
        except ValueError:
            pass #This means that words that do not appear in a class will use the class pr
ior
        try:
            #p_ham+=log(words[word]['p_ham']) #Log version - no longer used
            p_ham+=words[word]['p_ham']
        except ValueError:
            pass


    if p_spam>p_ham:
        y_pred=1

    else:
        y_pred=0
    y_true = email['y_true']
    if y_true == 1:
        prob_spam_list.append(math.exp(p_spam))
        if y_pred == 1:
            conf_matrix[0][0] +=1                    ### update confusion matrix
        else:
            conf_matrix[0][1] +=1

    else:
        prob_ham_list.append(math.exp(p_spam))
        if y_pred == 0:
            conf_matrix[1][1] +=1
        else:
            conf_matrix[1][0] +=1



    if y_pred != y_true:
```

```
        miss+= 1.0

    sample_size += 1.0
    accuracy = ((sample_size-miss)/sample_size)*100
    prob_spam = math.exp(p_spam)
    prob_ham = math.exp(p_ham)
    print  '%30s' %j, '%10d' %y_true, '%10d' %y_pred, '%18.2f %%' % accuracy, '%10.6f'%pro
b_spam, '%10.6f'%prob_ham
print 'See confusion matrix. Top Left is predicted correctly as spam. Bottom right predicte
d correctly as ham:'
print conf_matrix
print 'Error rate: %4.2f'%(100-accuracy),'%%'
```

```
Overwriting reducer.py
```

Hadoop streaming

In [922]:
```
!hadoop jar hadoop-*streaming*.jar -mapper mapper.py -reducer reducer.py -input enronemai
l_1h.txt -output NBLaplaceSmoothOutput
```

```
16/01/26 18:15:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
16/01/26 18:15:02 INFO Configuration.deprecation: session.id is deprecated. Instead, use d
fs.metrics.session-id
16/01/26 18:15:02 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracke
r, sessionId=
16/01/26 18:15:02 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobT
racker, sessionId= - already initialized
16/01/26 18:15:02 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/26 18:15:02 INFO mapreduce.JobSubmitter: number of splits:1
16/01/26 18:15:03 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1778598
902_0001
16/01/26 18:15:03 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/26 18:15:03 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/26 18:15:03 INFO mapreduce.Job: Running job: job_local1778598902_0001
16/01/26 18:15:03 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapre
d.FileOutputCommitter
16/01/26 18:15:03 INFO output.FileOutputCommitter: File Output Committer Algorithm version
is 1
16/01/26 18:15:03 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/26 18:15:03 INFO mapred.LocalJobRunner: Starting task: attempt_local1778598902_000
1_m_000000_0
16/01/26 18:15:03 INFO output.FileOutputCommitter: File Output Committer Algorithm version
is 1
16/01/26 18:15:03 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is su
pported only on Linux.
16/01/26 18:15:03 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 18:15:03 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/venama
x/enronemail_1h.txt:0+203981
16/01/26 18:15:03 INFO mapred.MapTask: numReduceTasks: 1
16/01/26 18:15:03 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/26 18:15:03 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/26 18:15:03 INFO mapred.MapTask: soft limit at 83886080
16/01/26 18:15:03 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/26 18:15:03 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/26 18:15:03 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapr
ed.MapTask$MapOutputBuffer
16/01/26 18:15:03 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/venamax/ucb-261/hw
2/./mapper.py]
16/01/26 18:15:03 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, us
```

```
e mapreduce.task.id
16/01/26 18:15:03 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead,
use mapreduce.cluster.local.dir
16/01/26 18:15:03 INFO Configuration.deprecation: map.input.file is deprecated. Instead, u
se mapreduce.map.input.file
16/01/26 18:15:03 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, u
se mapreduce.job.skiprecords
16/01/26 18:15:03 INFO Configuration.deprecation: map.input.length is deprecated. Instead,
use mapreduce.map.input.length
16/01/26 18:15:03 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. In
stead, use mapreduce.task.output.dir
16/01/26 18:15:03 INFO Configuration.deprecation: map.input.start is deprecated. Instead,
use mapreduce.map.input.start
16/01/26 18:15:03 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, us
e mapreduce.job.id
16/01/26 18:15:03 INFO Configuration.deprecation: user.name is deprecated. Instead, use ma
preduce.job.user.name
16/01/26 18:15:03 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instea
d, use mapreduce.task.ismap
16/01/26 18:15:03 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, u
se mapreduce.task.attempt.id
16/01/26 18:15:03 INFO Configuration.deprecation: mapred.task.partition is deprecated. Ins
tead, use mapreduce.task.partition
16/01/26 18:15:03 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:15:03 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:15:04 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:15:04 INFO streaming.PipeMapRed: Records R/W=101/1
16/01/26 18:15:04 INFO mapreduce.Job: Job job_local1778598902_0001 running in uber mode :
false
16/01/26 18:15:04 INFO mapreduce.Job:  map 0% reduce 0%
16/01/26 18:15:04 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 18:15:04 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 18:15:04 INFO mapred.LocalJobRunner:
16/01/26 18:15:04 INFO mapred.MapTask: Starting flush of map output
16/01/26 18:15:04 INFO mapred.MapTask: Spilling map output
16/01/26 18:15:04 INFO mapred.MapTask: bufstart = 0; bufend = 1172828; bufvoid = 104857600
16/01/26 18:15:04 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26072600(104
290400); length = 141797/6553600
16/01/26 18:15:05 INFO mapred.MapTask: Finished spill 0
16/01/26 18:15:05 INFO mapred.Task: Task:attempt_local1778598902_0001_m_000000_0 is done.
```

```
And is in the process of committing
16/01/26 18:15:05 INFO mapred.LocalJobRunner: Records R/W=101/1
16/01/26 18:15:05 INFO mapred.Task: Task 'attempt_local1778598902_0001_m_000000_0' done.
16/01/26 18:15:05 INFO mapred.LocalJobRunner: Finishing task: attempt_local1778598902_000
1_m_000000_0
16/01/26 18:15:05 INFO mapred.LocalJobRunner: map task executor complete.
16/01/26 18:15:05 INFO mapreduce.Job:  map 100% reduce 0%
16/01/26 18:15:05 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/26 18:15:05 INFO mapred.LocalJobRunner: Starting task: attempt_local1778598902_000
1_r_000000_0
16/01/26 18:15:05 INFO output.FileOutputCommitter: File Output Committer Algorithm version
is 1
16/01/26 18:15:05 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is su
pported only on Linux.
16/01/26 18:15:05 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 18:15:05 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.m
apreduce.task.reduce.Shuffle@5150e88d
16/01/26 18:15:05 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxS
ingleShuffleLimit=83584616, mergeThreshold=220663392, ioSortFactor=10, memToMemMergeOutput
sThreshold=10
16/01/26 18:15:05 INFO reduce.EventFetcher: attempt_local1778598902_0001_r_000000_0 Thread
started: EventFetcher for fetching Map Completion Events
16/01/26 18:15:05 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map
attempt_local1778598902_0001_m_000000_0 decomp: 1243730 len: 1243734 to MEMORY
16/01/26 18:15:05 INFO reduce.InMemoryMapOutput: Read 1243730 bytes from map-output for at
tempt_local1778598902_0001_m_000000_0
16/01/26 18:15:05 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1
243730, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->1243730
16/01/26 18:15:05 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/26 18:15:05 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 18:15:05 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-out
puts and 0 on-disk map-outputs
16/01/26 18:15:05 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 18:15:05 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of
total size: 1243727 bytes
16/01/26 18:15:05 INFO reduce.MergeManagerImpl: Merged 1 segments, 1243730 bytes to disk t
o satisfy reduce memory limit
16/01/26 18:15:05 INFO reduce.MergeManagerImpl: Merging 1 files, 1243734 bytes from disk
16/01/26 18:15:05 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory in
to reduce
```

```
16/01/26 18:15:05 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 18:15:05 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of
total size: 1243727 bytes
16/01/26 18:15:05 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 18:15:05 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/venamax/ucb-261/hw
2/./reducer.py]
16/01/26 18:15:05 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instea
d, use mapreduce.jobtracker.address
16/01/26 18:15:05 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead,
use mapreduce.job.maps
16/01/26 18:15:05 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:15:05 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:15:05 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:15:05 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:15:06 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:15:07 INFO streaming.PipeMapRed: Records R/W=35450/1
16/01/26 18:15:07 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 18:15:07 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 18:15:08 INFO mapred.Task: Task:attempt_local1778598902_0001_r_000000_0 is done.
And is in the process of committing
16/01/26 18:15:08 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 18:15:08 INFO mapred.Task: Task attempt_local1778598902_0001_r_000000_0 is allowe
d to commit now
16/01/26 18:15:08 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1778
598902_0001_r_000000_0' to hdfs://localhost:9000/user/venamax/NBLaplaceSmoothOutput/_tempo
rary/0/task_local1778598902_0001_r_000000
16/01/26 18:15:08 INFO mapred.LocalJobRunner: Records R/W=35450/1 > reduce
16/01/26 18:15:08 INFO mapred.Task: Task 'attempt_local1778598902_0001_r_000000_0' done.
16/01/26 18:15:08 INFO mapred.LocalJobRunner: Finishing task: attempt_local1778598902_000
1_r_000000_0
16/01/26 18:15:08 INFO mapred.LocalJobRunner: reduce task executor complete.
16/01/26 18:15:08 INFO mapreduce.Job:  map 100% reduce 100%
16/01/26 18:15:08 INFO mapreduce.Job: Job job_local1778598902_0001 completed successfully
16/01/26 18:15:08 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=2699304
                FILE: Number of bytes written=4531266
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
```

```
                HDFS: Number of bytes read=407962
                HDFS: Number of bytes written=9726
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=101
                Map output records=35450
                Map output bytes=1172828
                Map output materialized bytes=1243734
                Input split bytes=104
                Combine input records=0
                Combine output records=0
                Reduce input groups=7275
                Reduce shuffle bytes=1243734
                Reduce input records=35450
                Reduce output records=104
                Spilled Records=70900
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=5
                Total committed heap usage (bytes)=544210944
        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=203981
        File Output Format Counters
                Bytes Written=9726
16/01/26 18:15:08 INFO streaming.StreamJob: Output directory: NBLaplaceSmoothOutput
```

show the results

In [923]:
```
!hdfs dfs -cat NBLaplaceSmoothOutput/part-00000
```

```
16/01/26 18:15:17 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
Analyzed 96.0 using a vocabulary of 3077 terms
On 0 times we found spam probabilities that were zero
On 0 times we found not spam probabilities that were zero
                           ID      TRUTH      CLASS  CUMULATIVE ACCURACY     P-SPAM
P-HAM
            0010.2003-12-18.GP          1          0               0.00 %   0.447917   0.5
52083
     0010.2001-06-28.SA_and_HP          1          1              50.00 %   0.000000   0.0
00000
          0001.2000-01-17.beck          0          0              66.67 %   0.000000   0.0
00000
      0018.1999-12-14.kaminski          0          0              75.00 %   0.000000   0.0
00000
      0005.1999-12-12.kaminski          0          0              80.00 %   0.000000   0.0
00000
     0011.2001-06-29.SA_and_HP          1          1              83.33 %   0.000000   0.0
00000
            0008.2004-08-01.BG          1          1              85.71 %   0.000000   0.0
00000
        0009.1999-12-14.farmer          0          0              87.50 %   0.000000   0.0
00000
            0017.2003-12-18.GP          1          1              88.89 %   0.000000   0.0
00000
     0011.2001-06-28.SA_and_HP          1          1              90.00 %   0.000000   0.0
00000
     0015.2001-07-05.SA_and_HP          1          1              90.91 %   0.000000   0.0
00000
       0015.2001-02-12.kitchen          0          0              91.67 %   0.000000   0.0
00000
      0017.1999-12-14.kaminski          0          0              92.31 %   0.000000   0.0
00000
          0012.2000-01-17.beck          0          0              92.86 %   0.000000   0.0
00000
          0003.2000-01-17.beck          0          0              93.33 %   0.000000   0.0
00000
     0004.2001-06-12.SA_and_HP          1          1              93.75 %   0.000000   0.0
00000
     0008.2001-06-12.SA_and_HP          1          1              94.12 %   0.000000   0.0
```

```
00000
       0007.2001-02-09.kitchen          0          0               94.44 %   0.000000   0.0
00000
            0016.2004-08-01.BG          1          1               94.74 %   0.000000   0.0
00000
        0016.1999-12-15.farmer          0          0               95.00 %   0.000000   0.0
00000
            0013.2004-08-01.BG          1          1               95.24 %   0.000000   0.0
00000
            0005.2003-12-18.GP          1          1               95.45 %   0.000000   0.0
00000
       0012.2001-02-09.kitchen          0          0               95.65 %   0.000000   0.0
00000
        0011.1999-12-14.farmer          0          0               95.83 %   0.000000   0.0
00000
       0009.2001-02-09.kitchen          0          0               96.00 %   0.000000   0.0
00000
       0006.2001-02-08.kitchen          0          0               96.15 %   0.000000   0.0
00000
            0014.2003-12-19.GP          1          1               96.30 %   0.000048   0.0
00046
        0010.1999-12-14.farmer          0          0               96.43 %   0.000000   0.0
00000
            0010.2004-08-01.BG          1          1               96.55 %   0.000000   0.0
00000
      0014.1999-12-14.kaminski          0          0               96.67 %   0.000000   0.0
00000
      0006.1999-12-13.kaminski          0          0               96.77 %   0.000000   0.0
00000
        0005.1999-12-14.farmer          0          0               96.88 %   0.000000   0.0
00000
       0003.2001-02-08.kitchen          0          0               96.97 %   0.000000   0.0
00000
       0001.2001-02-07.kitchen          0          0               97.06 %   0.000000   0.0
00000
       0008.2001-02-09.kitchen          0          0               97.14 %   0.000000   0.0
00000
            0017.2004-08-02.BG          1          1               97.22 %   0.000000   0.0
00000
            0007.2003-12-18.GP          1          1               97.30 %   0.000000   0.0
```

```
00000
            0014.2004-08-01.BG          1          1               97.37 %   0.000000   0.0
00000
            0006.2003-12-18.GP          1          1               97.44 %   0.000000   0.0
00000
     0016.2001-07-05.SA_and_HP          1          1               97.50 %   0.000000   0.0
00000
         0005.2000-06-06.lokay          0          0               97.56 %   0.000000   0.0
00000
     0014.2001-07-04.SA_and_HP          1          1               97.62 %   0.000000   0.0
00000
      0001.2001-04-02.williams          0          0               97.67 %   0.000000   0.0
00000
         0012.2000-06-08.lokay          0          0               97.73 %   0.000000   0.0
00000
        0014.1999-12-15.farmer          0          0               97.78 %   0.000000   0.0
00000
         0009.2000-06-07.lokay          0          0               97.83 %   0.000000   0.0
00000
        0001.1999-12-10.farmer          0          0               97.87 %   0.447917   0.5
52083
     0008.2001-06-25.SA_and_HP          1          1               97.92 %   0.000000   0.0
00000
      0017.2001-04-03.williams          0          0               97.96 %   0.000000   0.0
00000
       0014.2001-02-12.kitchen          0          0               98.00 %   0.000000   0.0
00000
     0016.2001-07-06.SA_and_HP          1          1               98.04 %   0.000000   0.0
00000
        0015.1999-12-15.farmer          0          0               98.08 %   0.000000   0.0
00000
      0004.1999-12-10.kaminski          0          0               98.11 %   0.000000   0.0
00000
            0011.2004-08-01.BG          1          1               98.15 %   0.000000   0.0
00000
            0004.2004-08-01.BG          1          1               98.18 %   0.000000   0.0
00000
            0018.2003-12-18.GP          1          1               98.21 %   0.000000   0.0
00000
        0007.1999-12-14.farmer          0          0               98.25 %   0.000000   0.0
```

```
00000
            0016.2003-12-19.GP          1          1               98.28 %   0.000000   0.0
00000
        0004.1999-12-14.farmer          0          0               98.31 %   0.000000   0.0
00000
            0015.2003-12-19.GP          1          1               98.33 %   0.000000   0.0
00000
            0006.2004-08-01.BG          1          1               98.36 %   0.000000   0.0
00000
            0009.2003-12-18.GP          1          1               98.39 %   0.000000   0.0
00000
        0002.1999-12-13.farmer          0          0               98.41 %   0.000000   0.0
00000
            0008.2003-12-18.GP          1          1               98.44 %   0.000000   0.0
00000
      0010.1999-12-14.kaminski          0          0               98.46 %   0.447917   0.5
52083
          0007.2000-01-17.beck          0          0               98.48 %   0.000000   0.0
00000
        0003.1999-12-14.farmer          0          0               98.51 %   0.000000   0.0
00000
            0003.2004-08-01.BG          1          1               98.53 %   0.000000   0.0
00000
            0017.2004-08-01.BG          1          1               98.55 %   0.000000   0.0
00000
     0013.2001-06-30.SA_and_HP          1          1               98.57 %   0.000000   0.0
00000
      0003.1999-12-10.kaminski          0          0               98.59 %   0.000000   0.0
00000
        0012.1999-12-14.farmer          0          0               98.61 %   0.000000   0.0
00000
      0009.1999-12-13.kaminski          0          0               98.63 %   0.000000   0.0
00000
     0018.2001-07-13.SA_and_HP          1          1               98.65 %   0.000000   0.0
00000
       0002.2001-02-07.kitchen          0          0               98.67 %   0.000000   0.0
00000
            0007.2004-08-01.BG          1          1               98.68 %   0.000000   0.0
00000
      0012.1999-12-14.kaminski          0          0               98.70 %   0.000000   0.0
```

```
00000
     0005.2001-06-23.SA_and_HP          1          1              98.72 %   0.000000   0.0
00000
      0013.1999-12-14.kaminski          0          0              98.73 %   0.000000   0.0
00000
      0007.1999-12-13.kaminski          0          0              98.75 %   0.000000   0.0
00000
          0017.2000-01-17.beck          0          0              98.77 %   0.000000   0.0
00000
     0006.2001-06-25.SA_and_HP          1          1              98.78 %   0.000000   0.0
00000
      0006.2001-04-03.williams          0          0              98.80 %   0.000000   0.0
00000
       0005.2001-02-08.kitchen          0          0              98.81 %   0.000000   0.0
00000
            0002.2003-12-18.GP          1          1              98.82 %   0.000000   0.0
00000
            0003.2003-12-18.GP          1          1              98.84 %   0.000000   0.0
00000
      0013.2001-04-03.williams          0          0              98.85 %   0.000000   0.0
00000
      0004.2001-04-02.williams          0          0              98.86 %   0.000000   0.0
00000
       0010.2001-02-09.kitchen          0          0              98.88 %   0.000000   0.0
00000
        0013.1999-12-14.farmer          0          0              98.89 %   0.000000   0.0
00000
      0015.1999-12-14.kaminski          0          0              98.90 %   0.000000   0.0
00000
            0012.2003-12-19.GP          1          1              98.91 %   0.000000   0.0
00000
       0016.2001-02-12.kitchen          0          0              98.92 %   0.000000   0.0
00000
            0002.2004-08-01.BG          1          1              98.94 %   0.000000   0.0
00000
     0002.2001-05-25.SA_and_HP          1          1              98.95 %   0.000000   0.0
00000
            0011.2003-12-18.GP          1          1              98.96 %   0.000000   0.0
00000
See confusion matrix. Top Left is predicted correctly as spam. Bottom right predicted corr
```

```
ectly as ham:
[[ 42.   1.]
 [  0.  53.]]
Error rate: 1.04 %%
```

Report

Using Laplace allowed us to improve accuracy to 99% from 98%. The difference was on a single record that my prior algorithm predicted as ham. However, even in that prediction the estimated p_spam and p_ham were very close.

========== END OF HW 2.4

HW2.5. Repeat HW2.4. This time when modeling and classification ignore tokens with a frequency of less than three (3) in the training set.

How does it affect the misclassifcation error of learnt naive multinomial Bayesian Classifier on the training dataset:

Modified Reducer

In [924]:
```
%%writefile reducer.py
#!/usr/bin/python

import numpy as np
from operator import itemgetter
import sys
from itertools import groupby
import math


current_word, word = None, None
current_email, email = None, None
current_y_true, y_true = None, None


sum_records, sum_spamrecords, sum_hamrecords = 0,0,0
sum_spamwords, sum_hamwords = 0,0
spam_zeroes, ham_zeroes = 0,0


emails={} #Associative array to hold email data
words={} #Associative array for word data
conf_matrix = np.zeros((2,2))

# input comes from STDIN
for line in sys.stdin:

    line = line.strip()                        ### remove leading and trailing whitespace
    line = line.split('\t')                    ### parse the input we got from mapper.py
    if len(line) == 4:

        try:
            word = line[0]                          ### word we get from mapper.py
            count = line[1]                         ### corresponding count
            count = int(count)                      ### convert count (currently a string)
to int
            email = line[2]                         ### id that identifies each email
            y_true = line[3]                        ### spam truth
            y_true = int(y_true)                    ### spam truth as an integer
        except ValueError:                          ### any error then silently
```

```
            continue                                ### ignore/discard this line



        if current_word == word:                    ### this IF-switch only works because H
adoop sorts map output
            current_count += count                  ### by key (here: word) before it is pa
ssed to the reducer

            if current_word not in words.keys():     ### initialize word count of new words
                words[current_word]={'ham_count':0,'spam_count':0}

            if current_email not in emails.keys():          ### initialize word counts and
label of new email
                emails[current_email]={'y_true':current_y_true,'word_count':0,'words':[]}
                sum_records +=1.0
                if current_y_true == 1:                     ### identified new email as eit
her spam or ham
                    sum_spamrecords +=1.0

            elif current_email in emails.keys():
                emails[current_email]['word_count'] += 1
                emails[current_email]['words'].append(current_word)### store words in email

            if current_y_true == 1:                         ### if record where word is loc
ated is a spam
                current_spamcount += count         ### add to spam count of that word
                sum_spamwords += 1.0
            else:
                current_hamcount += count          ### if not add to ham count of thet word
                sum_hamwords +=1.0

        else:
            if current_word:
                try:
                    if current_word in words.keys():
                        words[current_word]['spam_count'] += current_spamcount ### update s
pam count for current word
                        words[current_word]['ham_count'] += current_hamcount ### update ham
count for current word
```

```
                except ValueError:                          ### if count was not a number t
hen silently
                    continue                                ### ignore/discard this line


            current_count = count                   ### set current count
            current_word = word                     ### set current number
            current_email = email                   ### set current id of email
            current_y_true = y_true                 ### set current spam truth
            current_spamcount, current_hamcount = 0,0



if current_word == word:                        ### do not forget to output the last word if
needed!
    emails[current_email]['word_count'] += 1
    emails[current_email]['words'].append(current_word)### store words in email
    words[current_word]['spam_count'] += current_spamcount ### update spam count for curren
t word
    words[current_word]['ham_count'] += current_hamcount ### update ham count for current w
ord



#Calculate stats for entire corpus

prior_spam= sum_spamrecords/sum_records
prior_ham= 1 - prior_spam
vocab_count=len(words)#number of unique words in the total vocabulary
prob_spamwords = sum_spamwords/(sum_spamwords+sum_hamwords)
prob_hamwords = 1-prob_spamwords



for k,word in words.iteritems():
    ##These versions calculate conditional probabilities WITH Laplace smoothing.
    if (word['spam_count'] + word['ham_count']) < 3:
        word['p_spam']= 1
        word['p_ham'] = 1
    else:
```

```
        word['p_spam']=(word['spam_count']+1)/(sum_spamwords+vocab_count)
        word['p_ham']=(word['ham_count']+1)/(sum_hamwords+vocab_count)

    #Compute conditional probabilities WITHOUT Laplace smoothing
    #word['p_spam']=(word['spam_count'])/(sum_spamwords)
    #word['p_ham']=(word['ham_count'])/(sum_hamwords)

    if word['p_spam'] == 0 and word['p_ham'] == 0:
        spam_zeroes +=1
        ham_zeroes +=1
        word['p_spam']= 1
        word['p_ham'] = 1

    elif word['p_spam'] == 0:
        spam_zeroes +=1
        word['p_ham'] = min(0.5 ** (1/word['ham_count']),0.99999999) ### back-calculates pr
obability
        word['p_spam']= 1-word['p_ham']              ### of the sequence of events assuming
                                                     ### sequence is 50% likely to happen
    elif word['p_ham'] == 0:
        ham_zeroes +=1
        word['p_spam'] = min(0.5 ** (1/word['spam_count']),0.99999999) ### Adjust probabili
ty
        word['p_ham']= 1-word['p_spam']               ### just as above

    word['p_spam'] = math.log(word['p_spam'])              ### convert to log scale
    word['p_ham'] = math.log(word['p_ham'])

#At this point the model is now trained, and we can use it to make our predictions

print 'Analyzed %s' % sum_records, 'using a vocabulary of %s' % vocab_count , 'terms'
print 'On %s'%spam_zeroes , 'times we found spam probabilities that were zero'
print 'On %s'%ham_zeroes, 'times we found not spam probabilities that were zero'
print '%30s' %'ID', '%10s' %'TRUTH', '%10s' %'CLASS', '%20s' %'CUMULATIVE ACCURACY', '%10
s'%'P-SPAM', '%10s'%'P-HAM'
miss, sample_size = 0,0
prob_spam_list, prob_ham_list = [], []
for j,email in emails.iteritems():

    #Log versions - no longer used
```

```
    p_spam= math.log(prior_spam)
    p_ham= math.log(prior_ham)

    #p_spam=prior_spam
    #p_ham=prior_ham

    for word in email['words']:

        try:
            #p_spam+=log(words[word]['p_spam']) #Log version - no longer used
            p_spam+=words[word]['p_spam']
        except ValueError:
            pass #This means that words that do not appear in a class will use the class pr
ior
        try:
            #p_ham+=log(words[word]['p_ham']) #Log version - no longer used
            p_ham+=words[word]['p_ham']
        except ValueError:
            pass


    if p_spam>p_ham:
        y_pred=1

    else:
        y_pred=0
    y_true = email['y_true']
    if y_true == 1:
        prob_spam_list.append(math.exp(p_spam))
        if y_pred == 1:
            conf_matrix[0][0] +=1                    ### update confusion matrix
        else:
            conf_matrix[0][1] +=1
    else:
        prob_ham_list.append(math.exp(p_spam))
        if y_pred == 0:
            conf_matrix[1][1] +=1
        else:
            conf_matrix[1][0] +=1
```

```
    if y_pred != y_true:
        miss+= 1.0

    sample_size += 1.0
    accuracy = ((sample_size-miss)/sample_size)*100
    prob_spam = math.exp(p_spam)
    prob_ham = math.exp(p_ham)
    print  '%30s' %j, '%10d' %y_true, '%10d' %y_pred, '%18.2f %%' % accuracy, '%10.6f'%pro
b_spam, '%10.6f'%prob_ham

print 'See confusion matrix. Top Left is predicted correctly as spam. Bottom right predicte
d correctly as ham:'
print conf_matrix
print 'Error rate: %4.2f'%(100-accuracy),'%%'
```

```
Overwriting reducer.py
```

Hadoop Streaming

In [925]:
```
!hadoop jar hadoop-*streaming*.jar -mapper mapper.py -reducer reducer.py -input enronemai
l_1h.txt -output NBLaplaceSmoothLess3Output
```

```
16/01/26 18:15:51 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
16/01/26 18:15:51 INFO Configuration.deprecation: session.id is deprecated. Instead, use d
fs.metrics.session-id
16/01/26 18:15:51 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracke
r, sessionId=
16/01/26 18:15:51 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobT
racker, sessionId= - already initialized
16/01/26 18:15:52 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/26 18:15:52 INFO mapreduce.JobSubmitter: number of splits:1
16/01/26 18:15:52 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1973410
96_0001
16/01/26 18:15:52 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/26 18:15:52 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/26 18:15:52 INFO mapreduce.Job: Running job: job_local197341096_0001
16/01/26 18:15:52 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapre
d.FileOutputCommitter
16/01/26 18:15:52 INFO output.FileOutputCommitter: File Output Committer Algorithm version
is 1
16/01/26 18:15:52 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/26 18:15:52 INFO mapred.LocalJobRunner: Starting task: attempt_local197341096_000
1_m_000000_0
16/01/26 18:15:52 INFO output.FileOutputCommitter: File Output Committer Algorithm version
is 1
16/01/26 18:15:52 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is su
pported only on Linux.
16/01/26 18:15:52 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 18:15:52 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/venama
x/enronemail_1h.txt:0+203981
16/01/26 18:15:52 INFO mapred.MapTask: numReduceTasks: 1
16/01/26 18:15:52 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/26 18:15:52 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/26 18:15:52 INFO mapred.MapTask: soft limit at 83886080
16/01/26 18:15:52 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/26 18:15:52 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/26 18:15:52 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapr
ed.MapTask$MapOutputBuffer
16/01/26 18:15:52 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/venamax/ucb-261/hw
2/./mapper.py]
16/01/26 18:15:52 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, us
```

```
e mapreduce.task.id
16/01/26 18:15:52 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead,
use mapreduce.cluster.local.dir
16/01/26 18:15:52 INFO Configuration.deprecation: map.input.file is deprecated. Instead, u
se mapreduce.map.input.file
16/01/26 18:15:52 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, u
se mapreduce.job.skiprecords
16/01/26 18:15:52 INFO Configuration.deprecation: map.input.length is deprecated. Instead,
use mapreduce.map.input.length
16/01/26 18:15:52 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. In
stead, use mapreduce.task.output.dir
16/01/26 18:15:52 INFO Configuration.deprecation: map.input.start is deprecated. Instead,
use mapreduce.map.input.start
16/01/26 18:15:52 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, us
e mapreduce.job.id
16/01/26 18:15:52 INFO Configuration.deprecation: user.name is deprecated. Instead, use ma
preduce.job.user.name
16/01/26 18:15:52 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instea
d, use mapreduce.task.ismap
16/01/26 18:15:52 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, u
se mapreduce.task.attempt.id
16/01/26 18:15:52 INFO Configuration.deprecation: mapred.task.partition is deprecated. Ins
tead, use mapreduce.task.partition
16/01/26 18:15:52 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:15:52 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:15:52 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:15:52 INFO streaming.PipeMapRed: Records R/W=101/1
16/01/26 18:15:53 INFO mapreduce.Job: Job job_local197341096_0001 running in uber mode : f
alse
16/01/26 18:15:53 INFO mapreduce.Job:  map 0% reduce 0%
16/01/26 18:15:53 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 18:15:53 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 18:15:53 INFO mapred.LocalJobRunner:
16/01/26 18:15:53 INFO mapred.MapTask: Starting flush of map output
16/01/26 18:15:53 INFO mapred.MapTask: Spilling map output
16/01/26 18:15:53 INFO mapred.MapTask: bufstart = 0; bufend = 1172828; bufvoid = 104857600
16/01/26 18:15:53 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26072600(104
290400); length = 141797/6553600
16/01/26 18:15:53 INFO mapred.MapTask: Finished spill 0
16/01/26 18:15:53 INFO mapred.Task: Task:attempt_local197341096_0001_m_000000_0 is done. A
```

```
nd is in the process of committing
16/01/26 18:15:53 INFO mapred.LocalJobRunner: Records R/W=101/1
16/01/26 18:15:53 INFO mapred.Task: Task 'attempt_local197341096_0001_m_000000_0' done.
16/01/26 18:15:53 INFO mapred.LocalJobRunner: Finishing task: attempt_local197341096_000
1_m_000000_0
16/01/26 18:15:53 INFO mapred.LocalJobRunner: map task executor complete.
16/01/26 18:15:53 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/26 18:15:53 INFO mapred.LocalJobRunner: Starting task: attempt_local197341096_000
1_r_000000_0
16/01/26 18:15:53 INFO output.FileOutputCommitter: File Output Committer Algorithm version
is 1
16/01/26 18:15:53 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is su
pported only on Linux.
16/01/26 18:15:53 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 18:15:53 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.m
apreduce.task.reduce.Shuffle@bed9885
16/01/26 18:15:53 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxS
ingleShuffleLimit=83584616, mergeThreshold=220663392, ioSortFactor=10, memToMemMergeOutput
sThreshold=10
16/01/26 18:15:53 INFO reduce.EventFetcher: attempt_local197341096_0001_r_000000_0 Thread
started: EventFetcher for fetching Map Completion Events
16/01/26 18:15:53 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map
attempt_local197341096_0001_m_000000_0 decomp: 1243730 len: 1243734 to MEMORY
16/01/26 18:15:53 INFO reduce.InMemoryMapOutput: Read 1243730 bytes from map-output for at
tempt_local197341096_0001_m_000000_0
16/01/26 18:15:53 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1
243730, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->1243730
16/01/26 18:15:53 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/26 18:15:53 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 18:15:53 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-out
puts and 0 on-disk map-outputs
16/01/26 18:15:53 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 18:15:53 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of
total size: 1243727 bytes
16/01/26 18:15:53 INFO reduce.MergeManagerImpl: Merged 1 segments, 1243730 bytes to disk t
o satisfy reduce memory limit
16/01/26 18:15:53 INFO reduce.MergeManagerImpl: Merging 1 files, 1243734 bytes from disk
16/01/26 18:15:53 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory in
to reduce
16/01/26 18:15:53 INFO mapred.Merger: Merging 1 sorted segments
```

```
16/01/26 18:15:53 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of
total size: 1243727 bytes
16/01/26 18:15:53 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 18:15:53 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/venamax/ucb-261/hw
2/./reducer.py]
16/01/26 18:15:53 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instea
d, use mapreduce.jobtracker.address
16/01/26 18:15:53 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead,
use mapreduce.job.maps
16/01/26 18:15:53 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:15:53 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:15:53 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:15:53 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:15:54 INFO mapreduce.Job:  map 100% reduce 0%
16/01/26 18:15:54 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 18:15:55 INFO streaming.PipeMapRed: Records R/W=35450/1
16/01/26 18:15:55 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 18:15:55 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 18:15:56 INFO mapred.Task: Task:attempt_local197341096_0001_r_000000_0 is done. A
nd is in the process of committing
16/01/26 18:15:56 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 18:15:56 INFO mapred.Task: Task attempt_local197341096_0001_r_000000_0 is allowed
to commit now
16/01/26 18:15:56 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1973
41096_0001_r_000000_0' to hdfs://localhost:9000/user/venamax/NBLaplaceSmoothLess3Output/_t
emporary/0/task_local197341096_0001_r_000000
16/01/26 18:15:56 INFO mapred.LocalJobRunner: Records R/W=35450/1 > reduce
16/01/26 18:15:56 INFO mapred.Task: Task 'attempt_local197341096_0001_r_000000_0' done.
16/01/26 18:15:56 INFO mapred.LocalJobRunner: Finishing task: attempt_local197341096_000
1_r_000000_0
16/01/26 18:15:56 INFO mapred.LocalJobRunner: reduce task executor complete.
16/01/26 18:15:56 INFO mapreduce.Job:  map 100% reduce 100%
16/01/26 18:15:56 INFO mapreduce.Job: Job job_local197341096_0001 completed successfully
16/01/26 18:15:56 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=2699304
                FILE: Number of bytes written=4528278
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
```

```
                HDFS: Number of bytes read=407962
                HDFS: Number of bytes written=9726
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=101
                Map output records=35450
                Map output bytes=1172828
                Map output materialized bytes=1243734
                Input split bytes=104
                Combine input records=0
                Combine output records=0
                Reduce input groups=7275
                Reduce shuffle bytes=1243734
                Reduce input records=35450
                Reduce output records=104
                Spilled Records=70900
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=4
                Total committed heap usage (bytes)=545259520
        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=203981
        File Output Format Counters
                Bytes Written=9726
16/01/26 18:15:56 INFO streaming.StreamJob: Output directory: NBLaplaceSmoothLess3Output
```

show results

In [926]:
```
!hdfs dfs -cat NBLaplaceSmoothLess3Output/part-00000
```

```
16/01/26 18:16:11 WARN util.NativeCodeLoader: Unable to load native-hadoop library for you
r platform... using builtin-java classes where applicable
Analyzed 96.0 using a vocabulary of 3077 terms
On 0 times we found spam probabilities that were zero
On 0 times we found not spam probabilities that were zero
                          ID      TRUTH      CLASS  CUMULATIVE ACCURACY     P-SPAM
P-HAM
            0010.2003-12-18.GP          1          0               0.00 %   0.447917   0.5
52083
     0010.2001-06-28.SA_and_HP          1          1              50.00 %   0.000000   0.0
00000
          0001.2000-01-17.beck          0          0              66.67 %   0.000000   0.0
00000
      0018.1999-12-14.kaminski          0          0              75.00 %   0.000000   0.0
00000
      0005.1999-12-12.kaminski          0          0              80.00 %   0.000000   0.0
00000
     0011.2001-06-29.SA_and_HP          1          1              83.33 %   0.000000   0.0
00000
            0008.2004-08-01.BG          1          1              85.71 %   0.000000   0.0
00000
        0009.1999-12-14.farmer          0          0              87.50 %   0.000000   0.0
00000
            0017.2003-12-18.GP          1          1              88.89 %   0.000000   0.0
00000
     0011.2001-06-28.SA_and_HP          1          1              90.00 %   0.000000   0.0
00000
     0015.2001-07-05.SA_and_HP          1          1              90.91 %   0.000000   0.0
00000
       0015.2001-02-12.kitchen          0          0              91.67 %   0.000000   0.0
00000
      0017.1999-12-14.kaminski          0          0              92.31 %   0.000000   0.0
00000
          0012.2000-01-17.beck          0          0              92.86 %   0.000000   0.0
00000
          0003.2000-01-17.beck          0          0              93.33 %   0.000000   0.0
00000
     0004.2001-06-12.SA_and_HP          1          1              93.75 %   0.000000   0.0
00000
     0008.2001-06-12.SA_and_HP          1          1              94.12 %   0.000000   0.0
```

```
00000
       0007.2001-02-09.kitchen          0          0               94.44 %   0.000000   0.0
00000
            0016.2004-08-01.BG          1          1               94.74 %   0.000000   0.0
00000
        0016.1999-12-15.farmer          0          0               95.00 %   0.000000   0.0
00000
            0013.2004-08-01.BG          1          1               95.24 %   0.000000   0.0
00000
            0005.2003-12-18.GP          1          1               95.45 %   0.000000   0.0
00000
       0012.2001-02-09.kitchen          0          0               95.65 %   0.000000   0.0
00000
        0011.1999-12-14.farmer          0          0               95.83 %   0.000000   0.0
00000
       0009.2001-02-09.kitchen          0          0               96.00 %   0.000000   0.0
00000
       0006.2001-02-08.kitchen          0          0               96.15 %   0.000000   0.0
00000
            0014.2003-12-19.GP          1          0               92.59 %   0.447917   0.5
52083
        0010.1999-12-14.farmer          0          0               92.86 %   0.000000   0.0
00000
            0010.2004-08-01.BG          1          1               93.10 %   0.000000   0.0
00000
      0014.1999-12-14.kaminski          0          0               93.33 %   0.000000   0.0
00000
      0006.1999-12-13.kaminski          0          0               93.55 %   0.000000   0.0
00000
        0005.1999-12-14.farmer          0          0               93.75 %   0.000000   0.0
00000
       0003.2001-02-08.kitchen          0          0               93.94 %   0.000000   0.0
00000
       0001.2001-02-07.kitchen          0          0               94.12 %   0.000000   0.0
00000
       0008.2001-02-09.kitchen          0          0               94.29 %   0.000000   0.0
00000
            0017.2004-08-02.BG          1          1               94.44 %   0.000000   0.0
00000
            0007.2003-12-18.GP          1          0               91.89 %   0.447917   0.5
```

```
52083
            0014.2004-08-01.BG          1          1               92.11 %   0.000000   0.0
00000
            0006.2003-12-18.GP          1          1               92.31 %   0.000000   0.0
00000
     0016.2001-07-05.SA_and_HP          1          1               92.50 %   0.000000   0.0
00000
         0005.2000-06-06.lokay          0          0               92.68 %   0.000000   0.0
00000
     0014.2001-07-04.SA_and_HP          1          1               92.86 %   0.000000   0.0
00000
      0001.2001-04-02.williams          0          0               93.02 %   0.000000   0.0
00000
         0012.2000-06-08.lokay          0          0               93.18 %   0.000000   0.0
00000
        0014.1999-12-15.farmer          0          0               93.33 %   0.000000   0.0
00000
         0009.2000-06-07.lokay          0          0               93.48 %   0.000000   0.0
00000
        0001.1999-12-10.farmer          0          0               93.62 %   0.447917   0.5
52083
     0008.2001-06-25.SA_and_HP          1          1               93.75 %   0.000000   0.0
00000
      0017.2001-04-03.williams          0          0               93.88 %   0.000000   0.0
00000
       0014.2001-02-12.kitchen          0          0               94.00 %   0.000000   0.0
00000
     0016.2001-07-06.SA_and_HP          1          1               94.12 %   0.000000   0.0
00000
        0015.1999-12-15.farmer          0          0               94.23 %   0.000000   0.0
00000
      0004.1999-12-10.kaminski          0          0               94.34 %   0.000000   0.0
00000
            0011.2004-08-01.BG          1          1               94.44 %   0.000000   0.0
00000
            0004.2004-08-01.BG          1          1               94.55 %   0.000000   0.0
00000
            0018.2003-12-18.GP          1          1               94.64 %   0.000000   0.0
00000
        0007.1999-12-14.farmer          0          0               94.74 %   0.000000   0.0
```

```
00000
            0016.2003-12-19.GP          1          1               94.83 %   0.000000   0.0
00000
        0004.1999-12-14.farmer          0          0               94.92 %   0.000000   0.0
00000
            0015.2003-12-19.GP          1          1               95.00 %   0.000000   0.0
00000
            0006.2004-08-01.BG          1          1               95.08 %   0.000000   0.0
00000
            0009.2003-12-18.GP          1          1               95.16 %   0.000000   0.0
00000
        0002.1999-12-13.farmer          0          0               95.24 %   0.000000   0.0
00000
            0008.2003-12-18.GP          1          1               95.31 %   0.000000   0.0
00000
      0010.1999-12-14.kaminski          0          0               95.38 %   0.447917   0.5
52083
          0007.2000-01-17.beck          0          0               95.45 %   0.000000   0.0
00000
        0003.1999-12-14.farmer          0          0               95.52 %   0.447917   0.5
52083
            0003.2004-08-01.BG          1          1               95.59 %   0.000000   0.0
00000
            0017.2004-08-01.BG          1          1               95.65 %   0.000000   0.0
00000
     0013.2001-06-30.SA_and_HP          1          1               95.71 %   0.000000   0.0
00000
      0003.1999-12-10.kaminski          0          0               95.77 %   0.000000   0.0
00000
        0012.1999-12-14.farmer          0          0               95.83 %   0.000000   0.0
00000
      0009.1999-12-13.kaminski          0          0               95.89 %   0.000000   0.0
00000
     0018.2001-07-13.SA_and_HP          1          1               95.95 %   0.000000   0.0
00000
       0002.2001-02-07.kitchen          0          0               96.00 %   0.000000   0.0
00000
            0007.2004-08-01.BG          1          1               96.05 %   0.000000   0.0
00000
      0012.1999-12-14.kaminski          0          0               96.10 %   0.000000   0.0
```

```
00000
     0005.2001-06-23.SA_and_HP          1          0              94.87 %   0.447917   0.5
52083
      0013.1999-12-14.kaminski          0          0              94.94 %   0.000000   0.0
00000
      0007.1999-12-13.kaminski          0          0              95.00 %   0.000000   0.0
00000
          0017.2000-01-17.beck          0          0              95.06 %   0.000000   0.0
00000
     0006.2001-06-25.SA_and_HP          1          0              93.90 %   0.447917   0.5
52083
      0006.2001-04-03.williams          0          0              93.98 %   0.000000   0.0
00000
       0005.2001-02-08.kitchen          0          0              94.05 %   0.000000   0.0
00000
            0002.2003-12-18.GP          1          1              94.12 %   0.000000   0.0
00000
            0003.2003-12-18.GP          1          1              94.19 %   0.000000   0.0
00000
      0013.2001-04-03.williams          0          0              94.25 %   0.000000   0.0
00000
      0004.2001-04-02.williams          0          0              94.32 %   0.447917   0.5
52083
       0010.2001-02-09.kitchen          0          0              94.38 %   0.000000   0.0
00000
        0013.1999-12-14.farmer          0          0              94.44 %   0.000000   0.0
00000
      0015.1999-12-14.kaminski          0          0              94.51 %   0.000000   0.0
00000
            0012.2003-12-19.GP          1          1              94.57 %   0.000000   0.0
00000
       0016.2001-02-12.kitchen          0          0              94.62 %   0.000000   0.0
00000
            0002.2004-08-01.BG          1          1              94.68 %   0.000000   0.0
00000
     0002.2001-05-25.SA_and_HP          1          1              94.74 %   0.000000   0.0
00000
            0011.2003-12-18.GP          1          1              94.79 %   0.000000   0.0
00000
See confusion matrix. Top Left is predicted correctly as spam. Bottom right predicted corr
```

```
ectly as ham:
[[ 38.   5.]
 [  0.  53.]]
Error rate: 5.21 %%
```

Report

Error rate went up because we're now discarding some words that are not that frequent but are great predictors. So I would not recommend to drop terms that are not that frequent. We should use the frequency to estimate how likely it is that term is a good predictor. That's what I did on 2.3. I adjusted zero probabilities by their frequency taking into account that as a term has higher frequency it is more likely to appear on both ham and spam lists.

========== END OF HW 2.5

HW2.6 Benchmark your code with the Python SciKit-Learn implementation of the multinomial Naive Bayes algorithm

It always a good idea to benchmark your solutions against publicly available libraries such as SciKit-Learn, The Machine Learning toolkit available in Python. In this exercise, we benchmark ourselves against the SciKit-Learn implementation of multinomial Naive Bayes. For more information on this implementation see: http://scikit-learn.org/stable/modules/naive_bayes.html (http://scikit-learn.org/stable/modules/naive_bayes.html) more

In this exercise, please complete the following:

— Run the Multinomial Naive Bayes algorithm (using default settings) from SciKit-Learn over the same training data used in HW2.5 and report the misclassification error (please note some data preparation might be needed to get the Multinomial Naive Bayes algorithm from SkiKit-Learn to run over this dataset)

- Prepare a table to present your results, where rows correspond to approach used (SkiKit-Learn versus your Hadoop implementation) and the column presents the training misclassification error — Explain/justify any differences in terms of training error rates over the dataset in HW2.5 between your Multinomial Naive Bayes implementation (in Map Reduce) versus the Multinomial Naive Bayes implementation in SciKit-Learn

In [854]:
```
import numpy as np
from sklearn.naive_bayes import MultinomialNB
from sklearn.feature_extraction.text import CountVectorizer
import sys
import re
# SK-learn libraries for evaluation.
from sklearn.metrics import confusion_matrix
from sklearn import metrics
from sklearn.metrics import classification_report

message, spam_truth, email_ids = [], [], []

WORD_RE = re.compile(r"[\w']+") #Compile regex to easily parse complete words
filename = 'enronemail_1h.txt'
with open (filename, "r") as myfile:
    for num,line in enumerate(myfile.readlines()):
        fields=line.split('\t') #parse line into separate fields
        subject_and_body=" ".join(fields[-2:]).strip()#parse the subject and body fields fr
om the line, and combine into one string
        if len(fields) == 4:
            message.append(subject_and_body)
            spam_truth.append(int(fields[1]))
            email_ids.append(fields[0])

for i in range (5):              ### print out data to make sure it displays expected forma
t
    print (message[i])
    print (spam_truth[i])
    print (email_ids[i])
```

christmas tree farm pictures NA
0
0001.1999-12-10.farmer
re: rankings thank you.
0
0001.1999-12-10.kaminski
leadership development pilot sally: what timing, ask and you shall receive. as per our d
iscussion, listed below is an update on the leadership pilot. your vendor selection team
will receive an update and even more information later in the week. on the lunch & learn
for energy operations, the audience and focus will be your group. we are ready to start u
p when appropriate. thank you for your time today. please call me if you have any questio
ns at x 33597. ----------forwarded by julie armstrong/corp/enron on 01/17/20
00 06:44 pm---------- from: susan runkel @ ect 01/17/2000 03:22 pm to:
cindy skinner/hou/ect @ ect, brad mcsherry/hou/ect @ ect, norma villarreal/hou/ect @ ect,
kimberly rizzi/hou/ect @ ect, fran l mayes/hou/ect @ ect, gary buck/hou/ect @ ect, robert
jones/corp/enron @ enron, sheila walton/hou/ect @ ect, philip conn/corp/enron @ enron, ma
ry overgaard/pdx/ect @ ect, kim melodick/hou/ect @ ect, valeria a hope/hou/ect @ ect cc:
david oxley/hou/ect @ ect, susan carrera/hou/ect @ ect, jane allen/hou/ect @ ect, christi
ne shenkman/enron_development @ enron_development, kathryn mclean/hou/ect @ ect, gracie s
presas/hou/ect @ ect, janice riedel/hou/ect @ ect, julie armstrong/corp/enron @ enron su
bject: leadership development pilot good news regarding the ena leadership curriculum! th
rough the help of a vendor selection team from eops, we've chosen southwest performance g
roup and wilson learning products as one of our primary vendors for the leadership curri
culum and programs. we are ready to conduct a pilot on february 8-10 of six modules. the
purpose of the pilot is to evaluate for fine-tuning the wilson learning materials and fac
ilitators and to present just a portion of the leadership curriculum. in order to evalua
te the materials thoroughly, it would be great to get a cross-section of ena to attend. w
e are asking that you invite several supervisors from your client groups to participate i
n any of the courses listed below. the sessions will be held in room 560 and times are li
sted below. also attached is a description of the modules. all are designed for supervis
ors only, with the exception being " communicating effectively ". this is open to any emp
loyee. as a benefit in attending the pilot, i will pick up the cost., so there will be no
charge back for their attendance. we are currently completing the curriculum design and w
ill have information on the full curriculum available in february. this will include opti
ons other than " classrom setting " for development. please respond back to gracie presa
s by february 1 with your names. if you have further questions, please contact me at 3-73
94. we are really excited that we have this available and hope that your clients will fin
d it to be valuable. the following are half-day sessions. supervisors may sign up for an
y or all depending on their need. it would be helpful if supervisors attend a minimum of
two modules. date module time target audience feb. 8 meeting leadership challenges 8-12

```
am supervisors with less  than 6 months experience  working styles 1-5 pm any supervisor
feb. 9 coaching to performance 8-12 am any supervisor  motivating for results 1-5 pm any s
upervisor  feb. 10 communicating effectively 8-12 am any employee  delegating and directin
g 1-5 pm any supervisor
0
0001.2000-01-17.beck
key hr issues going forward  a) year end reviews-report needs generating like mid-year doc
umenting business unit performance on review completion-david to john;  b) work out or pla
n generation for the nim/issues employees-david to john;  c) hpl transition issues-ongoin
g.  officially transferred.  regards  delainey
0
0001.2001-02-07.kitchen
re: quasi  good morning,  i'd love to go get some coffee with you, but remember that annoy
ing project that mike etringer wants me to work on for him? this morning i am kinda under
some pressure to hurry up and try to get some stuff figured out so i really don't have muc
h spare time right now. ja would flip out if i left for coffee now. maybe later this after
noon? or tomorrow morning? anyhow, another ride sounds really cool. i had lots of fun. and
yes, it would be cooler if i didn't have to worry about work. let me know when you have ex
tra time to go for a ride.  my weekend was pretty fun. i weed-wacked (is that a word?) my
yard for the first time. it looks so bad. i so don't know anything about lawn care. also i
planted some herbs and stuff in my yard which i am sure my dog will destroy, but it s wort
h a try. oh yeah, i also bought a snowboard. it's pretty cool. i bought some step-in switc
h boots, too. cool, eh?  so i'll talk to you later. have a great day.
0
0001.2001-04-02.williams
```

In [855]:

```
### Vectorize train_data and call it matrix X
vectorizer = CountVectorizer(min_df=1)
train = vectorizer.fit_transform(message)
print (train.shape)
```

```
(98, 5319)
```

In [856]:

```
labels = np.array(spam_truth)
print (labels.shape)
```

```
(98,)
```

In [900]:
```
#### Let's now define a MultiNomial classifier testing certain alphas and finding the optim
al
alphas = [0.0, 0.0001, 0.001, 0.01, 0.1, 0.5, 1.0, 2.0, 10.0]
max_mn_f1, max_alpha = 0,0
for i in range(len(alphas)):
    mn = MultinomialNB(alpha=alphas[i])
    mn.fit(train,labels)
    y_true = labels
    y_pred = mn.predict(train)
    if metrics.f1_score(y_true, y_pred, average ='weighted') > max_mn_f1:
        max_mn_f1 = metrics.f1_score(y_true, y_pred, average ='weighted')
        max_alpha = alphas[i]

mn = MultinomialNB(alpha=max_alpha)
mn.fit(train,labels)
y_true = labels
y_pred = mn.predict(train)
print ('SciKit-Learn Metrics:')
print ('F1:',(metrics.f1_score(y_true, y_pred, average ='weighted')))
print ('Confusion matrix:')
print ((confusion_matrix(y_true, y_pred)))
```

```
SciKit-Learn Metrics:
F1: 1.0
Confusion matrix:
[[55  0]
 [ 0 43]]
```

```
/Users/venamax/anaconda/lib/python3.5/site-packages/sklearn/naive_bayes.py:664: RuntimeWar
ning: divide by zero encountered in log
  self.feature_log_prob_ = (np.log(smoothed_fc)
/Users/venamax/anaconda/lib/python3.5/site-packages/sklearn/metrics/classification.py:976:
DeprecationWarning: From version 0.18, binary input will not be handled specially when usi
ng averaged precision/recall/F-score. Please use average='binary' to report only the posit
ive class performance.
  'positive class performance.', DeprecationWarning)
/Users/venamax/anaconda/lib/python3.5/site-packages/sklearn/metrics/classification.py:976:
DeprecationWarning: From version 0.18, binary input will not be handled specially when usi
ng averaged precision/recall/F-score. Please use average='binary' to report only the posit
ive class performance.
  'positive class performance.', DeprecationWarning)
/Users/venamax/anaconda/lib/python3.5/site-packages/sklearn/metrics/classification.py:976:
DeprecationWarning: From version 0.18, binary input will not be handled specially when usi
ng averaged precision/recall/F-score. Please use average='binary' to report only the posit
ive class performance.
  'positive class performance.', DeprecationWarning)
/Users/venamax/anaconda/lib/python3.5/site-packages/sklearn/metrics/classification.py:976:
DeprecationWarning: From version 0.18, binary input will not be handled specially when usi
ng averaged precision/recall/F-score. Please use average='binary' to report only the posit
ive class performance.
  'positive class performance.', DeprecationWarning)
/Users/venamax/anaconda/lib/python3.5/site-packages/sklearn/metrics/classification.py:976:
DeprecationWarning: From version 0.18, binary input will not be handled specially when usi
ng averaged precision/recall/F-score. Please use average='binary' to report only the posit
ive class performance.
  'positive class performance.', DeprecationWarning)
/Users/venamax/anaconda/lib/python3.5/site-packages/sklearn/metrics/classification.py:976:
DeprecationWarning: From version 0.18, binary input will not be handled specially when usi
ng averaged precision/recall/F-score. Please use average='binary' to report only the posit
ive class performance.
  'positive class performance.', DeprecationWarning)
/Users/venamax/anaconda/lib/python3.5/site-packages/sklearn/metrics/classification.py:976:
DeprecationWarning: From version 0.18, binary input will not be handled specially when usi
ng averaged precision/recall/F-score. Please use average='binary' to report only the posit
ive class performance.
  'positive class performance.', DeprecationWarning)
/Users/venamax/anaconda/lib/python3.5/site-packages/sklearn/metrics/classification.py:976:
DeprecationWarning: From version 0.18, binary input will not be handled specially when usi
ng averaged precision/recall/F-score. Please use average='binary' to report only the posit
ive class performance.
  'positive class performance.', DeprecationWarning)
/Users/venamax/anaconda/lib/python3.5/site-packages/sklearn/metrics/classification.py:976:
DeprecationWarning: From version 0.18, binary input will not be handled specially when usi
```

```
ng averaged precision/recall/F-score. Please use average='binary' to report only the posit
ive class performance.
  'positive class performance.', DeprecationWarning)
/Users/venamax/anaconda/lib/python3.5/site-packages/sklearn/metrics/classification.py:976:
DeprecationWarning: From version 0.18, binary input will not be handled specially when usi
ng averaged precision/recall/F-score. Please use average='binary' to report only the posit
ive class performance.
  'positive class performance.', DeprecationWarning)
/Users/venamax/anaconda/lib/python3.5/site-packages/sklearn/metrics/classification.py:976:
DeprecationWarning: From version 0.18, binary input will not be handled specially when usi
ng averaged precision/recall/F-score. Please use average='binary' to report only the posit
ive class performance.
  'positive class performance.', DeprecationWarning)
/Users/venamax/anaconda/lib/python3.5/site-packages/sklearn/metrics/classification.py:976:
DeprecationWarning: From version 0.18, binary input will not be handled specially when usi
ng averaged precision/recall/F-score. Please use average='binary' to report only the posit
ive class performance.
  'positive class performance.', DeprecationWarning)
```

Comparison Table

Method	Misclassifications	Rationale
MNB SciKit-Learn	None	model was optimized to fit train data and then run against that same data so it clearly overfits the data
MNB 2.5 - Removing low freq tokens	5	low frequency data provides critical info that reduces prediction power
MNB 2.4 Laplace Smoothing	2	works very well. it missed only an email with only one word
MNB 2.3 No Laplace Smoothing	1	almost as good as Laplace smoothing. Zero probabilities were adjusted based on frequency

========== END OF HW 2.6

HHW 2.6.1 OPTIONAL (note this exercise is a stretch HW and optional)

— Run the Bernoulli Naive Bayes algorithm from SciKit-Learn (using default settings) over the same training data used in HW2.6 and report the misclassification error

- Discuss the performance differences in terms of misclassification error rates over the dataset in HW2.5 between the Multinomial Naive Bayes implementation in SciKit-Learn with the Bernoulli Naive Bayes implementation in SciKit-Learn. Why such big differences. Explain.

Which approach to Naive Bayes would you recommend for SPAM detection? Justify your selection.

HW2.7 OPTIONAL (note this exercise is a stretch HW and optional)

The Enron SPAM data in the following folder enron1-Training-Data-RAW is in raw text form (with subfolders for SPAM and HAM that contain raw email messages in the following form:

--- Line 1 contains the subject --- The remaining lines contain the body of the email message.

In Python write a script to produce a TSV file called train-Enron-1.txt that has a similar format as the enronemail_1h.txt that you have been using so far. Please pay attend to funky characters and tabs. Check your resulting formated email data in Excel and in Python (e.g., count up the number of fields in each row; the number of SPAM mails and the number of HAM emails). Does each row correspond to an email record with four values? Note: use "NA" to denote empty field values.

HW2.8 OPTIONAL

Using Hadoop Map-Reduce write job(s) to perform the following: -- Train a multinomial Naive Bayes Classifier with Laplace plus one smoothing using the data extracted in HW2.7 (i.e., train-Enron-1.txt). Use all white-space delimitted tokens as independent input variables (assume spaces, fullstops, commas as delimiters). Drop tokens with a frequency of less than three (3). -- Test the learnt classifier using enronemail_1h.txt and report the misclassification error rate. Remember to use all white-space delimitted tokens as independent input variables (assume spaces, fullstops, commas as delimiters). How do we treat tokens in the test set that do not appear in the training set?

HW2.8.1 OPTIONAL

— Run both the Multinomial Naive Bayes and the Bernoulli Naive Bayes algorithms from SciKit-Learn (using default settings) over the same training data used in HW2.8 and report the misclassification error on both the training set and the testing set

- Prepare a table to present your results, where rows correspond to approach used (SciKit-Learn Multinomial NB; SciKit-Learn Bernouili NB; Your Hadoop implementation) and the columns presents the training misclassification error, and the misclassification error on the test data set
- Discuss the performance differences in terms of misclassification error rates over the test and training datasets by the different implementations. Which approch (Bernouili versus Multinomial) would you recommend for SPAM detection? Justify your selection.

========== END OF HOMEWORK

In [ ]: