Week 2 Homework 261 Spring 2016 - Glenn Dunmire

Question 2.0:

What is a race condition in the context of parallel computation? Give an example.

A race condition is a situation where the final output of running a program depends on the sequence of events. That is to say, the final value may be different depending on the order in which steps are executed.

A classic example of a race condition is where two threads want to increase a variable. Ideally one thread would increment the variable, then the other thread would. So if the original value was 0, the final output would be 2. However, if the threads access the variable at the same time or without a lock, the result could be 1. This would be because the threads overwrite each other. So A increments 0 -> 1 but then B overwrites the variable with 1.

What is MapReduce? How does it differ from Hadoop?

Broadly speaking, MapReduce is a programming framework while Hadoop is an implementation of MapReduce. MapReduce is a model for processing large datasets using a parallel, distributed algorithm on a cluster. Hadoop is a specific implementation of MapReduce in Java, which uses a special distributed file system (HDFS) and manages aspects of workflow like distribution and fault tolerance.

Which programming paradigm is Hadoop based on? Explain and give a simple example in code and show the code running.

Hadoop is based on the paradigm of functional programming. This paradigm is based on the evaluation of mathematical functions and avoids changing state or mutable data. An important point is that a functional language is the concept of a function that can take other functions as an argument, also known as higher-order functions. Map and Reduce are examples of this higher order function.

In [1]:
```
#Example of a functional program: using map to print the lengths of
strings in a list

states = ["Maryland", "Virginia", "Pennsylvania"]
states_length = map(len, states)
print states_length
```

```
[8, 8, 12]
```

Notice here I am providing the map function with another function, len(). Also here I am not changing the values inside the list nor am I relying on anything other than the input list to produce my output.

Question 2.1:

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

In [1]:
```
#Write a text file of form <integer, "NA">.
#use the random package to get random numbers
import random

N = 10000 #set size of list of numbers

#I chose to only include a list from 0 to 10000 to make it easy to
check if the numbers were sorted properly.
numbers = random.sample(range(0, 10000), N) #list of numbers at ran
dom
output = [] #store output
for number in numbers:
    output.append('<' + str(number) + ', ' + 'NA>') #properly forma
t strings

with open('integer.txt', 'w') as myfile: #write output to a text fi
le
    myfile.write("\n".join(output))
```

In [2]:
```
%%writefile mapper.py
#!/usr/bin/python
import sys
# input comes from STDIN (standard input)
for line in sys.stdin:
    line = line[1:] #remove beginning '<'
    num = line.split()[0] #split on whitespace and only keep number
    num = num[:-1] #remove trailing comma
    print '%s\t%s' % (num, 'NA') #print result to STDOUT for input
to reducer
```

```
Overwriting mapper.py
```

In [3]:
```
%%writefile reducer.py
#!/usr/bin/python
import sys

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

    # parse the input we got from mapper.py
    num, na = line.split('\t')

    print '<' + num + ', NA>'
```

```
Overwriting reducer.py
```

In [5]:
```
#Test mapper and reducer
#!cat integer.txt | python mapper.py | sort -n | python reducer.py
```

In [6]:
```
#Start hadoop yarn
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-yarn.sh
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-dfs.sh
```

```
starting yarn daemons
starting resourcemanager, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/yarn-dunmireg-resourcemanager-Glenns-Air.hom
e.out
localhost: starting nodemanager, logging to /usr/local/Cellar/hado
op/2.7.1/libexec/logs/yarn-dunmireg-nodemanager-Glenns-Air.home.ou
t
16/01/25 14:47:33 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/hadoop-dunmireg-namenode-Glenns-Air.home.out
localhost: starting datanode, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/hadoop-dunmireg-datanode-Glenns-Air.home.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/Cella
r/hadoop/2.7.1/libexec/logs/hadoop-dunmireg-secondarynamenode-Glen
ns-Air.home.out
16/01/25 14:47:49 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [8]:
```
#make directory
#!hdfs dfs -mkdir -p /user/dunmireg
```

```
16/01/23 13:24:02 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [7]:
```
#add input to hdfs
!hdfs dfs -put integer.txt /user/dunmireg
```

```
16/01/25 14:47:51 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [8]:
```
!hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/too
ls/lib/hadoop-streaming-2.7.1.jar \
-D mapred.output.key.comparator.class=org.apache.hadoop.mapred.li
b.KeyFieldBasedComparator \
-D  mapred.text.key.comparator.options=-n \
-mapper mapper.py \
-reducer reducer.py \
-input integer.txt \
-output integerOutput
```

```
16/01/25 14:47:56 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 14:47:57 INFO Configuration.deprecation: session.id is de
precated. Instead, use dfs.metrics.session-id
16/01/25 14:47:57 INFO jvm.JvmMetrics: Initializing JVM Metrics wi
th processName=JobTracker, sessionId=
16/01/25 14:47:57 INFO jvm.JvmMetrics: Cannot initialize JVM Metri
cs with processName=JobTracker, sessionId= - already initialized
16/01/25 14:47:57 INFO mapred.FileInputFormat: Total input paths t
o process : 1
16/01/25 14:47:57 INFO mapreduce.JobSubmitter: number of splits:1
16/01/25 14:47:57 INFO Configuration.deprecation: mapred.text.ke
y.comparator.options is deprecated. Instead, use mapreduce.partiti
on.keycomparator.options
16/01/25 14:47:57 INFO Configuration.deprecation: mapred.output.ke
y.comparator.class is deprecated. Instead, use mapreduce.job.outpu
t.key.comparator.class
16/01/25 14:47:58 INFO mapreduce.JobSubmitter: Submitting tokens f
or job: job_local264777542_0001
16/01/25 14:47:58 INFO mapreduce.Job: The url to track the job: ht
tp://localhost:8080/
16/01/25 14:47:58 INFO mapred.LocalJobRunner: OutputCommitter set
in config null
16/01/25 14:47:58 INFO mapred.LocalJobRunner: OutputCommitter is o
rg.apache.hadoop.mapred.FileOutputCommitter
16/01/25 14:47:58 INFO mapreduce.Job: Running job: job_local264777
542_0001
16/01/25 14:47:58 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/25 14:47:58 INFO mapred.LocalJobRunner: Waiting for map task
s
16/01/25 14:47:58 INFO mapred.LocalJobRunner: Starting task: attem
pt_local264777542_0001_m_000000_0
16/01/25 14:47:58 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/25 14:47:58 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/01/25 14:47:58 INFO mapred.Task:  Using ResourceCalculatorProce
ssTree : null
16/01/25 14:47:58 INFO mapred.MapTask: Processing split: hdfs://lo
calhost:9000/user/dunmireg/integer.txt:0+108889
16/01/25 14:47:58 INFO mapred.MapTask: numReduceTasks: 1
16/01/25 14:47:58 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(10
4857584)
16/01/25 14:47:58 INFO mapred.MapTask: mapreduce.task.io.sort.mb:
100
16/01/25 14:47:58 INFO mapred.MapTask: soft limit at 83886080
16/01/25 14:47:58 INFO mapred.MapTask: bufstart = 0; bufvoid = 104
857600
16/01/25 14:47:58 INFO mapred.MapTask: kvstart = 26214396; length
= 6553600
16/01/25 14:47:58 INFO mapred.MapTask: Map output collector class
```

```
= org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/25 14:47:58 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW2/./mapper.py]
16/01/25 14:47:58 INFO Configuration.deprecation: mapred.tip.id is
deprecated. Instead, use mapreduce.task.id
16/01/25 14:47:58 INFO Configuration.deprecation: mapred.local.dir
is deprecated. Instead, use mapreduce.cluster.local.dir
16/01/25 14:47:58 INFO Configuration.deprecation: map.input.file i
s deprecated. Instead, use mapreduce.map.input.file
16/01/25 14:47:58 INFO Configuration.deprecation: mapred.skip.on i
s deprecated. Instead, use mapreduce.job.skiprecords
16/01/25 14:47:58 INFO Configuration.deprecation: map.input.length
is deprecated. Instead, use mapreduce.map.input.length
16/01/25 14:47:58 INFO Configuration.deprecation: mapred.work.outp
ut.dir is deprecated. Instead, use mapreduce.task.output.dir
16/01/25 14:47:58 INFO Configuration.deprecation: map.input.start
is deprecated. Instead, use mapreduce.map.input.start
16/01/25 14:47:58 INFO Configuration.deprecation: mapred.job.id is
deprecated. Instead, use mapreduce.job.id
16/01/25 14:47:58 INFO Configuration.deprecation: user.name is dep
recated. Instead, use mapreduce.job.user.name
16/01/25 14:47:58 INFO Configuration.deprecation: mapred.task.is.m
ap is deprecated. Instead, use mapreduce.task.ismap
16/01/25 14:47:58 INFO Configuration.deprecation: mapred.task.id i
s deprecated. Instead, use mapreduce.task.attempt.id
16/01/25 14:47:58 INFO Configuration.deprecation: mapred.task.part
ition is deprecated. Instead, use mapreduce.task.partition
16/01/25 14:47:58 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/25 14:47:58 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/25 14:47:58 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/25 14:47:58 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/25 14:47:58 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/25 14:47:58 INFO streaming.PipeMapRed: Records R/W=10000/1
16/01/25 14:47:58 INFO streaming.PipeMapRed: MRErrorThread done
16/01/25 14:47:58 INFO streaming.PipeMapRed: mapRedFinished
16/01/25 14:47:58 INFO mapred.LocalJobRunner:
16/01/25 14:47:58 INFO mapred.MapTask: Starting flush of map outpu
t
16/01/25 14:47:58 INFO mapred.MapTask: Spilling map output
16/01/25 14:47:58 INFO mapred.MapTask: bufstart = 0; bufend = 7889
0; bufvoid = 104857600
16/01/25 14:47:58 INFO mapred.MapTask: kvstart = 26214396(10485758
4); kvend = 26174400(104697600); length = 39997/6553600
16/01/25 14:47:58 INFO mapred.MapTask: Finished spill 0
16/01/25 14:47:58 INFO mapred.Task: Task:attempt_local264777542_00
01_m_000000_0 is done. And is in the process of committing
16/01/25 14:47:58 INFO mapred.LocalJobRunner: Records R/W=10000/1
16/01/25 14:47:58 INFO mapred.Task: Task 'attempt_local264777542_0
001_m_000000_0' done.
```

```
16/01/25 14:47:58 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local264777542_0001_m_000000_0
16/01/25 14:47:58 INFO mapred.LocalJobRunner: map task executor co
mplete.
16/01/25 14:47:58 INFO mapred.LocalJobRunner: Waiting for reduce t
asks
16/01/25 14:47:58 INFO mapred.LocalJobRunner: Starting task: attem
pt_local264777542_0001_r_000000_0
16/01/25 14:47:58 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/25 14:47:58 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/01/25 14:47:58 INFO mapred.Task:  Using ResourceCalculatorProce
ssTree : null
16/01/25 14:47:58 INFO mapred.ReduceTask: Using ShuffleConsumerPlu
gin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@46526d0d
16/01/25 14:47:58 INFO reduce.MergeManagerImpl: MergerManager: mem
oryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshol
d=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/01/25 14:47:58 INFO reduce.EventFetcher: attempt_local26477754
2_0001_r_000000_0 Thread started: EventFetcher for fetching Map Co
mpletion Events
16/01/25 14:47:58 INFO reduce.LocalFetcher: localfetcher#1 about t
o shuffle output of map attempt_local264777542_0001_m_000000_0 dec
omp: 98892 len: 98896 to MEMORY
16/01/25 14:47:59 INFO reduce.InMemoryMapOutput: Read 98892 bytes
from map-output for attempt_local264777542_0001_m_000000_0
16/01/25 14:47:59 INFO reduce.MergeManagerImpl: closeInMemoryFile
-> map-output of size: 98892, inMemoryMapOutputs.size() -> 1, comm
itMemory -> 0, usedMemory ->98892
16/01/25 14:47:59 INFO reduce.EventFetcher: EventFetcher is interr
upted.. Returning
16/01/25 14:47:59 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 14:47:59 INFO reduce.MergeManagerImpl: finalMerge called
with 1 in-memory map-outputs and 0 on-disk map-outputs
16/01/25 14:47:59 INFO mapred.Merger: Merging 1 sorted segments
16/01/25 14:47:59 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 98888 bytes
16/01/25 14:47:59 INFO reduce.MergeManagerImpl: Merged 1 segments,
98892 bytes to disk to satisfy reduce memory limit
16/01/25 14:47:59 INFO reduce.MergeManagerImpl: Merging 1 files, 9
8896 bytes from disk
16/01/25 14:47:59 INFO reduce.MergeManagerImpl: Merging 0 segment
s, 0 bytes from memory into reduce
16/01/25 14:47:59 INFO mapred.Merger: Merging 1 sorted segments
16/01/25 14:47:59 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 98888 bytes
16/01/25 14:47:59 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 14:47:59 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW2/./reducer.py]
16/01/25 14:47:59 INFO Configuration.deprecation: mapred.job.track
er is deprecated. Instead, use mapreduce.jobtracker.address
16/01/25 14:47:59 INFO Configuration.deprecation: mapred.map.tasks
is deprecated. Instead, use mapreduce.job.maps
```

```
16/01/25 14:47:59 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/25 14:47:59 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/25 14:47:59 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/25 14:47:59 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/25 14:47:59 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/25 14:47:59 INFO streaming.PipeMapRed: Records R/W=10000/1
16/01/25 14:47:59 INFO streaming.PipeMapRed: MRErrorThread done
16/01/25 14:47:59 INFO streaming.PipeMapRed: mapRedFinished
16/01/25 14:47:59 INFO mapreduce.Job: Job job_local264777542_0001
running in uber mode : false
16/01/25 14:47:59 INFO mapreduce.Job:  map 100% reduce 0%
16/01/25 14:47:59 INFO mapred.Task: Task:attempt_local264777542_00
01_r_000000_0 is done. And is in the process of committing
16/01/25 14:47:59 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 14:47:59 INFO mapred.Task: Task attempt_local264777542_00
01_r_000000_0 is allowed to commit now
16/01/25 14:47:59 INFO output.FileOutputCommitter: Saved output of
task 'attempt_local264777542_0001_r_000000_0' to hdfs://localhos
t:9000/user/dunmireg/integerOutput/_temporary/0/task_local26477754
2_0001_r_000000
16/01/25 14:47:59 INFO mapred.LocalJobRunner: Records R/W=10000/1
> reduce
16/01/25 14:47:59 INFO mapred.Task: Task 'attempt_local264777542_0
001_r_000000_0' done.
16/01/25 14:47:59 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local264777542_0001_r_000000_0
16/01/25 14:47:59 INFO mapred.LocalJobRunner: reduce task executor
complete.
16/01/25 14:48:00 INFO mapreduce.Job:  map 100% reduce 100%
16/01/25 14:48:00 INFO mapreduce.Job: Job job_local264777542_0001
completed successfully
16/01/25 14:48:00 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=409896
                FILE: Number of bytes written=1097568
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=217778
                HDFS: Number of bytes written=118890
                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=78890
                Map output materialized bytes=98896
                Input split bytes=99
```

```
                Combine input records=0
                Combine output records=0
                Reduce input groups=10000
                Reduce shuffle bytes=98896
                Reduce input records=10000
                Reduce output records=10000
                Spilled Records=20000
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=5
                Total committed heap usage (bytes)=546308096
        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=108889
        File Output Format Counters
                Bytes Written=118890
16/01/25 14:48:00 INFO streaming.StreamJob: Output directory: inte
gerOutput
```

In [27]:
```
#show results
#!hdfs dfs -cat /user/dunmireg/integerOutput/part-00000
```

In [9]:
```
#move output to local directory
!hadoop fs -copyToLocal /user/dunmireg/integerOutput
```

```
16/01/25 14:49:01 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 14:49:02 WARN hdfs.DFSClient: DFSInputStream has been clo
sed already
16/01/25 14:49:02 WARN hdfs.DFSClient: DFSInputStream has been clo
sed already
```

In [10]:
```
#Remove output directory and stop yarn and hdfs
!hadoop fs -rmr /user/dunmireg/integer.txt #check
!hadoop fs -rmr /user/dunmireg/integerOutput
!/usr/local/Cellar/hadoop/2.7.1/sbin/stop-yarn.sh
!/usr/local/Cellar/hadoop/2.7.1/sbin/stop-dfs.sh
```

```
rmr: DEPRECATED: Please use 'rm -r' instead.
16/01/25 14:49:05 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 14:49:06 INFO fs.TrashPolicyDefault: Namenode trash confi
guration: Deletion interval = 0 minutes, Emptier interval = 0 minu
tes.
Deleted /user/dunmireg/integer.txt
rmr: DEPRECATED: Please use 'rm -r' instead.
16/01/25 14:49:07 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 14:49:07 INFO fs.TrashPolicyDefault: Namenode trash confi
guration: Deletion interval = 0 minutes, Emptier interval = 0 minu
tes.
Deleted /user/dunmireg/integerOutput
stopping yarn daemons
stopping resourcemanager
localhost: stopping nodemanager
no proxyserver to stop
16/01/25 14:49:19 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here 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 14:49:37 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [14]:
```
#display output
import os

with open(os.path.join('./integerOutput', 'part-00000'), 'r') as myfile: #get appropriate output
    lines = myfile.readlines() #read in lines
    print "Smallest 10:"
    for i in range(10): #get smallest 10, the first 10 numbers
        line = lines[i] #get right line
        line = line[1:] #remove '<'
        num = line.split()[0] #split on whitespace, keeping number
        num = num[:-1] #remove comma
        print num

    print "Largest 10"
    for i in range(9990, 10000): #repeat above with different range
        line = lines[i]
        line = line[1:]
        num = line.split()[0]
        num = num[:-1]
        print num
```

```
Smallest 10:
0
1
2
3
4
5
6
7
8
9
Largest 10
9990
9991
9992
9993
9994
9995
9996
9997
9998
9999
```

If I were to have multiple reducers, yes there would need to be an additional step. In this case I would need to include a partitioner to distribute the output of the mapper to the different reducers. This ensures that the sorted output of map is distributed to the correct reducers. For example, if I had 10 inputs the output of map would get passed to a partitioner which would distribute keys to the 2 reducers I have.

Question 2.2

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 to
ken occur? 10 times! This is the number we are looking for!
```

In [3]:
```
%%writefile mapper.py
#!/usr/bin/python
import sys
import re
WORD_RE = re.compile(r"[\w']+") #regex for string matching

for line in sys.stdin: #for each line
    components = line.split('\t')
    text = " ".join(components[-2:]).strip() #get text of subject a
nd content
    words = re.findall(WORD_RE, text) #match all words
    for word in words:
        print word + '\t' + '1'
```

```
Overwriting mapper.py
```

In [5]:
```
%%writefile reducer.py
#!/usr/bin/python
#credit to Professor Shanahan for the structure of this reducer
import sys

current_word = None
current_count = 0
word = None

#lines come from standard input
for line in sys.stdin:
    line = line.strip()
    line = line.split('\t')
    word = line[0]
    count = int(line[1])

    if current_word == word:
        current_count += count
    else:
        if current_word:
            print '%s\t%s' % (current_word, current_count)
        current_word = word
        current_count = count
#print last line
if current_word == word:
    print '%s\t%s' % (current_word, current_count)
```

```
Overwriting reducer.py
```

In [10]:
```
#!cat enronemail_1h.txt | python mapper.py | sort | python reduce
r.py
#confirm assistance = 10
```

In [11]:
```
#Start hadoop yarn
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-yarn.sh
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-dfs.sh
```

```
starting yarn daemons
starting resourcemanager, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/yarn-dunmireg-resourcemanager-Glenns-Air.hom
e.out
localhost: starting nodemanager, logging to /usr/local/Cellar/hado
op/2.7.1/libexec/logs/yarn-dunmireg-nodemanager-Glenns-Air.home.ou
t
16/01/25 18:17:27 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/hadoop-dunmireg-namenode-Glenns-Air.home.out
localhost: starting datanode, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/hadoop-dunmireg-datanode-Glenns-Air.home.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/Cella
r/hadoop/2.7.1/libexec/logs/hadoop-dunmireg-secondarynamenode-Glen
ns-Air.home.out
16/01/25 18:17:42 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [12]:
```
#add input to hdfs
!hdfs dfs -put enronemail_1h.txt /user/dunmireg
```

```
16/01/25 18:17:46 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [13]:
```
!hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/too
ls/lib/hadoop-streaming-2.7.1.jar \
-mapper mapper.py \
-reducer reducer.py \
-input enronemail_1h.txt \
-output enroneWordCount
```

```
16/01/25 18:17:50 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 18:17:50 INFO Configuration.deprecation: session.id is de
precated. Instead, use dfs.metrics.session-id
16/01/25 18:17:50 INFO jvm.JvmMetrics: Initializing JVM Metrics wi
th processName=JobTracker, sessionId=
16/01/25 18:17:50 INFO jvm.JvmMetrics: Cannot initialize JVM Metri
cs with processName=JobTracker, sessionId= - already initialized
16/01/25 18:17:50 INFO mapred.FileInputFormat: Total input paths t
o process : 1
16/01/25 18:17:51 INFO mapreduce.JobSubmitter: number of splits:1
16/01/25 18:17:51 INFO mapreduce.JobSubmitter: Submitting tokens f
or job: job_local1364191356_0001
16/01/25 18:17:51 INFO mapreduce.Job: The url to track the job: ht
tp://localhost:8080/
16/01/25 18:17:51 INFO mapred.LocalJobRunner: OutputCommitter set
in config null
16/01/25 18:17:51 INFO mapred.LocalJobRunner: OutputCommitter is o
rg.apache.hadoop.mapred.FileOutputCommitter
16/01/25 18:17:51 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/25 18:17:51 INFO mapreduce.Job: Running job: job_local136419
1356_0001
16/01/25 18:17:51 INFO mapred.LocalJobRunner: Waiting for map task
s
16/01/25 18:17:51 INFO mapred.LocalJobRunner: Starting task: attem
pt_local1364191356_0001_m_000000_0
16/01/25 18:17:51 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/25 18:17:51 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/01/25 18:17:51 INFO mapred.Task:  Using ResourceCalculatorProce
ssTree : null
16/01/25 18:17:51 INFO mapred.MapTask: Processing split: hdfs://lo
calhost:9000/user/dunmireg/enronemail_1h.txt:0+204658
16/01/25 18:17:51 INFO mapred.MapTask: numReduceTasks: 1
16/01/25 18:17:51 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(10
4857584)
16/01/25 18:17:51 INFO mapred.MapTask: mapreduce.task.io.sort.mb:
100
16/01/25 18:17:51 INFO mapred.MapTask: soft limit at 83886080
16/01/25 18:17:51 INFO mapred.MapTask: bufstart = 0; bufvoid = 104
857600
16/01/25 18:17:51 INFO mapred.MapTask: kvstart = 26214396; length
= 6553600
16/01/25 18:17:51 INFO mapred.MapTask: Map output collector class
= org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/25 18:17:51 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW2/./mapper.py]
16/01/25 18:17:51 INFO Configuration.deprecation: mapred.tip.id is
deprecated. Instead, use mapreduce.task.id
16/01/25 18:17:51 INFO Configuration.deprecation: mapred.local.dir
```

```
is deprecated. Instead, use mapreduce.cluster.local.dir
16/01/25 18:17:51 INFO Configuration.deprecation: map.input.file i
s deprecated. Instead, use mapreduce.map.input.file
16/01/25 18:17:51 INFO Configuration.deprecation: mapred.skip.on i
s deprecated. Instead, use mapreduce.job.skiprecords
16/01/25 18:17:51 INFO Configuration.deprecation: map.input.length
is deprecated. Instead, use mapreduce.map.input.length
16/01/25 18:17:51 INFO Configuration.deprecation: mapred.work.outp
ut.dir is deprecated. Instead, use mapreduce.task.output.dir
16/01/25 18:17:51 INFO Configuration.deprecation: map.input.start
is deprecated. Instead, use mapreduce.map.input.start
16/01/25 18:17:51 INFO Configuration.deprecation: mapred.job.id is
deprecated. Instead, use mapreduce.job.id
16/01/25 18:17:51 INFO Configuration.deprecation: user.name is dep
recated. Instead, use mapreduce.job.user.name
16/01/25 18:17:51 INFO Configuration.deprecation: mapred.task.is.m
ap is deprecated. Instead, use mapreduce.task.ismap
16/01/25 18:17:51 INFO Configuration.deprecation: mapred.task.id i
s deprecated. Instead, use mapreduce.task.attempt.id
16/01/25 18:17:51 INFO Configuration.deprecation: mapred.task.part
ition is deprecated. Instead, use mapreduce.task.partition
16/01/25 18:17:51 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/25 18:17:51 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/25 18:17:51 INFO streaming.PipeMapRed: Records R/W=72/1
16/01/25 18:17:51 INFO streaming.PipeMapRed: R/W/S=100/13450/0 i
n:NA [rec/s] out:NA [rec/s]
16/01/25 18:17:51 INFO streaming.PipeMapRed: MRErrorThread done
16/01/25 18:17:51 INFO streaming.PipeMapRed: mapRedFinished
16/01/25 18:17:51 INFO mapred.LocalJobRunner:
16/01/25 18:17:51 INFO mapred.MapTask: Starting flush of map outpu
t
16/01/25 18:17:51 INFO mapred.MapTask: Spilling map output
16/01/25 18:17:51 INFO mapred.MapTask: bufstart = 0; bufend = 2522
11; bufvoid = 104857600
16/01/25 18:17:51 INFO mapred.MapTask: kvstart = 26214396(10485758
4); kvend = 26082748(104330992); length = 131649/6553600
16/01/25 18:17:51 INFO mapred.MapTask: Finished spill 0
16/01/25 18:17:51 INFO mapred.Task: Task:attempt_local1364191356_0
001_m_000000_0 is done. And is in the process of committing
16/01/25 18:17:51 INFO mapred.LocalJobRunner: Records R/W=72/1
16/01/25 18:17:51 INFO mapred.Task: Task 'attempt_local136419135
6_0001_m_000000_0' done.
16/01/25 18:17:51 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local1364191356_0001_m_000000_0
16/01/25 18:17:51 INFO mapred.LocalJobRunner: map task executor co
mplete.
16/01/25 18:17:51 INFO mapred.LocalJobRunner: Waiting for reduce t
asks
16/01/25 18:17:51 INFO mapred.LocalJobRunner: Starting task: attem
pt_local1364191356_0001_r_000000_0
16/01/25 18:17:51 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
```

```
16/01/25 18:17:51 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/01/25 18:17:51 INFO mapred.Task:  Using ResourceCalculatorProce
ssTree : null
16/01/25 18:17:52 INFO mapred.ReduceTask: Using ShuffleConsumerPlu
gin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@76742bea
16/01/25 18:17:52 INFO reduce.MergeManagerImpl: MergerManager: mem
oryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshol
d=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/01/25 18:17:52 INFO reduce.EventFetcher: attempt_local136419135
6_0001_r_000000_0 Thread started: EventFetcher for fetching Map Co
mpletion Events
16/01/25 18:17:52 INFO reduce.LocalFetcher: localfetcher#1 about t
o shuffle output of map attempt_local1364191356_0001_m_000000_0 de
comp: 318039 len: 318043 to MEMORY
16/01/25 18:17:52 INFO reduce.InMemoryMapOutput: Read 318039 bytes
from map-output for attempt_local1364191356_0001_m_000000_0
16/01/25 18:17:52 INFO reduce.MergeManagerImpl: closeInMemoryFile
-> map-output of size: 318039, inMemoryMapOutputs.size() -> 1, com
mitMemory -> 0, usedMemory ->318039
16/01/25 18:17:52 INFO reduce.EventFetcher: EventFetcher is interr
upted.. Returning
16/01/25 18:17:52 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 18:17:52 INFO reduce.MergeManagerImpl: finalMerge called
with 1 in-memory map-outputs and 0 on-disk map-outputs
16/01/25 18:17:52 INFO mapred.Merger: Merging 1 sorted segments
16/01/25 18:17:52 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 318035 bytes
16/01/25 18:17:52 INFO reduce.MergeManagerImpl: Merged 1 segments,
318039 bytes to disk to satisfy reduce memory limit
16/01/25 18:17:52 INFO reduce.MergeManagerImpl: Merging 1 files, 3
18043 bytes from disk
16/01/25 18:17:52 INFO reduce.MergeManagerImpl: Merging 0 segment
s, 0 bytes from memory into reduce
16/01/25 18:17:52 INFO mapred.Merger: Merging 1 sorted segments
16/01/25 18:17:52 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 318035 bytes
16/01/25 18:17:52 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 18:17:52 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW2/./reducer.py]
16/01/25 18:17:52 INFO Configuration.deprecation: mapred.job.track
er is deprecated. Instead, use mapreduce.jobtracker.address
16/01/25 18:17:52 INFO Configuration.deprecation: mapred.map.tasks
is deprecated. Instead, use mapreduce.job.maps
16/01/25 18:17:52 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/25 18:17:52 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/25 18:17:52 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/25 18:17:52 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/25 18:17:52 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
```

```
16/01/25 18:17:52 INFO streaming.PipeMapRed: Records R/W=21196/1
16/01/25 18:17:52 INFO streaming.PipeMapRed: MRErrorThread done
16/01/25 18:17:52 INFO streaming.PipeMapRed: mapRedFinished
16/01/25 18:17:52 INFO mapreduce.Job: Job job_local1364191356_0001
running in uber mode : false
16/01/25 18:17:52 INFO mapreduce.Job:  map 100% reduce 0%
16/01/25 18:17:52 INFO mapred.Task: Task:attempt_local1364191356_0
001_r_000000_0 is done. And is in the process of committing
16/01/25 18:17:52 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 18:17:52 INFO mapred.Task: Task attempt_local1364191356_0
001_r_000000_0 is allowed to commit now
16/01/25 18:17:52 INFO output.FileOutputCommitter: Saved output of
task 'attempt_local1364191356_0001_r_000000_0' to hdfs://localhos
t:9000/user/dunmireg/enroneWordCount/_temporary/0/task_local136419
1356_0001_r_000000
16/01/25 18:17:52 INFO mapred.LocalJobRunner: Records R/W=21196/1
> reduce
16/01/25 18:17:52 INFO mapred.Task: Task 'attempt_local136419135
6_0001_r_000000_0' done.
16/01/25 18:17:52 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local1364191356_0001_r_000000_0
16/01/25 18:17:52 INFO mapred.LocalJobRunner: reduce task executor
complete.
16/01/25 18:17:53 INFO mapreduce.Job:  map 100% reduce 100%
16/01/25 18:17:53 INFO mapreduce.Job: Job job_local1364191356_0001
completed successfully
16/01/25 18:17:53 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=848202
                FILE: Number of bytes written=1755937
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=409316
                HDFS: Number of bytes written=53488
                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=100
                Map output records=32913
                Map output bytes=252211
                Map output materialized bytes=318043
                Input split bytes=105
                Combine input records=0
                Combine output records=0
                Reduce input groups=5491
                Reduce shuffle bytes=318043
                Reduce input records=32913
                Reduce output records=5491
                Spilled Records=65826
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
```

```
                GC time elapsed (ms)=5
                Total committed heap usage (bytes)=510656512
        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=204658
        File Output Format Counters
                Bytes Written=53488
16/01/25 18:17:53 INFO streaming.StreamJob: Output directory: enro
neWordCount
```

In [15]:
```
#show results
#!hdfs dfs -cat /user/dunmireg/enroneWordCount/part-00000
```

In [16]:
```
#move output to local directory
#bin/hadoop fs -copyToLocal /hdfs/source/path /localfs/destinatio
n/path
!hadoop fs -copyToLocal /user/dunmireg/enroneWordCount/part-00000
```

```
16/01/25 18:18:13 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 18:18:13 WARN hdfs.DFSClient: DFSInputStream has been clo
sed already
```

In [17]:
```
#Remove output directory and stop yarn and hdfs
!hadoop fs -rmr /user/dunmireg/enronemail_1h.txt
!hadoop fs -rmr /user/dunmireg/enroneWordCount
!/usr/local/Cellar/hadoop/2.7.1/sbin/stop-yarn.sh
!/usr/local/Cellar/hadoop/2.7.1/sbin/stop-dfs.sh
```

```
rmr: DEPRECATED: Please use 'rm -r' instead.
16/01/25 18:18:21 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 18:18:22 INFO fs.TrashPolicyDefault: Namenode trash confi
guration: Deletion interval = 0 minutes, Emptier interval = 0 minu
tes.
Deleted /user/dunmireg/enronemail_1h.txt
rmr: DEPRECATED: Please use 'rm -r' instead.
16/01/25 18:18:23 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 18:18:23 INFO fs.TrashPolicyDefault: Namenode trash confi
guration: Deletion interval = 0 minutes, Emptier interval = 0 minu
tes.
Deleted /user/dunmireg/enroneWordCount
stopping yarn daemons
stopping resourcemanager
localhost: stopping nodemanager
no proxyserver to stop
16/01/25 18:18:35 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here 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 18:18:53 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [21]:
```
#Rename output file for convenience and print results of assistance
#import os
#os.rename('part-00000', 'wordCount') #only needs to run once

#open file and read contents
with open('wordCount', 'r') as myfile:
    lines = myfile.readlines()
    for line in lines:
        components = line.split('\t') #parse input
        if components[0] == 'assistance': #check if found the right
word
            print "Number of times assistance occurs: " + component
s[1] #print results
            break  #break loop
```

```
Number of times assistance occurs: 10
```

HW2.2.1

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

In [103]:
```
%%writefile mapper.py
#!/usr/bin/python
import sys
# input comes from STDIN (standard input)
for line in sys.stdin:
    components = line.split('\t')
    #reverse input, so instead of word, count it now becomes count,
word with count serving as key
    #note convert number to an int to remove new line character, th
en turn to string
    print components[1].rstrip() + '\t' + components[0] #print resu
lt to STDOUT for input to reducer
```

```
Overwriting mapper.py
```

In [104]:
```
%%writefile reducer.py
#!/usr/bin/python
import sys

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

    # parse the input we got from mapper.py
    line = line.split('\t')
    count = line[0]
    word = line[1].rstrip()

    #reverse order, relying on hadoop shuffling to get into proper
order
    print word + '\t' + count
```

```
Overwriting reducer.py
```

In [105]:
```
#Start hadoop yarn
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-yarn.sh
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-dfs.sh
```

```
starting yarn daemons
starting resourcemanager, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/yarn-dunmireg-resourcemanager-Glenns-Air.hom
e.out
localhost: starting nodemanager, logging to /usr/local/Cellar/hado
op/2.7.1/libexec/logs/yarn-dunmireg-nodemanager-Glenns-Air.home.ou
t
16/01/25 19:11:09 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/hadoop-dunmireg-namenode-Glenns-Air.home.out
localhost: starting datanode, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/hadoop-dunmireg-datanode-Glenns-Air.home.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/Cella
r/hadoop/2.7.1/libexec/logs/hadoop-dunmireg-secondarynamenode-Glen
ns-Air.home.out
16/01/25 19:11:25 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [106]:
```
#add input to hdfs
!hdfs dfs -put wordCount /user/dunmireg
```

```
16/01/25 19:11:44 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [107]:
```
!hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/too
ls/lib/hadoop-streaming-2.7.1.jar \
-D mapred.output.key.comparator.class=org.apache.hadoop.mapred.li
b.KeyFieldBasedComparator \
-D  mapred.text.key.comparator.options=-n \
-mapper mapper.py \
-reducer reducer.py \
-input wordCount \
-output sortedWordCount
```

```
16/01/25 19:11:47 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 19:11:47 INFO Configuration.deprecation: session.id is de
precated. Instead, use dfs.metrics.session-id
16/01/25 19:11:47 INFO jvm.JvmMetrics: Initializing JVM Metrics wi
th processName=JobTracker, sessionId=
16/01/25 19:11:47 INFO jvm.JvmMetrics: Cannot initialize JVM Metri
cs with processName=JobTracker, sessionId= - already initialized
16/01/25 19:11:47 INFO mapred.FileInputFormat: Total input paths t
o process : 1
16/01/25 19:11:47 INFO mapreduce.JobSubmitter: number of splits:1
16/01/25 19:11:48 INFO Configuration.deprecation: mapred.text.ke
y.comparator.options is deprecated. Instead, use mapreduce.partiti
on.keycomparator.options
16/01/25 19:11:48 INFO Configuration.deprecation: mapred.output.ke
y.comparator.class is deprecated. Instead, use mapreduce.job.outpu
t.key.comparator.class
16/01/25 19:11:48 INFO mapreduce.JobSubmitter: Submitting tokens f
or job: job_local327801418_0001
16/01/25 19:11:48 INFO mapreduce.Job: The url to track the job: ht
tp://localhost:8080/
16/01/25 19:11:48 INFO mapred.LocalJobRunner: OutputCommitter set
in config null
16/01/25 19:11:48 INFO mapreduce.Job: Running job: job_local327801
418_0001
16/01/25 19:11:48 INFO mapred.LocalJobRunner: OutputCommitter is o
rg.apache.hadoop.mapred.FileOutputCommitter
16/01/25 19:11:48 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/25 19:11:48 INFO mapred.LocalJobRunner: Waiting for map task
s
16/01/25 19:11:48 INFO mapred.LocalJobRunner: Starting task: attem
pt_local327801418_0001_m_000000_0
16/01/25 19:11:48 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/25 19:11:48 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/01/25 19:11:48 INFO mapred.Task:  Using ResourceCalculatorProce
ssTree : null
16/01/25 19:11:48 INFO mapred.MapTask: Processing split: hdfs://lo
calhost:9000/user/dunmireg/wordCount:0+53488
16/01/25 19:11:48 INFO mapred.MapTask: numReduceTasks: 1
16/01/25 19:11:48 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(10
4857584)
16/01/25 19:11:48 INFO mapred.MapTask: mapreduce.task.io.sort.mb:
100
16/01/25 19:11:48 INFO mapred.MapTask: soft limit at 83886080
16/01/25 19:11:48 INFO mapred.MapTask: bufstart = 0; bufvoid = 104
857600
16/01/25 19:11:48 INFO mapred.MapTask: kvstart = 26214396; length
= 6553600
16/01/25 19:11:48 INFO mapred.MapTask: Map output collector class
```

```
= org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/25 19:11:48 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW2/./mapper.py]
16/01/25 19:11:48 INFO Configuration.deprecation: mapred.tip.id is
deprecated. Instead, use mapreduce.task.id
16/01/25 19:11:48 INFO Configuration.deprecation: mapred.local.dir
is deprecated. Instead, use mapreduce.cluster.local.dir
16/01/25 19:11:48 INFO Configuration.deprecation: map.input.file i
s deprecated. Instead, use mapreduce.map.input.file
16/01/25 19:11:48 INFO Configuration.deprecation: mapred.skip.on i
s deprecated. Instead, use mapreduce.job.skiprecords
16/01/25 19:11:48 INFO Configuration.deprecation: map.input.length
is deprecated. Instead, use mapreduce.map.input.length
16/01/25 19:11:48 INFO Configuration.deprecation: mapred.work.outp
ut.dir is deprecated. Instead, use mapreduce.task.output.dir
16/01/25 19:11:48 INFO Configuration.deprecation: map.input.start
is deprecated. Instead, use mapreduce.map.input.start
16/01/25 19:11:48 INFO Configuration.deprecation: mapred.job.id is
deprecated. Instead, use mapreduce.job.id
16/01/25 19:11:48 INFO Configuration.deprecation: user.name is dep
recated. Instead, use mapreduce.job.user.name
16/01/25 19:11:48 INFO Configuration.deprecation: mapred.task.is.m
ap is deprecated. Instead, use mapreduce.task.ismap
16/01/25 19:11:48 INFO Configuration.deprecation: mapred.task.id i
s deprecated. Instead, use mapreduce.task.attempt.id
16/01/25 19:11:48 INFO Configuration.deprecation: mapred.task.part
ition is deprecated. Instead, use mapreduce.task.partition
16/01/25 19:11:48 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/25 19:11:48 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/25 19:11:48 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/25 19:11:48 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/25 19:11:48 INFO streaming.PipeMapRed: Records R/W=5491/1
16/01/25 19:11:48 INFO streaming.PipeMapRed: MRErrorThread done
16/01/25 19:11:48 INFO streaming.PipeMapRed: mapRedFinished
16/01/25 19:11:48 INFO mapred.LocalJobRunner:
16/01/25 19:11:48 INFO mapred.MapTask: Starting flush of map outpu
t
16/01/25 19:11:48 INFO mapred.MapTask: Spilling map output
16/01/25 19:11:48 INFO mapred.MapTask: bufstart = 0; bufend = 5348
8; bufvoid = 104857600
16/01/25 19:11:48 INFO mapred.MapTask: kvstart = 26214396(10485758
4); kvend = 26192436(104769744); length = 21961/6553600
16/01/25 19:11:48 INFO mapred.MapTask: Finished spill 0
16/01/25 19:11:48 INFO mapred.Task: Task:attempt_local327801418_00
01_m_000000_0 is done. And is in the process of committing
16/01/25 19:11:48 INFO mapred.LocalJobRunner: Records R/W=5491/1
16/01/25 19:11:48 INFO mapred.Task: Task 'attempt_local327801418_0
001_m_000000_0' done.
16/01/25 19:11:48 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local327801418_0001_m_000000_0
```

```
16/01/25 19:11:48 INFO mapred.LocalJobRunner: map task executor co
mplete.
16/01/25 19:11:48 INFO mapred.LocalJobRunner: Waiting for reduce t
asks
16/01/25 19:11:48 INFO mapred.LocalJobRunner: Starting task: attem
pt_local327801418_0001_r_000000_0
16/01/25 19:11:48 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/25 19:11:48 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/01/25 19:11:48 INFO mapred.Task:  Using ResourceCalculatorProce
ssTree : null
16/01/25 19:11:48 INFO mapred.ReduceTask: Using ShuffleConsumerPlu
gin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@68c5d7a0
16/01/25 19:11:48 INFO reduce.MergeManagerImpl: MergerManager: mem
oryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshol
d=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/01/25 19:11:48 INFO reduce.EventFetcher: attempt_local32780141
8_0001_r_000000_0 Thread started: EventFetcher for fetching Map Co
mpletion Events
16/01/25 19:11:48 INFO reduce.LocalFetcher: localfetcher#1 about t
o shuffle output of map attempt_local327801418_0001_m_000000_0 dec
omp: 64472 len: 64476 to MEMORY
16/01/25 19:11:48 INFO reduce.InMemoryMapOutput: Read 64472 bytes
from map-output for attempt_local327801418_0001_m_000000_0
16/01/25 19:11:48 INFO reduce.MergeManagerImpl: closeInMemoryFile
-> map-output of size: 64472, inMemoryMapOutputs.size() -> 1, comm
itMemory -> 0, usedMemory ->64472
16/01/25 19:11:48 INFO reduce.EventFetcher: EventFetcher is interr
upted.. Returning
16/01/25 19:11:48 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 19:11:48 INFO reduce.MergeManagerImpl: finalMerge called
with 1 in-memory map-outputs and 0 on-disk map-outputs
16/01/25 19:11:48 INFO mapred.Merger: Merging 1 sorted segments
16/01/25 19:11:48 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 64468 bytes
16/01/25 19:11:48 INFO reduce.MergeManagerImpl: Merged 1 segments,
64472 bytes to disk to satisfy reduce memory limit
16/01/25 19:11:48 INFO reduce.MergeManagerImpl: Merging 1 files, 6
4476 bytes from disk
16/01/25 19:11:48 INFO reduce.MergeManagerImpl: Merging 0 segment
s, 0 bytes from memory into reduce
16/01/25 19:11:48 INFO mapred.Merger: Merging 1 sorted segments
16/01/25 19:11:48 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 64468 bytes
16/01/25 19:11:48 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 19:11:48 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW2/./reducer.py]
16/01/25 19:11:48 INFO Configuration.deprecation: mapred.job.track
er is deprecated. Instead, use mapreduce.jobtracker.address
16/01/25 19:11:48 INFO Configuration.deprecation: mapred.map.tasks
is deprecated. Instead, use mapreduce.job.maps
16/01/25 19:11:48 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
```

```
16/01/25 19:11:48 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/25 19:11:48 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/25 19:11:48 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/25 19:11:48 INFO streaming.PipeMapRed: Records R/W=5491/1
16/01/25 19:11:48 INFO streaming.PipeMapRed: MRErrorThread done
16/01/25 19:11:48 INFO streaming.PipeMapRed: mapRedFinished
16/01/25 19:11:49 INFO mapred.Task: Task:attempt_local327801418_00
01_r_000000_0 is done. And is in the process of committing
16/01/25 19:11:49 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 19:11:49 INFO mapred.Task: Task attempt_local327801418_00
01_r_000000_0 is allowed to commit now
16/01/25 19:11:49 INFO output.FileOutputCommitter: Saved output of
task 'attempt_local327801418_0001_r_000000_0' to hdfs://localhos
t:9000/user/dunmireg/sortedWordCount/_temporary/0/task_local327801
418_0001_r_000000
16/01/25 19:11:49 INFO mapred.LocalJobRunner: Records R/W=5491/1 >
reduce
16/01/25 19:11:49 INFO mapred.Task: Task 'attempt_local327801418_0
001_r_000000_0' done.
16/01/25 19:11:49 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local327801418_0001_r_000000_0
16/01/25 19:11:49 INFO mapred.LocalJobRunner: reduce task executor
complete.
16/01/25 19:11:49 INFO mapreduce.Job: Job job_local327801418_0001
running in uber mode : false
16/01/25 19:11:49 INFO mapreduce.Job:  map 100% reduce 100%
16/01/25 19:11:49 INFO mapreduce.Job: Job job_local327801418_0001
completed successfully
16/01/25 19:11:49 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=341050
                FILE: Number of bytes written=994302
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=106976
                HDFS: Number of bytes written=53488
                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=5491
                Map output records=5491
                Map output bytes=53488
                Map output materialized bytes=64476
                Input split bytes=97
                Combine input records=0
                Combine output records=0
                Reduce input groups=107
                Reduce shuffle bytes=64476
                Reduce input records=5491
```

```
                Reduce output records=5491
                Spilled Records=10982
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=6
                Total committed heap usage (bytes)=542113792
        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=53488
        File Output Format Counters
                Bytes Written=53488
16/01/25 19:11:49 INFO streaming.StreamJob: Output directory: sort
edWordCount
```

In [81]:
```
#show results
#!hdfs dfs -cat /user/dunmireg/sortedWordCount/part-00000
```

In [108]:
```
#move output to local directory
!hadoop fs -copyToLocal /user/dunmireg/sortedWordCount
```

```
16/01/25 19:11:57 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 19:11:57 WARN hdfs.DFSClient: DFSInputStream has been clo
sed already
16/01/25 19:11:57 WARN hdfs.DFSClient: DFSInputStream has been clo
sed already
```

In [109]:
```
#Remove output directory and stop yarn and hdfs
!hadoop fs -rmr /user/dunmireg/wordCount
!hadoop fs -rmr /user/dunmireg/sortedWordCount
!/usr/local/Cellar/hadoop/2.7.1/sbin/stop-yarn.sh
!/usr/local/Cellar/hadoop/2.7.1/sbin/stop-dfs.sh
```

```
rmr: DEPRECATED: Please use 'rm -r' instead.
16/01/25 19:11:59 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 19:12:00 INFO fs.TrashPolicyDefault: Namenode trash confi
guration: Deletion interval = 0 minutes, Emptier interval = 0 minu
tes.
Deleted /user/dunmireg/wordCount
rmr: DEPRECATED: Please use 'rm -r' instead.
16/01/25 19:12:01 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 19:12:01 INFO fs.TrashPolicyDefault: Namenode trash confi
guration: Deletion interval = 0 minutes, Emptier interval = 0 minu
tes.
Deleted /user/dunmireg/sortedWordCount
stopping yarn daemons
stopping resourcemanager
localhost: stopping nodemanager
no proxyserver to stop
16/01/25 19:12:13 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here 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 19:12:31 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [111]:
```
#Display outputs from results file
import os

with open(os.path.join('./sortedWordCount', 'part-00000'), 'r') as
myfile:
    lines = myfile.readlines() #read file
    lines = lines[-10:] #get last 10 lines
    for line in lines:
        print line #print results
```

```
for     373

ect     382

your    394

in      417

you     432

a       542

of      566

and     668

to      963

the     1247
```

Above you can see the top 10 most frequently occurring words. This would make sense, most of these words are extremely common.

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 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"

In [1]:
```
%%writefile mapper.py
#!/usr/bin/python

import sys
import re
WORD_RE = re.compile(r"[\w']+") #regex for string matching

for line in sys.stdin: #for each line
    line = line.strip()
    line = line.rstrip()
    components = line.split('\t')
    text = " ".join(components[-2:]).strip() #get text of subject a
nd content
    words = re.findall(WORD_RE, text) #match all words
    for word in words:
        print components[0] + '\t' + word + '\t' + components[1] #p
rint email ID + word + spam flag
```

```
Overwriting mapper.py
```

In [2]:
```
!chmod a+x mapper.py
```

In [2]:
```
%%writefile reducer.py
#!/usr/bin/python
import sys

emails = set() #hold email IDs
words = {} #hold words and associated counts
spam_emails = 0 #how many emails are marked as spam
spam_word_count = 0 #how many words appear in spam
ham_word_count = 0 #how many words appear in ham

for line in sys.stdin:
    line = line.strip()
    line = line.rstrip()
    components = line.split('\t') #split input

    ID = components[0] #put input variables into fields to make eas
ier
    word = components[1]
    spam = int(components[2])

    if word not in words.keys(): #if a word is not in the words dic
tionary, add it and initialize counts to 0
        words[word] = {'spam_count': 0, 'ham_count': 0}
    if ID not in emails: #add email to set to store unique IDs
        emails.add(ID)
        if spam == 1: #increment spam counter
            spam_emails += 1


    if spam == 1: #if the flag if spam, increment the word spam_cou
nt value by 1, else do the same for ham
        words[word]['spam_count'] += 1
        spam_word_count += 1
    else:
        words[word]['ham_count'] += 1
        ham_word_count += 1


prior_spam = float(spam_emails)/len(emails) #get prior probabilitie
s
prior_ham = 1-prior_spam

for i, word in words.iteritems(): #calculate conditional probabilit
ies: number of times word appears in class/number of words in class
    word['spam_like'] = float(word['spam_count'])/(spam_word_count)
    word['ham_like'] = float(word['ham_count'])/(ham_word_count)


print prior_spam #print priors
print prior_ham
for word in words.keys():
    #Word "\t" spam likelihood '\t' ham likelihood written to file
    print word + '\t' + str(words[word]['spam_like']) + '\t' + st
```

```
r(words[word]['ham_like']) #print each word along with spam and ham
conditional probabilities
```

```
Overwriting reducer.py
```

In [4]:
```
!chmod a+x reducer.py
```

In [60]:
```
#Examine output
#!cat enronemail_1h.txt | python mapper.py | python reducer.py
```

In [3]:
```
#Start hadoop yarn
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-yarn.sh
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-dfs.sh
```

```
starting yarn daemons
starting resourcemanager, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/yarn-dunmireg-resourcemanager-Glenns-Air.hom
e.out
localhost: starting nodemanager, logging to /usr/local/Cellar/hado
op/2.7.1/libexec/logs/yarn-dunmireg-nodemanager-Glenns-Air.home.ou
t
16/01/25 21:16:06 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/hadoop-dunmireg-namenode-Glenns-Air.home.out
localhost: starting datanode, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/hadoop-dunmireg-datanode-Glenns-Air.home.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/Cella
r/hadoop/2.7.1/libexec/logs/hadoop-dunmireg-secondarynamenode-Glen
ns-Air.home.out
16/01/25 21:16:22 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [6]:
```
#make directory
#!hdfs dfs -mkdir -p /user/dunmireg
```

```
16/01/25 00:14:43 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [4]:
```
#add input to hdfs
!hdfs dfs -put enronemail_1h.txt /user/dunmireg
```

```
16/01/25 21:16:28 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [5]:
```
!hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/too
ls/lib/hadoop-streaming-2.7.1.jar \
-mapper mapper.py \
-reducer reducer.py \
-input enronemail_1h.txt \
-output enroneEmailCondProbs
```

```
16/01/25 21:16:31 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:16:32 INFO Configuration.deprecation: session.id is de
precated. Instead, use dfs.metrics.session-id
16/01/25 21:16:32 INFO jvm.JvmMetrics: Initializing JVM Metrics wi
th processName=JobTracker, sessionId=
16/01/25 21:16:32 INFO jvm.JvmMetrics: Cannot initialize JVM Metri
cs with processName=JobTracker, sessionId= - already initialized
16/01/25 21:16:32 INFO mapred.FileInputFormat: Total input paths t
o process : 1
16/01/25 21:16:32 INFO mapreduce.JobSubmitter: number of splits:1
16/01/25 21:16:32 INFO mapreduce.JobSubmitter: Submitting tokens f
or job: job_local1443500456_0001
16/01/25 21:16:32 INFO mapreduce.Job: The url to track the job: ht
tp://localhost:8080/
16/01/25 21:16:32 INFO mapred.LocalJobRunner: OutputCommitter set
in config null
16/01/25 21:16:32 INFO mapred.LocalJobRunner: OutputCommitter is o
rg.apache.hadoop.mapred.FileOutputCommitter
16/01/25 21:16:32 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/25 21:16:32 INFO mapreduce.Job: Running job: job_local144350
0456_0001
16/01/25 21:16:32 INFO mapred.LocalJobRunner: Waiting for map task
s
16/01/25 21:16:32 INFO mapred.LocalJobRunner: Starting task: attem
pt_local1443500456_0001_m_000000_0
16/01/25 21:16:33 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/25 21:16:33 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/01/25 21:16:33 INFO mapred.Task:  Using ResourceCalculatorProce
ssTree : null
16/01/25 21:16:33 INFO mapred.MapTask: Processing split: hdfs://lo
calhost:9000/user/dunmireg/enronemail_1h.txt:0+204658
16/01/25 21:16:33 INFO mapred.MapTask: numReduceTasks: 1
16/01/25 21:16:33 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(10
4857584)
16/01/25 21:16:33 INFO mapred.MapTask: mapreduce.task.io.sort.mb:
100
16/01/25 21:16:33 INFO mapred.MapTask: soft limit at 83886080
16/01/25 21:16:33 INFO mapred.MapTask: bufstart = 0; bufvoid = 104
857600
16/01/25 21:16:33 INFO mapred.MapTask: kvstart = 26214396; length
= 6553600
16/01/25 21:16:33 INFO mapred.MapTask: Map output collector class
= org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/25 21:16:33 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW2/./mapper.py]
16/01/25 21:16:33 INFO Configuration.deprecation: mapred.tip.id is
deprecated. Instead, use mapreduce.task.id
16/01/25 21:16:33 INFO Configuration.deprecation: mapred.local.dir
```

```
is deprecated. Instead, use mapreduce.cluster.local.dir
16/01/25 21:16:33 INFO Configuration.deprecation: map.input.file i
s deprecated. Instead, use mapreduce.map.input.file
16/01/25 21:16:33 INFO Configuration.deprecation: mapred.skip.on i
s deprecated. Instead, use mapreduce.job.skiprecords
16/01/25 21:16:33 INFO Configuration.deprecation: map.input.length
is deprecated. Instead, use mapreduce.map.input.length
16/01/25 21:16:33 INFO Configuration.deprecation: mapred.work.outp
ut.dir is deprecated. Instead, use mapreduce.task.output.dir
16/01/25 21:16:33 INFO Configuration.deprecation: map.input.start
is deprecated. Instead, use mapreduce.map.input.start
16/01/25 21:16:33 INFO Configuration.deprecation: mapred.job.id is
deprecated. Instead, use mapreduce.job.id
16/01/25 21:16:33 INFO Configuration.deprecation: user.name is dep
recated. Instead, use mapreduce.job.user.name
16/01/25 21:16:33 INFO Configuration.deprecation: mapred.task.is.m
ap is deprecated. Instead, use mapreduce.task.ismap
16/01/25 21:16:33 INFO Configuration.deprecation: mapred.task.id i
s deprecated. Instead, use mapreduce.task.attempt.id
16/01/25 21:16:33 INFO Configuration.deprecation: mapred.task.part
ition is deprecated. Instead, use mapreduce.task.partition
16/01/25 21:16:33 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/25 21:16:33 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/25 21:16:33 INFO streaming.PipeMapRed: Records R/W=72/1
16/01/25 21:16:33 INFO streaming.PipeMapRed: R/W/S=100/19097/0 i
n:NA [rec/s] out:NA [rec/s]
16/01/25 21:16:33 INFO streaming.PipeMapRed: MRErrorThread done
16/01/25 21:16:33 INFO streaming.PipeMapRed: mapRedFinished
16/01/25 21:16:33 INFO mapred.LocalJobRunner:
16/01/25 21:16:33 INFO mapred.MapTask: Starting flush of map outpu
t
16/01/25 21:16:33 INFO mapred.MapTask: Spilling map output
16/01/25 21:16:33 INFO mapred.MapTask: bufstart = 0; bufend = 1032
108; bufvoid = 104857600
16/01/25 21:16:33 INFO mapred.MapTask: kvstart = 26214396(10485758
4); kvend = 26082748(104330992); length = 131649/6553600
16/01/25 21:16:33 INFO mapred.MapTask: Finished spill 0
16/01/25 21:16:33 INFO mapred.Task: Task:attempt_local1443500456_0
001_m_000000_0 is done. And is in the process of committing
16/01/25 21:16:33 INFO mapred.LocalJobRunner: Records R/W=72/1
16/01/25 21:16:33 INFO mapred.Task: Task 'attempt_local144350045
6_0001_m_000000_0' done.
16/01/25 21:16:33 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local1443500456_0001_m_000000_0
16/01/25 21:16:33 INFO mapred.LocalJobRunner: map task executor co
mplete.
16/01/25 21:16:33 INFO mapred.LocalJobRunner: Waiting for reduce t
asks
16/01/25 21:16:33 INFO mapred.LocalJobRunner: Starting task: attem
pt_local1443500456_0001_r_000000_0
16/01/25 21:16:33 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
```

```
16/01/25 21:16:33 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/01/25 21:16:33 INFO mapred.Task:  Using ResourceCalculatorProce
ssTree : null
16/01/25 21:16:33 INFO mapred.ReduceTask: Using ShuffleConsumerPlu
gin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@d7d5f7
16/01/25 21:16:33 INFO reduce.MergeManagerImpl: MergerManager: mem
oryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshol
d=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/01/25 21:16:33 INFO reduce.EventFetcher: attempt_local144350045
6_0001_r_000000_0 Thread started: EventFetcher for fetching Map Co
mpletion Events
16/01/25 21:16:33 INFO reduce.LocalFetcher: localfetcher#1 about t
o shuffle output of map attempt_local1443500456_0001_m_000000_0 de
comp: 1097936 len: 1097940 to MEMORY
16/01/25 21:16:33 INFO reduce.InMemoryMapOutput: Read 1097936 byte
s from map-output for attempt_local1443500456_0001_m_000000_0
16/01/25 21:16:33 INFO reduce.MergeManagerImpl: closeInMemoryFile
-> map-output of size: 1097936, inMemoryMapOutputs.size() -> 1, co
mmitMemory -> 0, usedMemory ->1097936
16/01/25 21:16:33 INFO reduce.EventFetcher: EventFetcher is interr
upted.. Returning
16/01/25 21:16:33 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 21:16:33 INFO reduce.MergeManagerImpl: finalMerge called
with 1 in-memory map-outputs and 0 on-disk map-outputs
16/01/25 21:16:33 INFO mapred.Merger: Merging 1 sorted segments
16/01/25 21:16:33 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 1097911 bytes
16/01/25 21:16:33 INFO reduce.MergeManagerImpl: Merged 1 segments,
1097936 bytes to disk to satisfy reduce memory limit
16/01/25 21:16:33 INFO reduce.MergeManagerImpl: Merging 1 files, 1
097940 bytes from disk
16/01/25 21:16:33 INFO reduce.MergeManagerImpl: Merging 0 segment
s, 0 bytes from memory into reduce
16/01/25 21:16:33 INFO mapred.Merger: Merging 1 sorted segments
16/01/25 21:16:33 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 1097911 bytes
16/01/25 21:16:33 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 21:16:33 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW2/./reducer.py]
16/01/25 21:16:33 INFO Configuration.deprecation: mapred.job.track
er is deprecated. Instead, use mapreduce.jobtracker.address
16/01/25 21:16:33 INFO Configuration.deprecation: mapred.map.tasks
is deprecated. Instead, use mapreduce.job.maps
16/01/25 21:16:33 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/25 21:16:33 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/25 21:16:33 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/25 21:16:33 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/25 21:16:33 INFO mapreduce.Job: Job job_local1443500456_0001
running in uber mode : false
```

```
16/01/25 21:16:33 INFO mapreduce.Job:  map 100% reduce 0%
16/01/25 21:16:34 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/25 21:16:36 INFO streaming.PipeMapRed: Records R/W=32913/1
16/01/25 21:16:36 INFO streaming.PipeMapRed: MRErrorThread done
16/01/25 21:16:36 INFO streaming.PipeMapRed: mapRedFinished
16/01/25 21:16:36 INFO mapred.Task: Task:attempt_local1443500456_0
001_r_000000_0 is done. And is in the process of committing
16/01/25 21:16:36 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 21:16:36 INFO mapred.Task: Task attempt_local1443500456_0
001_r_000000_0 is allowed to commit now
16/01/25 21:16:36 INFO output.FileOutputCommitter: Saved output of
task 'attempt_local1443500456_0001_r_000000_0' to hdfs://localhos
t:9000/user/dunmireg/enroneEmailCondProbs/_temporary/0/task_local1
443500456_0001_r_000000
16/01/25 21:16:36 INFO mapred.LocalJobRunner: Records R/W=32913/1
> reduce
16/01/25 21:16:36 INFO mapred.Task: Task 'attempt_local144350045
6_0001_r_000000_0' done.
16/01/25 21:16:36 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local1443500456_0001_r_000000_0
16/01/25 21:16:36 INFO mapred.LocalJobRunner: reduce task executor
complete.
16/01/25 21:16:36 INFO mapreduce.Job:  map 100% reduce 100%
16/01/25 21:16:36 INFO mapreduce.Job: Job job_local1443500456_0001
completed successfully
16/01/25 21:16:36 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=2407996
                FILE: Number of bytes written=4095648
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=409316
                HDFS: Number of bytes written=172513
                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=100
                Map output records=32913
                Map output bytes=1032108
                Map output materialized bytes=1097940
                Input split bytes=105
                Combine input records=0
                Combine output records=0
                Reduce input groups=100
                Reduce shuffle bytes=1097940
                Reduce input records=32913
                Reduce output records=5493
                Spilled Records=65826
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
```

```
                GC time elapsed (ms)=5
                Total committed heap usage (bytes)=509607936
        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=204658
        File Output Format Counters
                Bytes Written=172513
16/01/25 21:16:36 INFO streaming.StreamJob: Output directory: enro
neEmailCondProbs
```

In [31]:
```
#show results
#!hdfs dfs -cat /user/dunmireg/enroneEmailCondProbs/part-00000
```

In [6]:
```
#move output to local directory
#bin/hadoop fs -copyToLocal /hdfs/source/path /localfs/destinatio
n/path
!hadoop fs -copyToLocal /user/dunmireg/enroneEmailCondProbs
```

```
16/01/25 21:17:25 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:17:25 WARN hdfs.DFSClient: DFSInputStream has been clo
sed already
16/01/25 21:17:25 WARN hdfs.DFSClient: DFSInputStream has been clo
sed already
```

In [7]:
```
#Remove output directory and stop yarn and hdfs
!hadoop fs -rmr /user/dunmireg/enronemail_1h.txt #check
!hadoop fs -rmr /user/dunmireg/enroneEmailCondProbs
!/usr/local/Cellar/hadoop/2.7.1/sbin/stop-yarn.sh
!/usr/local/Cellar/hadoop/2.7.1/sbin/stop-dfs.sh
```

```
rmr: DEPRECATED: Please use 'rm -r' instead.
16/01/25 21:17:28 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:17:28 INFO fs.TrashPolicyDefault: Namenode trash confi
guration: Deletion interval = 0 minutes, Emptier interval = 0 minu
tes.
Deleted /user/dunmireg/enronemail_1h.txt
rmr: DEPRECATED: Please use 'rm -r' instead.
16/01/25 21:17:29 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:17:30 INFO fs.TrashPolicyDefault: Namenode trash confi
guration: Deletion interval = 0 minutes, Emptier interval = 0 minu
tes.
Deleted /user/dunmireg/enroneEmailCondProbs
stopping yarn daemons
stopping resourcemanager
localhost: stopping nodemanager
localhost: nodemanager did not stop gracefully after 5 seconds: ki
lling with kill -9
no proxyserver to stop
16/01/25 21:17:42 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here 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 21:18:00 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [8]:
```
%%writefile mapper.py
#!/usr/bin/python

#The bulk of the classification work happens in the mapper. This wi
ll read in the file of conditional probabilities
#and then compute the conditional probability of each word and perf
orms the classification. The classification output
#is then sent to the reducer
import sys
import re
import os
from math import log
from math import exp

priorSpam = 0 #prior probabilities from file
priorHam = 0
words = {} #dictionary to hold word conditional probabilities

with open(os.path.join('./enroneEmailCondProbs', 'part-00000'),
'r') as myfile: #read file
    lines = myfile.readlines()
    priorSpam = float(lines[0]) #grab prior probabilities
    priorHam = float(lines[1])
    for line in lines[2:]:#parse lines for word with conditional pr
obabilities
        line = line.strip()
        line = line.rstrip()
        components = line.split('\t')
        words[components[0]] = {'spam_like': float(components[1]),
'ham_like': float(components[2])} #remove new line


WORD_RE = re.compile(r"[\w']+")

spamSkip = 0 #how many times did a skip occur in spam and ham
hamSkip = 0


#NB: I decided to add a large negative number to the probability of
each class if the word did not appear
#in that class. If I skipped over the word, the accuracy was 0%. I
decided that although this resembles smoothing
#it is still appropriate. If I were to skip a word that means that
the conditional probability of a word appearing
#in the class it did not appear in is 0, which is not true. Instead
I set it to a small number.
#Other methods have been discussed in class but I believe this is t
he most appropriate.
for line in sys.stdin:
    line = line.strip()
    line = line.rstrip()
    components = line.split('\t') #split line
    text = " ".join(components[-2:]).strip() #combine subject and t
```

```
ext
    text = re.findall(WORD_RE, text)

    spamScore = log(priorSpam) #take logs
    hamScore = log(priorHam)
    for word in text:
        if word in words.keys():
            if float(words[word]['spam_like']) != 0: #this checks i
f a word has occurred in a class
                spamScore += log(float(words[word]['spam_like']))
#increment probability
            else:
                spamScore += -300
                spamSkip += 1 #skipped over a word in spam
            if float(words[word]['ham_like']) != 0: #repeat procedu
re for ham
                hamScore += log(float(words[word]['ham_like']))
            else:
                hamScore += -300
                hamSkip += 1
        pred = 0 #predicted class
    if spamScore > hamScore:
        pred = 1
    #output is email ID (key), true flag, predicted class, posterio
r probabilities (exponentiated) and skip counts
    #When I tried to print the skip counts by themselves there was
an error. I do not know the cause of this error
    #and I know it is inefficient and wrong, but this allows me to
at least process it
    print components[0] + '\t' + components[1] + '\t' + str(pred) +
'\t' + str(exp(spamScore)) + '\t' + str(exp(hamScore)) + '\t' + st
r(spamSkip) + '\t' + str(hamSkip)
```

```
Overwriting mapper.py
```

In [12]:
```
!chmod a+x mapper.py
```

In [9]:
```
%%writefile reducer.py
#!/usr/bin/python
import sys

misclassified = 0 #number of emails misclassified
skipSpam = 0 #number of times skip a word in spam
skipHam = 0

for line in sys.stdin:
    line = line.strip()
    line = line.rstrip()
    components = line.split('\t') #parse intput
    if int(components[1]) != int(components[2]): #if the true class
ification and the predicted do not agree, increment
            misclassified += 1
    skipSpam = int(components[5])
    skipHam = int(components[6])
    print line #print results
#print output
print "Misclassified: " + str(misclassified) + " which means this h
as an accuracy of " + str(100-misclassified) + "%"
print "Skipped " + str(skipSpam) + " words in spam"
print "Skipped " + str(skipHam) + " words in ham"
```

```
Overwriting reducer.py
```

In [14]:
```
!chmod a+x reducer.py
```

In [67]:
```
#check results
#!cat enronemail_1h.txt | python mapper.py | python reducer.py > ou
tput.txt
```

In [10]:
```
#Start hadoop yarn
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-yarn.sh
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-dfs.sh
```

```
starting yarn daemons
starting resourcemanager, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/yarn-dunmireg-resourcemanager-Glenns-Air.hom
e.out
localhost: starting nodemanager, logging to /usr/local/Cellar/hado
op/2.7.1/libexec/logs/yarn-dunmireg-nodemanager-Glenns-Air.home.ou
t
16/01/25 21:18:13 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/hadoop-dunmireg-namenode-Glenns-Air.home.out
localhost: starting datanode, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/hadoop-dunmireg-datanode-Glenns-Air.home.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/Cella
r/hadoop/2.7.1/libexec/logs/hadoop-dunmireg-secondarynamenode-Glen
ns-Air.home.out
16/01/25 21:18:29 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [ ]:
```
#make directory
#!hdfs dfs -mkdir -p /user/dunmireg
```

In [11]:
```
#add input to hdfs
!hdfs dfs -put enronemail_1h.txt /user/dunmireg
```

```
16/01/25 21:18:38 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [12]:
```
!hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/too
ls/lib/hadoop-streaming-2.7.1.jar \
-mapper mapper.py \
-reducer reducer.py \
-input enronemail_1h.txt \
-output enroneEmailClassificationNoSmoothing
```

```
16/01/25 21:18:42 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:18:43 INFO Configuration.deprecation: session.id is de
precated. Instead, use dfs.metrics.session-id
16/01/25 21:18:43 INFO jvm.JvmMetrics: Initializing JVM Metrics wi
th processName=JobTracker, sessionId=
16/01/25 21:18:43 INFO jvm.JvmMetrics: Cannot initialize JVM Metri
cs with processName=JobTracker, sessionId= - already initialized
16/01/25 21:18:43 INFO mapred.FileInputFormat: Total input paths t
o process : 1
16/01/25 21:18:43 INFO mapreduce.JobSubmitter: number of splits:1
16/01/25 21:18:44 INFO mapreduce.JobSubmitter: Submitting tokens f
or job: job_local762733417_0001
16/01/25 21:18:44 INFO mapred.LocalJobRunner: OutputCommitter set
in config null
16/01/25 21:18:44 INFO mapreduce.Job: The url to track the job: ht
tp://localhost:8080/
16/01/25 21:18:44 INFO mapred.LocalJobRunner: OutputCommitter is o
rg.apache.hadoop.mapred.FileOutputCommitter
16/01/25 21:18:44 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/25 21:18:44 INFO mapreduce.Job: Running job: job_local762733
417_0001
16/01/25 21:18:44 INFO mapred.LocalJobRunner: Waiting for map task
s
16/01/25 21:18:44 INFO mapred.LocalJobRunner: Starting task: attem
pt_local762733417_0001_m_000000_0
16/01/25 21:18:44 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/25 21:18:44 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/01/25 21:18:44 INFO mapred.Task:  Using ResourceCalculatorProce
ssTree : null
16/01/25 21:18:44 INFO mapred.MapTask: Processing split: hdfs://lo
calhost:9000/user/dunmireg/enronemail_1h.txt:0+204658
16/01/25 21:18:44 INFO mapred.MapTask: numReduceTasks: 1
16/01/25 21:18:44 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(10
4857584)
16/01/25 21:18:44 INFO mapred.MapTask: mapreduce.task.io.sort.mb:
100
16/01/25 21:18:44 INFO mapred.MapTask: soft limit at 83886080
16/01/25 21:18:44 INFO mapred.MapTask: bufstart = 0; bufvoid = 104
857600
16/01/25 21:18:44 INFO mapred.MapTask: kvstart = 26214396; length
= 6553600
16/01/25 21:18:44 INFO mapred.MapTask: Map output collector class
= org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/25 21:18:44 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW2/./mapper.py]
16/01/25 21:18:44 INFO Configuration.deprecation: mapred.tip.id is
deprecated. Instead, use mapreduce.task.id
16/01/25 21:18:44 INFO Configuration.deprecation: mapred.local.dir
```

```
is deprecated. Instead, use mapreduce.cluster.local.dir
16/01/25 21:18:44 INFO Configuration.deprecation: map.input.file i
s deprecated. Instead, use mapreduce.map.input.file
16/01/25 21:18:44 INFO Configuration.deprecation: mapred.skip.on i
s deprecated. Instead, use mapreduce.job.skiprecords
16/01/25 21:18:44 INFO Configuration.deprecation: map.input.length
is deprecated. Instead, use mapreduce.map.input.length
16/01/25 21:18:44 INFO Configuration.deprecation: mapred.work.outp
ut.dir is deprecated. Instead, use mapreduce.task.output.dir
16/01/25 21:18:44 INFO Configuration.deprecation: map.input.start
is deprecated. Instead, use mapreduce.map.input.start
16/01/25 21:18:44 INFO Configuration.deprecation: mapred.job.id is
deprecated. Instead, use mapreduce.job.id
16/01/25 21:18:44 INFO Configuration.deprecation: user.name is dep
recated. Instead, use mapreduce.job.user.name
16/01/25 21:18:44 INFO Configuration.deprecation: mapred.task.is.m
ap is deprecated. Instead, use mapreduce.task.ismap
16/01/25 21:18:44 INFO Configuration.deprecation: mapred.task.id i
s deprecated. Instead, use mapreduce.task.attempt.id
16/01/25 21:18:44 INFO Configuration.deprecation: mapred.task.part
ition is deprecated. Instead, use mapreduce.task.partition
16/01/25 21:18:44 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/25 21:18:44 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/25 21:18:45 INFO mapreduce.Job: Job job_local762733417_0001
running in uber mode : false
16/01/25 21:18:45 INFO mapreduce.Job:  map 0% reduce 0%
16/01/25 21:18:46 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:10
0=100/1 [rec/s] out:0=0/1 [rec/s]
16/01/25 21:18:49 INFO streaming.PipeMapRed: MRErrorThread done
16/01/25 21:18:49 INFO streaming.PipeMapRed: Records R/W=100/1
16/01/25 21:18:49 INFO streaming.PipeMapRed: mapRedFinished
16/01/25 21:18:49 INFO mapred.LocalJobRunner:
16/01/25 21:18:49 INFO mapred.MapTask: Starting flush of map outpu
t
16/01/25 21:18:49 INFO mapred.MapTask: Spilling map output
16/01/25 21:18:49 INFO mapred.MapTask: bufstart = 0; bufend = 485
5; bufvoid = 104857600
16/01/25 21:18:49 INFO mapred.MapTask: kvstart = 26214396(10485758
4); kvend = 26214000(104856000); length = 397/6553600
16/01/25 21:18:49 INFO mapred.MapTask: Finished spill 0
16/01/25 21:18:49 INFO mapred.Task: Task:attempt_local762733417_00
01_m_000000_0 is done. And is in the process of committing
16/01/25 21:18:49 INFO mapred.LocalJobRunner: Records R/W=100/1
16/01/25 21:18:49 INFO mapred.Task: Task 'attempt_local762733417_0
001_m_000000_0' done.
16/01/25 21:18:49 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local762733417_0001_m_000000_0
16/01/25 21:18:49 INFO mapred.LocalJobRunner: map task executor co
mplete.
16/01/25 21:18:49 INFO mapred.LocalJobRunner: Waiting for reduce t
asks
16/01/25 21:18:49 INFO mapred.LocalJobRunner: Starting task: attem
```

```
pt_local762733417_0001_r_000000_0
16/01/25 21:18:49 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/25 21:18:49 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/01/25 21:18:49 INFO mapred.Task:  Using ResourceCalculatorProce
ssTree : null
16/01/25 21:18:49 INFO mapred.ReduceTask: Using ShuffleConsumerPlu
gin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@60045d35
16/01/25 21:18:49 INFO reduce.MergeManagerImpl: MergerManager: mem
oryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshol
d=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/01/25 21:18:49 INFO reduce.EventFetcher: attempt_local76273341
7_0001_r_000000_0 Thread started: EventFetcher for fetching Map Co
mpletion Events
16/01/25 21:18:49 INFO reduce.LocalFetcher: localfetcher#1 about t
o shuffle output of map attempt_local762733417_0001_m_000000_0 dec
omp: 5057 len: 5061 to MEMORY
16/01/25 21:18:49 INFO reduce.InMemoryMapOutput: Read 5057 bytes f
rom map-output for attempt_local762733417_0001_m_000000_0
16/01/25 21:18:49 INFO reduce.MergeManagerImpl: closeInMemoryFile
-> map-output of size: 5057, inMemoryMapOutputs.size() -> 1, commi
tMemory -> 0, usedMemory ->5057
16/01/25 21:18:49 INFO reduce.EventFetcher: EventFetcher is interr
upted.. Returning
16/01/25 21:18:49 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 21:18:49 INFO reduce.MergeManagerImpl: finalMerge called
with 1 in-memory map-outputs and 0 on-disk map-outputs
16/01/25 21:18:49 INFO mapred.Merger: Merging 1 sorted segments
16/01/25 21:18:49 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 5032 bytes
16/01/25 21:18:49 INFO reduce.MergeManagerImpl: Merged 1 segments,
5057 bytes to disk to satisfy reduce memory limit
16/01/25 21:18:49 INFO reduce.MergeManagerImpl: Merging 1 files, 5
061 bytes from disk
16/01/25 21:18:49 INFO reduce.MergeManagerImpl: Merging 0 segment
s, 0 bytes from memory into reduce
16/01/25 21:18:49 INFO mapred.Merger: Merging 1 sorted segments
16/01/25 21:18:49 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 5032 bytes
16/01/25 21:18:49 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 21:18:49 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW2/./reducer.py]
16/01/25 21:18:49 INFO Configuration.deprecation: mapred.job.track
er is deprecated. Instead, use mapreduce.jobtracker.address
16/01/25 21:18:49 INFO Configuration.deprecation: mapred.map.tasks
is deprecated. Instead, use mapreduce.job.maps
16/01/25 21:18:49 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/25 21:18:49 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/25 21:18:49 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/25 21:18:49 INFO streaming.PipeMapRed: MRErrorThread done
```

```
16/01/25 21:18:49 INFO streaming.PipeMapRed: Records R/W=100/1
16/01/25 21:18:49 INFO streaming.PipeMapRed: mapRedFinished
16/01/25 21:18:49 INFO mapred.Task: Task:attempt_local762733417_00
01_r_000000_0 is done. And is in the process of committing
16/01/25 21:18:49 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 21:18:49 INFO mapred.Task: Task attempt_local762733417_00
01_r_000000_0 is allowed to commit now
16/01/25 21:18:49 INFO output.FileOutputCommitter: Saved output of
task 'attempt_local762733417_0001_r_000000_0' to hdfs://localhos
t:9000/user/dunmireg/enroneEmailClassificationNoSmoothing/_tempora
ry/0/task_local762733417_0001_r_000000
16/01/25 21:18:49 INFO mapred.LocalJobRunner: Records R/W=100/1 >
reduce
16/01/25 21:18:49 INFO mapred.Task: Task 'attempt_local762733417_0
001_r_000000_0' done.
16/01/25 21:18:49 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local762733417_0001_r_000000_0
16/01/25 21:18:49 INFO mapred.LocalJobRunner: reduce task executor
complete.
16/01/25 21:18:50 INFO mapreduce.Job:  map 100% reduce 100%
16/01/25 21:18:50 INFO mapreduce.Job: Job job_local762733417_0001
completed successfully
16/01/25 21:18:50 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=222238
                FILE: Number of bytes written=814067
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=409316
                HDFS: Number of bytes written=4969
                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=100
                Map output records=100
                Map output bytes=4855
                Map output materialized bytes=5061
                Input split bytes=105
                Combine input records=0
                Combine output records=0
                Reduce input groups=100
                Reduce shuffle bytes=5061
                Reduce input records=100
                Reduce output records=103
                Spilled Records=200
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=619708416
        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=204658
        File Output Format Counters
                Bytes Written=4969
16/01/25 21:18:50 INFO streaming.StreamJob: Output directory: enro
neEmailClassificationNoSmoothing
```

In [49]:
```
#!hdfs dfs -cat /user/dunmireg/enroneEmailClassificationNoSmoothin
g/part-00000
```

In [13]:
```
#move output to local directory - makes easier to process for analy
sis
!hadoop fs -copyToLocal /user/dunmireg/enroneEmailClassificationNoS
moothing
```

```
16/01/25 21:19:23 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:19:24 WARN hdfs.DFSClient: DFSInputStream has been clo
sed already
16/01/25 21:19:24 WARN hdfs.DFSClient: DFSInputStream has been clo
sed already
```

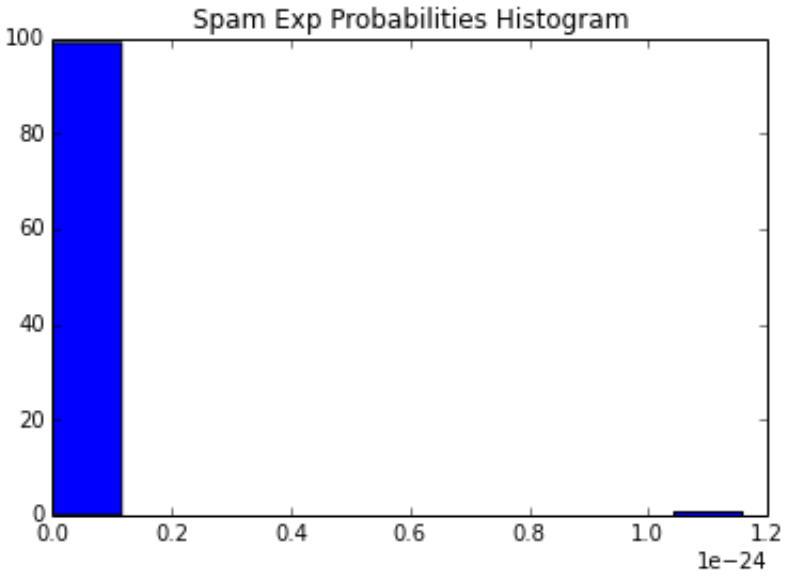
In [14]:
```
#Remove output directory and stop yarn and hdfs
!hadoop fs -rmr /user/dunmireg/enronemail_1h.txt #check
!hadoop fs -rmr /user/dunmireg/enroneEmailClassificationNoSmoothing
!/usr/local/Cellar/hadoop/2.7.1/sbin/stop-yarn.sh
!/usr/local/Cellar/hadoop/2.7.1/sbin/stop-dfs.sh
```

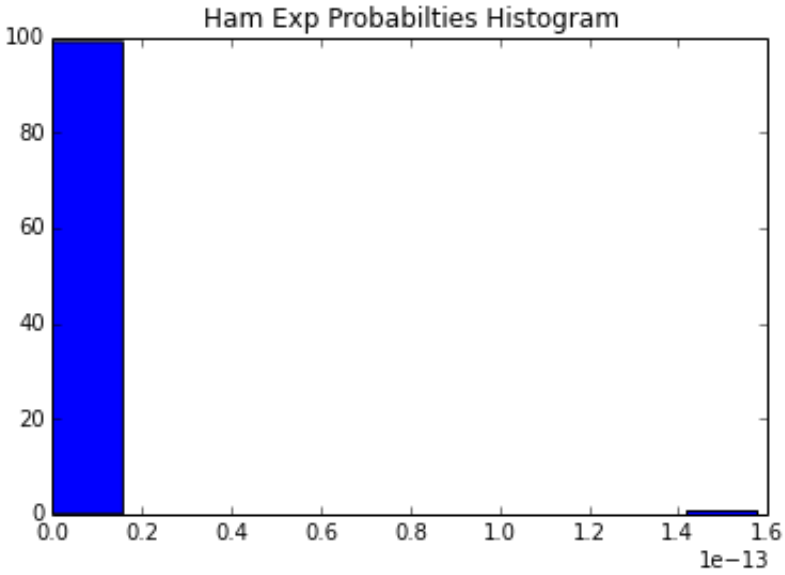
```
rmr: DEPRECATED: Please use 'rm -r' instead.
16/01/25 21:19:27 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:19:27 INFO fs.TrashPolicyDefault: Namenode trash confi
guration: Deletion interval = 0 minutes, Emptier interval = 0 minu
tes.
Deleted /user/dunmireg/enronemail_1h.txt
rmr: DEPRECATED: Please use 'rm -r' instead.
16/01/25 21:19:28 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:19:29 INFO fs.TrashPolicyDefault: Namenode trash confi
guration: Deletion interval = 0 minutes, Emptier interval = 0 minu
tes.
Deleted /user/dunmireg/enroneEmailClassificationNoSmoothing
stopping yarn daemons
stopping resourcemanager
localhost: stopping nodemanager
no proxyserver to stop
16/01/25 21:19:40 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here 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 21:19:58 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [15]:
```
#Display outputs from results file
import os

with open(os.path.join('./enroneEmailClassificationNoSmoothing', 'p
art-00000'), 'r') as myfile:
    lines = myfile.readlines() #read file
    lines = lines[-3:] #get last 3 lines
    for line in lines:
        print line #print results
```

```
Misclassified: 0 which means this has an accuracy of 100%

Skipped 4961 words in spam

Skipped 5694 words in ham
```

In [7]:
```
%matplotlib inline
#Make histogram
import os
import matplotlib.pyplot as plt

spam_probs = [] #list of spam probabilities
ham_probs = [] #list of ham probabilities
with open(os.path.join('./enroneEmailClassificationNoSmoothing', 'p
art-00000'), 'r') as myfile: #read file
    lines = myfile.readlines()
    for line in lines[:-3]: #exclude last 3 lines which have result
s
        components = line.split('\t')
        spam_probs.append(float(components[3]))
        ham_probs.append(float(components[4]))

s = plt.figure(1)
plt.hist(spam_probs)
plt.xlabel = "Probability"
plt.ylabel = "Frequency"
plt.title("Spam Exp Probabilities Histogram")

h = plt.figure(2)
plt.hist(ham_probs)
plt.xlabel = "Probability"
plt.ylabel = "Frequency"
plt.title("Ham Exp Probabilties Histogram")
plt.show()
```


Spam Exp Probabilities Histogram
100
80
60
40
20
0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e−24


Ham Exp Probabilties Histogram
100
80
60
40
20
0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1e−13

Summary

Overall, I am seeing an accuracy of 100%, meaning none of the emails are misclassified. This would probably suggest my model is overfitting the data. However, in this case we are using the training set as our testing set so there is a reason for that.

I found earlier I:

Skipped 4961 words in spam

Skipped 5694 words in ham

In observing the exponentiated probabilities we can see that the vast majority of words contribute a very tiny amount to both spam and ham classification. However, there looks to be a very small number of words that contribute large probabilities to either spam or ham. This would seem to suggest that a small subset of words really make the difference in classification. That is, the presence of these "classifier words" would probably be a key component in classifying an email as ham or spam.

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)

In [20]:
```
%%writefile mapper.py
#!/usr/bin/python

import sys
import re
WORD_RE = re.compile(r"[\w']+") #regex for string matching

for line in sys.stdin:
    components = line.split('\t') #split input file
    text = " ".join(components[-2:]).strip() #combine to produce su
bject and content together
    words = re.findall(WORD_RE, text)
    for word in words:
        print components[0] + '\t' + word + '\t' + components[1] #p
rint email ID, word, spam flag
```

```
Overwriting mapper.py
```

In [21]:
```
%%writefile reducer.py
#!/usr/bin/python
import sys

emails = set() #hold email IDs
words = {} #hold words and associated counts
spam_emails = 0 #how many emails are marked as spam
spam_word_count = 0 #count of words in spam
ham_word_count = 0 #count of words in ham
vocab = set() #set of unique words in all text

for line in sys.stdin:
    line = line.strip()
    line = line.rstrip()
    components = line.split('\t')

    ID = components[0] #parse components to appropriate variables
    word = components[1]
    spam = int(components[2])

    if word not in words.keys():
        words[word] = {'spam_count': 0, 'ham_count': 0} #add word t
o dictionary if not there already
        vocab.add(word) #add word to vocab
    if ID not in emails:
        emails.add(ID)
        if spam == 1:
            spam_emails += 1 #increment spam emails counter


    if spam == 1:
        words[word]['spam_count'] += 1 #if email is spam, increment
spam counter by 1
        spam_word_count += 1
    else:
        words[word]['ham_count'] += 1 #repeat for ham
        ham_word_count += 1


prior_spam = float(spam_emails)/len(emails) #get prior probabilitie
s
prior_ham = 1-prior_spam

for i, word in words.iteritems():
    #This calculation uses a laplace smoother, +1 to numerator and
+ vocab in denominator
    #See wikipedia entry
    word['spam_like'] = float(word['spam_count'] + 1)/(spam_word_co
unt + len(vocab)) #calculate conditional probs
    word['ham_like'] = float(word['ham_count'] + 1)/(ham_word_count
+ len(vocab))
```

```
print prior_spam
print prior_ham
for word in words.keys():
    #Word "\t" spam likelihood '\t' ham likelihood written to file
    print word + '\t' + str(words[word]['spam_like']) + '\t' + st
r(words[word]['ham_like']) #print results
```

```
Overwriting reducer.py
```

In [22]:
```
#Start hadoop yarn
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-yarn.sh
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-dfs.sh
```

```
starting yarn daemons
starting resourcemanager, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/yarn-dunmireg-resourcemanager-Glenns-Air.hom
e.out
localhost: starting nodemanager, logging to /usr/local/Cellar/hado
op/2.7.1/libexec/logs/yarn-dunmireg-nodemanager-Glenns-Air.home.ou
t
16/01/25 21:22:26 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/hadoop-dunmireg-namenode-Glenns-Air.home.out
localhost: starting datanode, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/hadoop-dunmireg-datanode-Glenns-Air.home.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/Cella
r/hadoop/2.7.1/libexec/logs/hadoop-dunmireg-secondarynamenode-Glen
ns-Air.home.out
16/01/25 21:22:41 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [4]:
```
#make directory
#!hdfs dfs -mkdir -p /user/dunmireg
```

```
16/01/24 22:46:13 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [23]:
```
#add input to hdfs
!hdfs dfs -put enronemail_1h.txt /user/dunmireg
```

```
16/01/25 21:22:47 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [24]:
```
!hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/too
ls/lib/hadoop-streaming-2.7.1.jar \
-mapper mapper.py \
-reducer reducer.py \
-input enronemail_1h.txt \
-output enroneEmailCondProbLaplace
```

```
16/01/25 21:22:51 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:22:52 INFO Configuration.deprecation: session.id is de
precated. Instead, use dfs.metrics.session-id
16/01/25 21:22:52 INFO jvm.JvmMetrics: Initializing JVM Metrics wi
th processName=JobTracker, sessionId=
16/01/25 21:22:52 INFO jvm.JvmMetrics: Cannot initialize JVM Metri
cs with processName=JobTracker, sessionId= - already initialized
16/01/25 21:22:52 INFO mapred.FileInputFormat: Total input paths t
o process : 1
16/01/25 21:22:52 INFO mapreduce.JobSubmitter: number of splits:1
16/01/25 21:22:52 INFO mapreduce.JobSubmitter: Submitting tokens f
or job: job_local1220185851_0001
16/01/25 21:22:53 INFO mapred.LocalJobRunner: OutputCommitter set
in config null
16/01/25 21:22:53 INFO mapred.LocalJobRunner: OutputCommitter is o
rg.apache.hadoop.mapred.FileOutputCommitter
16/01/25 21:22:53 INFO mapreduce.Job: The url to track the job: ht
tp://localhost:8080/
16/01/25 21:22:53 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/25 21:22:53 INFO mapreduce.Job: Running job: job_local122018
5851_0001
16/01/25 21:22:53 INFO mapred.LocalJobRunner: Waiting for map task
s
16/01/25 21:22:53 INFO mapred.LocalJobRunner: Starting task: attem
pt_local1220185851_0001_m_000000_0
16/01/25 21:22:53 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/25 21:22:53 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/01/25 21:22:53 INFO mapred.Task:  Using ResourceCalculatorProce
ssTree : null
16/01/25 21:22:53 INFO mapred.MapTask: Processing split: hdfs://lo
calhost:9000/user/dunmireg/enronemail_1h.txt:0+204658
16/01/25 21:22:53 INFO mapred.MapTask: numReduceTasks: 1
16/01/25 21:22:53 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(10
4857584)
16/01/25 21:22:53 INFO mapred.MapTask: mapreduce.task.io.sort.mb:
100
16/01/25 21:22:53 INFO mapred.MapTask: soft limit at 83886080
16/01/25 21:22:53 INFO mapred.MapTask: bufstart = 0; bufvoid = 104
857600
16/01/25 21:22:53 INFO mapred.MapTask: kvstart = 26214396; length
= 6553600
16/01/25 21:22:53 INFO mapred.MapTask: Map output collector class
= org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/25 21:22:53 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW2/./mapper.py]
16/01/25 21:22:53 INFO Configuration.deprecation: mapred.tip.id is
deprecated. Instead, use mapreduce.task.id
16/01/25 21:22:53 INFO Configuration.deprecation: mapred.local.dir
```

```
is deprecated. Instead, use mapreduce.cluster.local.dir
16/01/25 21:22:53 INFO Configuration.deprecation: map.input.file i
s deprecated. Instead, use mapreduce.map.input.file
16/01/25 21:22:53 INFO Configuration.deprecation: mapred.skip.on i
s deprecated. Instead, use mapreduce.job.skiprecords
16/01/25 21:22:53 INFO Configuration.deprecation: map.input.length
is deprecated. Instead, use mapreduce.map.input.length
16/01/25 21:22:53 INFO Configuration.deprecation: mapred.work.outp
ut.dir is deprecated. Instead, use mapreduce.task.output.dir
16/01/25 21:22:53 INFO Configuration.deprecation: map.input.start
is deprecated. Instead, use mapreduce.map.input.start
16/01/25 21:22:53 INFO Configuration.deprecation: mapred.job.id is
deprecated. Instead, use mapreduce.job.id
16/01/25 21:22:53 INFO Configuration.deprecation: user.name is dep
recated. Instead, use mapreduce.job.user.name
16/01/25 21:22:53 INFO Configuration.deprecation: mapred.task.is.m
ap is deprecated. Instead, use mapreduce.task.ismap
16/01/25 21:22:53 INFO Configuration.deprecation: mapred.task.id i
s deprecated. Instead, use mapreduce.task.attempt.id
16/01/25 21:22:53 INFO Configuration.deprecation: mapred.task.part
ition is deprecated. Instead, use mapreduce.task.partition
16/01/25 21:22:53 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/25 21:22:53 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/25 21:22:53 INFO streaming.PipeMapRed: Records R/W=72/1
16/01/25 21:22:53 INFO streaming.PipeMapRed: R/W/S=100/8828/0 in:N
A [rec/s] out:NA [rec/s]
16/01/25 21:22:53 INFO streaming.PipeMapRed: MRErrorThread done
16/01/25 21:22:53 INFO streaming.PipeMapRed: mapRedFinished
16/01/25 21:22:53 INFO mapred.LocalJobRunner:
16/01/25 21:22:53 INFO mapred.MapTask: Starting flush of map outpu
t
16/01/25 21:22:53 INFO mapred.MapTask: Spilling map output
16/01/25 21:22:53 INFO mapred.MapTask: bufstart = 0; bufend = 1032
108; bufvoid = 104857600
16/01/25 21:22:53 INFO mapred.MapTask: kvstart = 26214396(10485758
4); kvend = 26082748(104330992); length = 131649/6553600
16/01/25 21:22:53 INFO mapred.MapTask: Finished spill 0
16/01/25 21:22:53 INFO mapred.Task: Task:attempt_local1220185851_0
001_m_000000_0 is done. And is in the process of committing
16/01/25 21:22:53 INFO mapred.LocalJobRunner: Records R/W=72/1
16/01/25 21:22:53 INFO mapred.Task: Task 'attempt_local122018585
1_0001_m_000000_0' done.
16/01/25 21:22:53 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local1220185851_0001_m_000000_0
16/01/25 21:22:53 INFO mapred.LocalJobRunner: map task executor co
mplete.
16/01/25 21:22:53 INFO mapred.LocalJobRunner: Waiting for reduce t
asks
16/01/25 21:22:53 INFO mapred.LocalJobRunner: Starting task: attem
pt_local1220185851_0001_r_000000_0
16/01/25 21:22:53 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
```

```
16/01/25 21:22:53 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/01/25 21:22:53 INFO mapred.Task:  Using ResourceCalculatorProce
ssTree : null
16/01/25 21:22:53 INFO mapred.ReduceTask: Using ShuffleConsumerPlu
gin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@52df4dfc
16/01/25 21:22:53 INFO reduce.MergeManagerImpl: MergerManager: mem
oryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshol
d=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/01/25 21:22:53 INFO reduce.EventFetcher: attempt_local122018585
1_0001_r_000000_0 Thread started: EventFetcher for fetching Map Co
mpletion Events
16/01/25 21:22:53 INFO reduce.LocalFetcher: localfetcher#1 about t
o shuffle output of map attempt_local1220185851_0001_m_000000_0 de
comp: 1097936 len: 1097940 to MEMORY
16/01/25 21:22:53 INFO reduce.InMemoryMapOutput: Read 1097936 byte
s from map-output for attempt_local1220185851_0001_m_000000_0
16/01/25 21:22:53 INFO reduce.MergeManagerImpl: closeInMemoryFile
-> map-output of size: 1097936, inMemoryMapOutputs.size() -> 1, co
mmitMemory -> 0, usedMemory ->1097936
16/01/25 21:22:53 INFO reduce.EventFetcher: EventFetcher is interr
upted.. Returning
16/01/25 21:22:53 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 21:22:53 INFO reduce.MergeManagerImpl: finalMerge called
with 1 in-memory map-outputs and 0 on-disk map-outputs
16/01/25 21:22:53 INFO mapred.Merger: Merging 1 sorted segments
16/01/25 21:22:53 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 1097911 bytes
16/01/25 21:22:53 INFO reduce.MergeManagerImpl: Merged 1 segments,
1097936 bytes to disk to satisfy reduce memory limit
16/01/25 21:22:53 INFO reduce.MergeManagerImpl: Merging 1 files, 1
097940 bytes from disk
16/01/25 21:22:53 INFO reduce.MergeManagerImpl: Merging 0 segment
s, 0 bytes from memory into reduce
16/01/25 21:22:53 INFO mapred.Merger: Merging 1 sorted segments
16/01/25 21:22:53 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 1097911 bytes
16/01/25 21:22:53 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 21:22:53 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW2/./reducer.py]
16/01/25 21:22:53 INFO Configuration.deprecation: mapred.job.track
er is deprecated. Instead, use mapreduce.jobtracker.address
16/01/25 21:22:53 INFO Configuration.deprecation: mapred.map.tasks
is deprecated. Instead, use mapreduce.job.maps
16/01/25 21:22:53 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/25 21:22:53 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/25 21:22:53 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/25 21:22:54 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/25 21:22:54 INFO mapreduce.Job: Job job_local1220185851_0001
running in uber mode : false
```

```
16/01/25 21:22:54 INFO mapreduce.Job:  map 100% reduce 0%
16/01/25 21:22:54 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/25 21:22:56 INFO streaming.PipeMapRed: Records R/W=32913/1
16/01/25 21:22:56 INFO streaming.PipeMapRed: MRErrorThread done
16/01/25 21:22:56 INFO streaming.PipeMapRed: mapRedFinished
16/01/25 21:22:56 INFO mapred.Task: Task:attempt_local1220185851_0
001_r_000000_0 is done. And is in the process of committing
16/01/25 21:22:56 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 21:22:56 INFO mapred.Task: Task attempt_local1220185851_0
001_r_000000_0 is allowed to commit now
16/01/25 21:22:56 INFO output.FileOutputCommitter: Saved output of
task 'attempt_local1220185851_0001_r_000000_0' to hdfs://localhos
t:9000/user/dunmireg/enroneEmailCondProbLaplace/_temporary/0/tas
k_local1220185851_0001_r_000000
16/01/25 21:22:56 INFO mapred.LocalJobRunner: Records R/W=32913/1
> reduce
16/01/25 21:22:56 INFO mapred.Task: Task 'attempt_local122018585
1_0001_r_000000_0' done.
16/01/25 21:22:56 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local1220185851_0001_r_000000_0
16/01/25 21:22:56 INFO mapred.LocalJobRunner: reduce task executor
complete.
16/01/25 21:22:57 INFO mapreduce.Job:  map 100% reduce 100%
16/01/25 21:22:57 INFO mapreduce.Job: Job job_local1220185851_0001
completed successfully
16/01/25 21:22:57 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=2407996
                FILE: Number of bytes written=4095672
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=409316
                HDFS: Number of bytes written=237609
                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=100
                Map output records=32913
                Map output bytes=1032108
                Map output materialized bytes=1097940
                Input split bytes=105
                Combine input records=0
                Combine output records=0
                Reduce input groups=100
                Reduce shuffle bytes=1097940
                Reduce input records=32913
                Reduce output records=5493
                Spilled Records=65826
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
```

```
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=618659840
        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=204658
        File Output Format Counters
                Bytes Written=237609
16/01/25 21:22:57 INFO streaming.StreamJob: Output directory: enro
neEmailCondProbLaplace
```

In [33]:
```
#Check output
#!hdfs dfs -cat /user/dunmireg/enroneEmailCondProbLaplace/part-0000
0
```

In [25]:
```
!hadoop fs -copyToLocal /user/dunmireg/enroneEmailCondProbLaplace
```

```
16/01/25 21:23:03 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:23:04 WARN hdfs.DFSClient: DFSInputStream has been clo
sed already
16/01/25 21:23:04 WARN hdfs.DFSClient: DFSInputStream has been clo
sed already
```

In [26]:
```
#Remove output directory and stop yarn and hdfs
!hadoop fs -rmr /user/dunmireg/enronemail_1h.txt
!hadoop fs -rmr /user/dunmireg/enroneEmailCondProbLaplace
!/usr/local/Cellar/hadoop/2.7.1/sbin/stop-yarn.sh
!/usr/local/Cellar/hadoop/2.7.1/sbin/stop-dfs.sh
```

```
rmr: DEPRECATED: Please use 'rm -r' instead.
16/01/25 21:23:06 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:23:06 INFO fs.TrashPolicyDefault: Namenode trash confi
guration: Deletion interval = 0 minutes, Emptier interval = 0 minu
tes.
Deleted /user/dunmireg/enronemail_1h.txt
rmr: DEPRECATED: Please use 'rm -r' instead.
16/01/25 21:23:07 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:23:08 INFO fs.TrashPolicyDefault: Namenode trash confi
guration: Deletion interval = 0 minutes, Emptier interval = 0 minu
tes.
Deleted /user/dunmireg/enroneEmailCondProbLaplace
stopping yarn daemons
stopping resourcemanager
localhost: stopping nodemanager
no proxyserver to stop
16/01/25 21:23:19 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here 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 21:23:38 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [27]:
```
%%writefile mapper.py
#!/usr/bin/python

#mapper for classification. This is essentially the same procedure
as in 2.3.
import sys
import re
import os
from math import log
from math import exp

priorSpam = 0 #hold priors
priorHam = 0
words = {}

with open(os.path.join('./enroneEmailCondProbLaplace', 'part-0000
0'), 'r') as myfile: #read file
    lines = myfile.readlines()
    priorSpam = float(lines[0]) #parse first lines for priors
    priorHam = float(lines[1])
    for line in lines[2:]:
        line = line.strip()
        line = line.rstrip()
        components = line.split('\t')
        #add conditional probabilities to words dictionary
        words[components[0]] = {'spam_like': float(components[1]),
'ham_like': float(components[2])}


WORD_RE = re.compile(r"[\w']+")

for line in sys.stdin:
    line = line.strip()
    line = line.rstrip()
    components = line.split('\t')
    text = " ".join(components[-2:]).strip()
    text = re.findall(WORD_RE, text)

    spamScore = log(priorSpam)
    hamScore = log(priorHam)
    for word in text:
        if word in words.keys():
            #add conditional probabilities to scores
            spamScore += log(float(words[word]['spam_like']))
            hamScore += log(float(words[word]['ham_like']))
        pred = 0 #assign prediction
    if spamScore > hamScore:
        pred = 1
    #output ID, true classification, prediction, and conditional pr
obabilities
    print components[0] + '\t' + components[1] + '\t' + str(pred) +
'\t' + str(exp(spamScore)) + '\t' + str(exp(hamScore))
```

```
Overwriting mapper.py
```

In [28]:
```
%%writefile reducer.py
#!/usr/bin/python
import sys

misclassified = 0 #count of how many emails are misclassified

for line in sys.stdin:
    line = line.strip()
    line = line.rstrip()
    components = line.split('\t') #split components
    if int(components[1]) != int(components[2]): #if true classific
ation and prediction don't agree, increment
            misclassified += 1
    print line
#print final output
print "Misclassified: " + str(misclassified) + " which means this h
as an accuracy of " + str(100-misclassified) + "%"
```

```
Overwriting reducer.py
```

In [36]:
```
#Check code
#!cat enronemail_1h.txt | python mapper.py | python reducer.py
```

In [29]:
```
#Start hadoop yarn
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-yarn.sh
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-dfs.sh
```

```
starting yarn daemons
starting resourcemanager, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/yarn-dunmireg-resourcemanager-Glenns-Air.hom
e.out
localhost: starting nodemanager, logging to /usr/local/Cellar/hado
op/2.7.1/libexec/logs/yarn-dunmireg-nodemanager-Glenns-Air.home.ou
t
16/01/25 21:24:55 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/hadoop-dunmireg-namenode-Glenns-Air.home.out
localhost: starting datanode, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/hadoop-dunmireg-datanode-Glenns-Air.home.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/Cella
r/hadoop/2.7.1/libexec/logs/hadoop-dunmireg-secondarynamenode-Glen
ns-Air.home.out
16/01/25 21:25:10 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [14]:
```
#make directory
#!hdfs dfs -mkdir -p /user/dunmireg
```

```
16/01/24 22:47:57 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [30]:
```
#add input to hdfs
!hdfs dfs -put enronemail_1h.txt /user/dunmireg
```

```
16/01/25 21:25:15 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [31]:
```
!hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/too
ls/lib/hadoop-streaming-2.7.1.jar \
-mapper mapper.py \
-reducer reducer.py \
-input enronemail_1h.txt \
-output enroneEmailClassLaplace
```

```
16/01/25 21:25:18 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:25:19 INFO Configuration.deprecation: session.id is de
precated. Instead, use dfs.metrics.session-id
16/01/25 21:25:19 INFO jvm.JvmMetrics: Initializing JVM Metrics wi
th processName=JobTracker, sessionId=
16/01/25 21:25:19 INFO jvm.JvmMetrics: Cannot initialize JVM Metri
cs with processName=JobTracker, sessionId= - already initialized
16/01/25 21:25:19 INFO mapred.FileInputFormat: Total input paths t
o process : 1
16/01/25 21:25:19 INFO mapreduce.JobSubmitter: number of splits:1
16/01/25 21:25:19 INFO mapreduce.JobSubmitter: Submitting tokens f
or job: job_local1852274356_0001
16/01/25 21:25:19 INFO mapreduce.Job: The url to track the job: ht
tp://localhost:8080/
16/01/25 21:25:19 INFO mapreduce.Job: Running job: job_local185227
4356_0001
16/01/25 21:25:19 INFO mapred.LocalJobRunner: OutputCommitter set
in config null
16/01/25 21:25:19 INFO mapred.LocalJobRunner: OutputCommitter is o
rg.apache.hadoop.mapred.FileOutputCommitter
16/01/25 21:25:19 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/25 21:25:20 INFO mapred.LocalJobRunner: Waiting for map task
s
16/01/25 21:25:20 INFO mapred.LocalJobRunner: Starting task: attem
pt_local1852274356_0001_m_000000_0
16/01/25 21:25:20 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/25 21:25:20 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/01/25 21:25:20 INFO mapred.Task:  Using ResourceCalculatorProce
ssTree : null
16/01/25 21:25:20 INFO mapred.MapTask: Processing split: hdfs://lo
calhost:9000/user/dunmireg/enronemail_1h.txt:0+204658
16/01/25 21:25:20 INFO mapred.MapTask: numReduceTasks: 1
16/01/25 21:25:20 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(10
4857584)
16/01/25 21:25:20 INFO mapred.MapTask: mapreduce.task.io.sort.mb:
100
16/01/25 21:25:20 INFO mapred.MapTask: soft limit at 83886080
16/01/25 21:25:20 INFO mapred.MapTask: bufstart = 0; bufvoid = 104
857600
16/01/25 21:25:20 INFO mapred.MapTask: kvstart = 26214396; length
= 6553600
16/01/25 21:25:20 INFO mapred.MapTask: Map output collector class
= org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/25 21:25:20 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW2/./mapper.py]
16/01/25 21:25:20 INFO Configuration.deprecation: mapred.tip.id is
deprecated. Instead, use mapreduce.task.id
16/01/25 21:25:20 INFO Configuration.deprecation: mapred.local.dir
```

```
is deprecated. Instead, use mapreduce.cluster.local.dir
16/01/25 21:25:20 INFO Configuration.deprecation: map.input.file i
s deprecated. Instead, use mapreduce.map.input.file
16/01/25 21:25:20 INFO Configuration.deprecation: mapred.skip.on i
s deprecated. Instead, use mapreduce.job.skiprecords
16/01/25 21:25:20 INFO Configuration.deprecation: map.input.length
is deprecated. Instead, use mapreduce.map.input.length
16/01/25 21:25:20 INFO Configuration.deprecation: mapred.work.outp
ut.dir is deprecated. Instead, use mapreduce.task.output.dir
16/01/25 21:25:20 INFO Configuration.deprecation: map.input.start
is deprecated. Instead, use mapreduce.map.input.start
16/01/25 21:25:20 INFO Configuration.deprecation: mapred.job.id is
deprecated. Instead, use mapreduce.job.id
16/01/25 21:25:20 INFO Configuration.deprecation: user.name is dep
recated. Instead, use mapreduce.job.user.name
16/01/25 21:25:20 INFO Configuration.deprecation: mapred.task.is.m
ap is deprecated. Instead, use mapreduce.task.ismap
16/01/25 21:25:20 INFO Configuration.deprecation: mapred.task.id i
s deprecated. Instead, use mapreduce.task.attempt.id
16/01/25 21:25:20 INFO Configuration.deprecation: mapred.task.part
ition is deprecated. Instead, use mapreduce.task.partition
16/01/25 21:25:20 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/25 21:25:20 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/25 21:25:20 INFO mapreduce.Job: Job job_local1852274356_0001
running in uber mode : false
16/01/25 21:25:20 INFO mapreduce.Job:  map 0% reduce 0%
16/01/25 21:25:21 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:10
0=100/1 [rec/s] out:0=0/1 [rec/s]
16/01/25 21:25:24 INFO streaming.PipeMapRed: MRErrorThread done
16/01/25 21:25:24 INFO streaming.PipeMapRed: Records R/W=100/1
16/01/25 21:25:24 INFO streaming.PipeMapRed: mapRedFinished
16/01/25 21:25:24 INFO mapred.LocalJobRunner:
16/01/25 21:25:24 INFO mapred.MapTask: Starting flush of map outpu
t
16/01/25 21:25:24 INFO mapred.MapTask: Spilling map output
16/01/25 21:25:24 INFO mapred.MapTask: bufstart = 0; bufend = 428
4; bufvoid = 104857600
16/01/25 21:25:24 INFO mapred.MapTask: kvstart = 26214396(10485758
4); kvend = 26214000(104856000); length = 397/6553600
16/01/25 21:25:24 INFO mapred.MapTask: Finished spill 0
16/01/25 21:25:24 INFO mapred.Task: Task:attempt_local1852274356_0
001_m_000000_0 is done. And is in the process of committing
16/01/25 21:25:24 INFO mapred.LocalJobRunner: Records R/W=100/1
16/01/25 21:25:24 INFO mapred.Task: Task 'attempt_local185227435
6_0001_m_000000_0' done.
16/01/25 21:25:24 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local1852274356_0001_m_000000_0
16/01/25 21:25:24 INFO mapred.LocalJobRunner: map task executor co
mplete.
16/01/25 21:25:24 INFO mapred.LocalJobRunner: Waiting for reduce t
asks
16/01/25 21:25:24 INFO mapred.LocalJobRunner: Starting task: attem
```

```
pt_local1852274356_0001_r_000000_0
16/01/25 21:25:24 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/25 21:25:24 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/01/25 21:25:24 INFO mapred.Task:  Using ResourceCalculatorProce
ssTree : null
16/01/25 21:25:24 INFO mapred.ReduceTask: Using ShuffleConsumerPlu
gin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@1e469b74
16/01/25 21:25:24 INFO reduce.MergeManagerImpl: MergerManager: mem
oryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshol
d=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/01/25 21:25:24 INFO reduce.EventFetcher: attempt_local185227435
6_0001_r_000000_0 Thread started: EventFetcher for fetching Map Co
mpletion Events
16/01/25 21:25:24 INFO reduce.LocalFetcher: localfetcher#1 about t
o shuffle output of map attempt_local1852274356_0001_m_000000_0 de
comp: 4486 len: 4490 to MEMORY
16/01/25 21:25:24 INFO reduce.InMemoryMapOutput: Read 4486 bytes f
rom map-output for attempt_local1852274356_0001_m_000000_0
16/01/25 21:25:24 INFO reduce.MergeManagerImpl: closeInMemoryFile
-> map-output of size: 4486, inMemoryMapOutputs.size() -> 1, commi
tMemory -> 0, usedMemory ->4486
16/01/25 21:25:24 INFO reduce.EventFetcher: EventFetcher is interr
upted.. Returning
16/01/25 21:25:24 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 21:25:24 INFO reduce.MergeManagerImpl: finalMerge called
with 1 in-memory map-outputs and 0 on-disk map-outputs
16/01/25 21:25:24 INFO mapred.Merger: Merging 1 sorted segments
16/01/25 21:25:24 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 4461 bytes
16/01/25 21:25:24 INFO reduce.MergeManagerImpl: Merged 1 segments,
4486 bytes to disk to satisfy reduce memory limit
16/01/25 21:25:24 INFO reduce.MergeManagerImpl: Merging 1 files, 4
490 bytes from disk
16/01/25 21:25:24 INFO reduce.MergeManagerImpl: Merging 0 segment
s, 0 bytes from memory into reduce
16/01/25 21:25:24 INFO mapred.Merger: Merging 1 sorted segments
16/01/25 21:25:24 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 4461 bytes
16/01/25 21:25:24 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 21:25:24 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW2/./reducer.py]
16/01/25 21:25:24 INFO Configuration.deprecation: mapred.job.track
er is deprecated. Instead, use mapreduce.jobtracker.address
16/01/25 21:25:24 INFO Configuration.deprecation: mapred.map.tasks
is deprecated. Instead, use mapreduce.job.maps
16/01/25 21:25:24 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/25 21:25:24 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/25 21:25:24 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/25 21:25:24 INFO streaming.PipeMapRed: MRErrorThread done
```

```
16/01/25 21:25:24 INFO streaming.PipeMapRed: Records R/W=100/1
16/01/25 21:25:24 INFO streaming.PipeMapRed: mapRedFinished
16/01/25 21:25:24 INFO mapred.Task: Task:attempt_local1852274356_0
001_r_000000_0 is done. And is in the process of committing
16/01/25 21:25:24 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 21:25:24 INFO mapred.Task: Task attempt_local1852274356_0
001_r_000000_0 is allowed to commit now
16/01/25 21:25:24 INFO output.FileOutputCommitter: Saved output of
task 'attempt_local1852274356_0001_r_000000_0' to hdfs://localhos
t:9000/user/dunmireg/enroneEmailClassLaplace/_temporary/0/task_loc
al1852274356_0001_r_000000
16/01/25 21:25:24 INFO mapred.LocalJobRunner: Records R/W=100/1 >
reduce
16/01/25 21:25:24 INFO mapred.Task: Task 'attempt_local185227435
6_0001_r_000000_0' done.
16/01/25 21:25:24 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local1852274356_0001_r_000000_0
16/01/25 21:25:24 INFO mapred.LocalJobRunner: reduce task executor
complete.
16/01/25 21:25:24 INFO mapreduce.Job:  map 100% reduce 100%
16/01/25 21:25:24 INFO mapreduce.Job: Job job_local1852274356_0001
completed successfully
16/01/25 21:25:24 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=221096
                FILE: Number of bytes written=815310
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=409316
                HDFS: Number of bytes written=4343
                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=100
                Map output records=100
                Map output bytes=4284
                Map output materialized bytes=4490
                Input split bytes=105
                Combine input records=0
                Combine output records=0
                Reduce input groups=100
                Reduce shuffle bytes=4490
                Reduce input records=100
                Reduce output records=101
                Spilled Records=200
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=622854144
        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=204658
        File Output Format Counters
                Bytes Written=4343
16/01/25 21:25:24 INFO streaming.StreamJob: Output directory: enro
neEmailClassLaplace
```

In [35]:
```
#Check output
#!hdfs dfs -cat /user/dunmireg/enroneEmailClassLaplace/part-00000
```

In [32]:
```
!hadoop fs -copyToLocal /user/dunmireg/enroneEmailClassLaplace
```

```
16/01/25 21:25:30 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:25:30 WARN hdfs.DFSClient: DFSInputStream has been clo
sed already
16/01/25 21:25:30 WARN hdfs.DFSClient: DFSInputStream has been clo
sed already
```

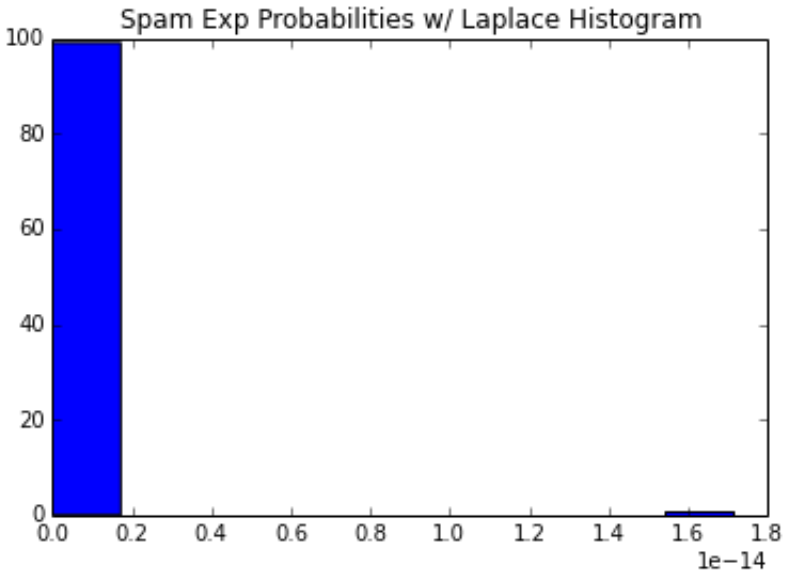
In [33]:
```
#Remove output directory and stop yarn and hdfs
!hadoop fs -rmr /user/dunmireg/enronemail_1h.txt
!hadoop fs -rmr /user/dunmireg/enroneEmailClassLaplace
!/usr/local/Cellar/hadoop/2.7.1/sbin/stop-yarn.sh
!/usr/local/Cellar/hadoop/2.7.1/sbin/stop-dfs.sh
```

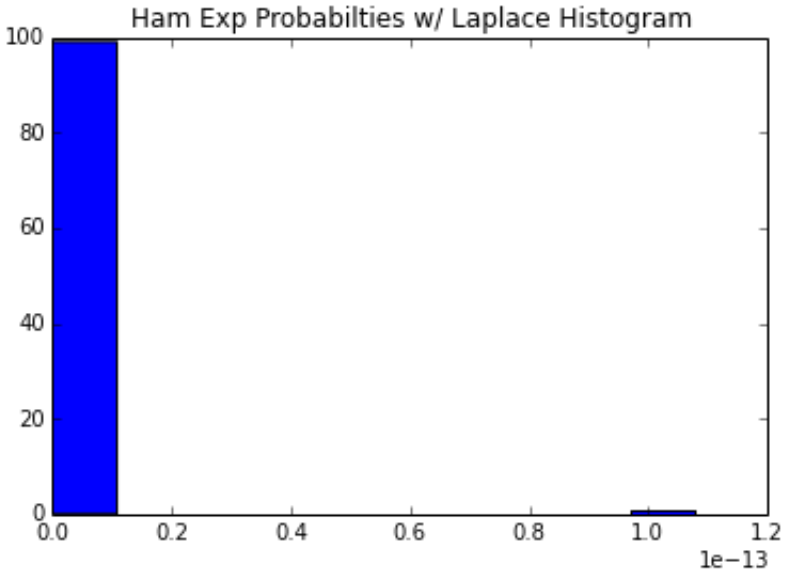
```
rmr: DEPRECATED: Please use 'rm -r' instead.
16/01/25 21:25:32 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:25:33 INFO fs.TrashPolicyDefault: Namenode trash confi
guration: Deletion interval = 0 minutes, Emptier interval = 0 minu
tes.
Deleted /user/dunmireg/enronemail_1h.txt
rmr: DEPRECATED: Please use 'rm -r' instead.
16/01/25 21:25:34 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:25:35 INFO fs.TrashPolicyDefault: Namenode trash confi
guration: Deletion interval = 0 minutes, Emptier interval = 0 minu
tes.
Deleted /user/dunmireg/enroneEmailClassLaplace
stopping yarn daemons
stopping resourcemanager
localhost: stopping nodemanager
no proxyserver to stop
16/01/25 21:25:46 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here 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 21:26:05 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [34]:
```
#Display outputs from results file
import os

with open(os.path.join('./enroneEmailClassLaplace', 'part-00000'),
'r') as myfile:
    lines = myfile.readlines() #read file
    lines = lines[-1:] #get last 3 lines
    for line in lines:
        print line #print results
```

```
Misclassified: 0 which means this has an accuracy of 100%
```

In [8]:
```
%matplotlib inline
#Make histogram
import os
import matplotlib.pyplot as plt

spam_probs = [] #list of spam probabilities
ham_probs = [] #list of ham probabilities
with open(os.path.join('./enroneEmailClassLaplace', 'part-00000'),
'r') as myfile: #read file
    lines = myfile.readlines()
    for line in lines[:-1]: #exclude last line which has results
        components = line.split('\t')
        spam_probs.append(float(components[3]))
        ham_probs.append(float(components[4]))

s = plt.figure(1)
plt.hist(spam_probs)
plt.xlabel = "Probability"
plt.ylabel = "Frequency"
plt.title("Spam Exp Probabilities w/ Laplace Histogram")

h = plt.figure(2)
plt.hist(ham_probs)
plt.xlabel = "Probability"
plt.ylabel = "Frequency"
plt.title("Ham Exp Probabilties w/ Laplace Histogram")
plt.show()
```





Summary

The misclassification rate of the no smoothing classifier was 0, which is the same as the misclassification rate of this laplace classifier. This probably has to do with the fact that in my earlier classifier I was doing a form of smoothing by adding a very small number to the probability for a word that did not appear in a class.

In this case I am seeing broadly the same trends. That is, the majority of words have a very small probability of either class but a handful of words seem to have an outsized impact on classification. It is interesting to note that in the previous (no smoothing) classifier the 'classifier words' in ham actually had a larger impact on score than the 'classifier' words in spam. In this case, the situation is reversed, with the 'classifier' words in spam having a larger impact than the words in ham

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:

In [36]:
```
%%writefile mapper.py
#!/usr/bin/python

import sys
import re
WORD_RE = re.compile(r"[\w']+") #regex for word classification

for line in sys.stdin:
    components = line.split('\t') #split inpput
    text = " ".join(components[-2:]).strip() #combine subject and c
ontent into text field
    words = re.findall(WORD_RE, text)
    for word in words:
        print components[0] + '\t' + word + '\t' + components[1] #p
rint ID, word, spam flag
```

```
Overwriting mapper.py
```

In [37]:
```
%%writefile reducer.py
#!/usr/bin/python
import sys

emails = set() #hold email IDs
words = {} #hold words and associated counts
spam_emails = 0 #how many emails are marked as spam
spam_word_count = 0 #how many words in spam
ham_word_count = 0 #how many words in ham
vocab = set() #unique words in all text

for line in sys.stdin:
    line = line.strip()
    line = line.rstrip()
    components = line.split('\t') #split input

    ID = components[0] #put input into appropriate variables
    word = components[1]
    spam = int(components[2])

    if word not in words.keys(): #add word to words dictionary, giv
e it spam and ham counts
        words[word] = {'spam_count': 0, 'ham_count': 0}
        vocab.add(word)
    if ID not in emails:
        emails.add(ID)
        if spam == 1: #increment spam counter
            spam_emails += 1


    if spam == 1: #if email is spam, increment the spam counter, ot
herwise increment ham counter
        words[word]['spam_count'] += 1
        spam_word_count += 1
    else:
        words[word]['ham_count'] += 1
        ham_word_count += 1


prior_spam = float(spam_emails)/len(emails) #get prior probabilitie
s
prior_ham = 1-prior_spam

for word in words.keys(): #remove words that have less than 3 count
s from dictionary
    if words[word]['spam_count'] + words[word]['ham_count'] < 3:
        del words[word]

for i, word in words.iteritems(): #use laplace smoother to get cond
itional probabilities
    word['spam_like'] = float(word['spam_count'] + 1)/(spam_word_co
unt + len(vocab))
    word['ham_like'] = float(word['ham_count'] + 1)/(ham_word_count
```

```
+ len(vocab))


print prior_spam #output ham and spam priors
print prior_ham
for word in words.keys():
    #Word "\t" spam likelihood '\t' ham likelihood written to file
    print word + '\t' + str(words[word]['spam_like']) + '\t' + st
r(words[word]['ham_like'])
```

```
Overwriting reducer.py
```

In [38]:
```
#Start hadoop yarn
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-yarn.sh
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-dfs.sh
```

```
starting yarn daemons
starting resourcemanager, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/yarn-dunmireg-resourcemanager-Glenns-Air.hom
e.out
localhost: starting nodemanager, logging to /usr/local/Cellar/hado
op/2.7.1/libexec/logs/yarn-dunmireg-nodemanager-Glenns-Air.home.ou
t
16/01/25 21:27:14 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/hadoop-dunmireg-namenode-Glenns-Air.home.out
localhost: starting datanode, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/hadoop-dunmireg-datanode-Glenns-Air.home.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/Cella
r/hadoop/2.7.1/libexec/logs/hadoop-dunmireg-secondarynamenode-Glen
ns-Air.home.out
16/01/25 21:27:29 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [39]:
```
#add input to hdfs
!hdfs dfs -put enronemail_1h.txt /user/dunmireg
```

```
16/01/25 21:27:50 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [40]:
```
!hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/too
ls/lib/hadoop-streaming-2.7.1.jar \
-mapper mapper.py \
-reducer reducer.py \
-input enronemail_1h.txt \
-output enroneEmailCondProb3
```

```
16/01/25 21:27:54 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:27:54 INFO Configuration.deprecation: session.id is de
precated. Instead, use dfs.metrics.session-id
16/01/25 21:27:54 INFO jvm.JvmMetrics: Initializing JVM Metrics wi
th processName=JobTracker, sessionId=
16/01/25 21:27:54 INFO jvm.JvmMetrics: Cannot initialize JVM Metri
cs with processName=JobTracker, sessionId= - already initialized
16/01/25 21:27:55 INFO mapred.FileInputFormat: Total input paths t
o process : 1
16/01/25 21:27:55 INFO mapreduce.JobSubmitter: number of splits:1
16/01/25 21:27:55 INFO mapreduce.JobSubmitter: Submitting tokens f
or job: job_local1432961765_0001
16/01/25 21:27:55 INFO mapreduce.Job: The url to track the job: ht
tp://localhost:8080/
16/01/25 21:27:55 INFO mapreduce.Job: Running job: job_local143296
1765_0001
16/01/25 21:27:55 INFO mapred.LocalJobRunner: OutputCommitter set
in config null
16/01/25 21:27:55 INFO mapred.LocalJobRunner: OutputCommitter is o
rg.apache.hadoop.mapred.FileOutputCommitter
16/01/25 21:27:55 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/25 21:27:55 INFO mapred.LocalJobRunner: Waiting for map task
s
16/01/25 21:27:55 INFO mapred.LocalJobRunner: Starting task: attem
pt_local1432961765_0001_m_000000_0
16/01/25 21:27:55 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/25 21:27:55 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/01/25 21:27:55 INFO mapred.Task:  Using ResourceCalculatorProce
ssTree : null
16/01/25 21:27:55 INFO mapred.MapTask: Processing split: hdfs://lo
calhost:9000/user/dunmireg/enronemail_1h.txt:0+204658
16/01/25 21:27:55 INFO mapred.MapTask: numReduceTasks: 1
16/01/25 21:27:55 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(10
4857584)
16/01/25 21:27:55 INFO mapred.MapTask: mapreduce.task.io.sort.mb:
100
16/01/25 21:27:55 INFO mapred.MapTask: soft limit at 83886080
16/01/25 21:27:55 INFO mapred.MapTask: bufstart = 0; bufvoid = 104
857600
16/01/25 21:27:55 INFO mapred.MapTask: kvstart = 26214396; length
= 6553600
16/01/25 21:27:55 INFO mapred.MapTask: Map output collector class
= org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/25 21:27:55 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW2/./mapper.py]
16/01/25 21:27:55 INFO Configuration.deprecation: mapred.tip.id is
deprecated. Instead, use mapreduce.task.id
16/01/25 21:27:55 INFO Configuration.deprecation: mapred.local.dir
```

```
is deprecated. Instead, use mapreduce.cluster.local.dir
16/01/25 21:27:55 INFO Configuration.deprecation: map.input.file i
s deprecated. Instead, use mapreduce.map.input.file
16/01/25 21:27:55 INFO Configuration.deprecation: mapred.skip.on i
s deprecated. Instead, use mapreduce.job.skiprecords
16/01/25 21:27:55 INFO Configuration.deprecation: map.input.length
is deprecated. Instead, use mapreduce.map.input.length
16/01/25 21:27:55 INFO Configuration.deprecation: mapred.work.outp
ut.dir is deprecated. Instead, use mapreduce.task.output.dir
16/01/25 21:27:55 INFO Configuration.deprecation: map.input.start
is deprecated. Instead, use mapreduce.map.input.start
16/01/25 21:27:55 INFO Configuration.deprecation: mapred.job.id is
deprecated. Instead, use mapreduce.job.id
16/01/25 21:27:55 INFO Configuration.deprecation: user.name is dep
recated. Instead, use mapreduce.job.user.name
16/01/25 21:27:55 INFO Configuration.deprecation: mapred.task.is.m
ap is deprecated. Instead, use mapreduce.task.ismap
16/01/25 21:27:55 INFO Configuration.deprecation: mapred.task.id i
s deprecated. Instead, use mapreduce.task.attempt.id
16/01/25 21:27:55 INFO Configuration.deprecation: mapred.task.part
ition is deprecated. Instead, use mapreduce.task.partition
16/01/25 21:27:55 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/25 21:27:55 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/25 21:27:55 INFO streaming.PipeMapRed: Records R/W=72/1
16/01/25 21:27:56 INFO streaming.PipeMapRed: R/W/S=100/19097/0 i
n:NA [rec/s] out:NA [rec/s]
16/01/25 21:27:56 INFO streaming.PipeMapRed: MRErrorThread done
16/01/25 21:27:56 INFO streaming.PipeMapRed: mapRedFinished
16/01/25 21:27:56 INFO mapred.LocalJobRunner:
16/01/25 21:27:56 INFO mapred.MapTask: Starting flush of map outpu
t
16/01/25 21:27:56 INFO mapred.MapTask: Spilling map output
16/01/25 21:27:56 INFO mapred.MapTask: bufstart = 0; bufend = 1032
108; bufvoid = 104857600
16/01/25 21:27:56 INFO mapred.MapTask: kvstart = 26214396(10485758
4); kvend = 26082748(104330992); length = 131649/6553600
16/01/25 21:27:56 INFO mapred.MapTask: Finished spill 0
16/01/25 21:27:56 INFO mapred.Task: Task:attempt_local1432961765_0
001_m_000000_0 is done. And is in the process of committing
16/01/25 21:27:56 INFO mapred.LocalJobRunner: Records R/W=72/1
16/01/25 21:27:56 INFO mapred.Task: Task 'attempt_local143296176
5_0001_m_000000_0' done.
16/01/25 21:27:56 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local1432961765_0001_m_000000_0
16/01/25 21:27:56 INFO mapred.LocalJobRunner: map task executor co
mplete.
16/01/25 21:27:56 INFO mapred.LocalJobRunner: Waiting for reduce t
asks
16/01/25 21:27:56 INFO mapred.LocalJobRunner: Starting task: attem
pt_local1432961765_0001_r_000000_0
16/01/25 21:27:56 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
```

```
16/01/25 21:27:56 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/01/25 21:27:56 INFO mapred.Task:  Using ResourceCalculatorProce
ssTree : null
16/01/25 21:27:56 INFO mapred.ReduceTask: Using ShuffleConsumerPlu
gin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@4a12a357
16/01/25 21:27:56 INFO reduce.MergeManagerImpl: MergerManager: mem
oryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshol
d=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/01/25 21:27:56 INFO reduce.EventFetcher: attempt_local143296176
5_0001_r_000000_0 Thread started: EventFetcher for fetching Map Co
mpletion Events
16/01/25 21:27:56 INFO reduce.LocalFetcher: localfetcher#1 about t
o shuffle output of map attempt_local1432961765_0001_m_000000_0 de
comp: 1097936 len: 1097940 to MEMORY
16/01/25 21:27:56 INFO reduce.InMemoryMapOutput: Read 1097936 byte
s from map-output for attempt_local1432961765_0001_m_000000_0
16/01/25 21:27:56 INFO reduce.MergeManagerImpl: closeInMemoryFile
-> map-output of size: 1097936, inMemoryMapOutputs.size() -> 1, co
mmitMemory -> 0, usedMemory ->1097936
16/01/25 21:27:56 INFO reduce.EventFetcher: EventFetcher is interr
upted.. Returning
16/01/25 21:27:56 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 21:27:56 INFO reduce.MergeManagerImpl: finalMerge called
with 1 in-memory map-outputs and 0 on-disk map-outputs
16/01/25 21:27:56 INFO mapred.Merger: Merging 1 sorted segments
16/01/25 21:27:56 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 1097911 bytes
16/01/25 21:27:56 INFO reduce.MergeManagerImpl: Merged 1 segments,
1097936 bytes to disk to satisfy reduce memory limit
16/01/25 21:27:56 INFO reduce.MergeManagerImpl: Merging 1 files, 1
097940 bytes from disk
16/01/25 21:27:56 INFO reduce.MergeManagerImpl: Merging 0 segment
s, 0 bytes from memory into reduce
16/01/25 21:27:56 INFO mapred.Merger: Merging 1 sorted segments
16/01/25 21:27:56 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 1097911 bytes
16/01/25 21:27:56 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 21:27:56 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW2/./reducer.py]
16/01/25 21:27:56 INFO Configuration.deprecation: mapred.job.track
er is deprecated. Instead, use mapreduce.jobtracker.address
16/01/25 21:27:56 INFO Configuration.deprecation: mapred.map.tasks
is deprecated. Instead, use mapreduce.job.maps
16/01/25 21:27:56 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/25 21:27:56 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/25 21:27:56 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/25 21:27:56 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/25 21:27:56 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA
[rec/s] out:NA [rec/s]
```

```
16/01/25 21:27:56 INFO mapreduce.Job: Job job_local1432961765_0001
running in uber mode : false
16/01/25 21:27:56 INFO mapreduce.Job:  map 100% reduce 0%
16/01/25 21:27:58 INFO streaming.PipeMapRed: Records R/W=32913/1
16/01/25 21:27:58 INFO streaming.PipeMapRed: MRErrorThread done
16/01/25 21:27:58 INFO streaming.PipeMapRed: mapRedFinished
16/01/25 21:27:58 INFO mapred.Task: Task:attempt_local1432961765_0
001_r_000000_0 is done. And is in the process of committing
16/01/25 21:27:58 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 21:27:58 INFO mapred.Task: Task attempt_local1432961765_0
001_r_000000_0 is allowed to commit now
16/01/25 21:27:58 INFO output.FileOutputCommitter: Saved output of
task 'attempt_local1432961765_0001_r_000000_0' to hdfs://localhos
t:9000/user/dunmireg/enroneEmailCondProb3/_temporary/0/task_local1
432961765_0001_r_000000
16/01/25 21:27:58 INFO mapred.LocalJobRunner: Records R/W=32913/1
> reduce
16/01/25 21:27:58 INFO mapred.Task: Task 'attempt_local143296176
5_0001_r_000000_0' done.
16/01/25 21:27:58 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local1432961765_0001_r_000000_0
16/01/25 21:27:58 INFO mapred.LocalJobRunner: reduce task executor
complete.
16/01/25 21:27:59 INFO mapreduce.Job:  map 100% reduce 100%
16/01/25 21:27:59 INFO mapreduce.Job: Job job_local1432961765_0001
completed successfully
16/01/25 21:27:59 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=2407996
                FILE: Number of bytes written=4095648
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=409316
                HDFS: Number of bytes written=79886
                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=100
                Map output records=32913
                Map output bytes=1032108
                Map output materialized bytes=1097940
                Input split bytes=105
                Combine input records=0
                Combine output records=0
                Reduce input groups=100
                Reduce shuffle bytes=1097940
                Reduce input records=32913
                Reduce output records=1883
                Spilled Records=65826
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
```

```
                GC time elapsed (ms)=8
                Total committed heap usage (bytes)=511705088
        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=204658
        File Output Format Counters
                Bytes Written=79886
16/01/25 21:27:59 INFO streaming.StreamJob: Output directory: enro
neEmailCondProb3
```

In [41]:
```
!hadoop fs -copyToLocal /user/dunmireg/enroneEmailCondProb3
```

```
16/01/25 21:28:03 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:28:03 WARN hdfs.DFSClient: DFSInputStream has been clo
sed already
16/01/25 21:28:03 WARN hdfs.DFSClient: DFSInputStream has been clo
sed already
```

In [42]:
```
#Remove output directory and stop yarn and hdfs
!hadoop fs -rmr /user/dunmireg/enronemail_1h.txt
!hadoop fs -rmr /user/dunmireg/enroneEmailCondProb3
!/usr/local/Cellar/hadoop/2.7.1/sbin/stop-yarn.sh
!/usr/local/Cellar/hadoop/2.7.1/sbin/stop-dfs.sh
```

```
rmr: DEPRECATED: Please use 'rm -r' instead.
16/01/25 21:28:06 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:28:07 INFO fs.TrashPolicyDefault: Namenode trash confi
guration: Deletion interval = 0 minutes, Emptier interval = 0 minu
tes.
Deleted /user/dunmireg/enronemail_1h.txt
rmr: DEPRECATED: Please use 'rm -r' instead.
16/01/25 21:28:08 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:28:08 INFO fs.TrashPolicyDefault: Namenode trash confi
guration: Deletion interval = 0 minutes, Emptier interval = 0 minu
tes.
Deleted /user/dunmireg/enroneEmailCondProb3
stopping yarn daemons
stopping resourcemanager
localhost: stopping nodemanager
no proxyserver to stop
16/01/25 21:28:20 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here 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 21:28:38 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [43]:
```
%%writefile mapper.py
#!/usr/bin/python

#I have placed the mapper here but have not modified it in any way
from the previous mapper. It will still
#produce ID + \t + word + \t + true spam flag to send to the reduce
r.
import sys
import re
import os
from math import log
from math import exp

priorSpam = 0 #priors
priorHam = 0
words = {}

with open(os.path.join('./enroneEmailCondProbLaplace', 'part-0000
0'), 'r') as myfile: #read input file
    lines = myfile.readlines() #parse lines
    priorSpam = float(lines[0]) #get priors
    priorHam = float(lines[1])
    for line in lines[2:]:
        line = line.strip()
        line = line.rstrip()
        components = line.split('\t') #split remaining lines and ad
d word and likelihoods to dictionary
        words[components[0]] = {'spam_like': float(components[1]),
'ham_like': float(components[2])}


WORD_RE = re.compile(r"[\w']+")

for line in sys.stdin: #read input
    line = line.strip()
    line = line.rstrip()
    components = line.split('\t')
    text = " ".join(components[-2:]).strip() #get subject and conte
nt together as text
    text = re.findall(WORD_RE, text)

    spamScore = log(priorSpam) #get priors
    hamScore = log(priorHam)
    for word in text:
        if word in words.keys(): #increment scores based on word co
nditional probabilities
            spamScore += log(float(words[word]['spam_like']))
            hamScore += log(float(words[word]['ham_like']))
        pred = 0 #predicted class
    if spamScore > hamScore:
        pred = 1
    #output ID, spam flag, predicted class, and exponentiated condi
tional probabilities for document
```

```
    print components[0] + '\t' + components[1] + '\t' + str(pred) +
'\t' + str(exp(spamScore)) + '\t' + str(exp(hamScore))
```

```
Overwriting mapper.py
```

In [45]:
```
%%writefile reducer.py
#!/usr/bin/python
import sys

misclassified = 0 #keep track of how many are misclassified

for line in sys.stdin:
    line = line.strip()
    line = line.rstrip()
    components = line.split('\t')
    if int(components[1]) != int(components[2]):
            misclassified += 1 #if predicted and true flag disagree
increment counter
    print line
print "Misclassified: " + str(misclassified) + " which means this h
as an accuracy of " + str(100-misclassified) + "%"
```

```
Overwriting reducer.py
```

In [30]:
```
#Check code
#!cat enronemail_1h.txt | python mapper.py | python reducer.py
```

In [46]:
```
#Start hadoop yarn
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-yarn.sh
!/usr/local/Cellar/hadoop/2.7.1/sbin/start-dfs.sh
```

```
starting yarn daemons
starting resourcemanager, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/yarn-dunmireg-resourcemanager-Glenns-Air.hom
e.out
localhost: starting nodemanager, logging to /usr/local/Cellar/hado
op/2.7.1/libexec/logs/yarn-dunmireg-nodemanager-Glenns-Air.home.ou
t
16/01/25 21:29:49 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/hadoop-dunmireg-namenode-Glenns-Air.home.out
localhost: starting datanode, logging to /usr/local/Cellar/hadoo
p/2.7.1/libexec/logs/hadoop-dunmireg-datanode-Glenns-Air.home.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/Cella
r/hadoop/2.7.1/libexec/logs/hadoop-dunmireg-secondarynamenode-Glen
ns-Air.home.out
16/01/25 21:30:04 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [47]:
```
#add input to hdfs
!hdfs dfs -put enronemail_1h.txt /user/dunmireg
```

```
16/01/25 21:30:08 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [48]:
```
!hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/too
ls/lib/hadoop-streaming-2.7.1.jar \
-mapper mapper.py \
-reducer reducer.py \
-input enronemail_1h.txt \
-output enroneEmailClass3
```

```
16/01/25 21:30:11 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:30:12 INFO Configuration.deprecation: session.id is de
precated. Instead, use dfs.metrics.session-id
16/01/25 21:30:12 INFO jvm.JvmMetrics: Initializing JVM Metrics wi
th processName=JobTracker, sessionId=
16/01/25 21:30:12 INFO jvm.JvmMetrics: Cannot initialize JVM Metri
cs with processName=JobTracker, sessionId= - already initialized
16/01/25 21:30:12 INFO mapred.FileInputFormat: Total input paths t
o process : 1
16/01/25 21:30:12 INFO mapreduce.JobSubmitter: number of splits:1
16/01/25 21:30:12 INFO mapreduce.JobSubmitter: Submitting tokens f
or job: job_local1804261376_0001
16/01/25 21:30:12 INFO mapreduce.Job: The url to track the job: ht
tp://localhost:8080/
16/01/25 21:30:12 INFO mapred.LocalJobRunner: OutputCommitter set
in config null
16/01/25 21:30:12 INFO mapreduce.Job: Running job: job_local180426
1376_0001
16/01/25 21:30:12 INFO mapred.LocalJobRunner: OutputCommitter is o
rg.apache.hadoop.mapred.FileOutputCommitter
16/01/25 21:30:12 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/25 21:30:13 INFO mapred.LocalJobRunner: Waiting for map task
s
16/01/25 21:30:13 INFO mapred.LocalJobRunner: Starting task: attem
pt_local1804261376_0001_m_000000_0
16/01/25 21:30:13 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/25 21:30:13 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/01/25 21:30:13 INFO mapred.Task:  Using ResourceCalculatorProce
ssTree : null
16/01/25 21:30:13 INFO mapred.MapTask: Processing split: hdfs://lo
calhost:9000/user/dunmireg/enronemail_1h.txt:0+204658
16/01/25 21:30:13 INFO mapred.MapTask: numReduceTasks: 1
16/01/25 21:30:13 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(10
4857584)
16/01/25 21:30:13 INFO mapred.MapTask: mapreduce.task.io.sort.mb:
100
16/01/25 21:30:13 INFO mapred.MapTask: soft limit at 83886080
16/01/25 21:30:13 INFO mapred.MapTask: bufstart = 0; bufvoid = 104
857600
16/01/25 21:30:13 INFO mapred.MapTask: kvstart = 26214396; length
= 6553600
16/01/25 21:30:13 INFO mapred.MapTask: Map output collector class
= org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/25 21:30:13 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW2/./mapper.py]
16/01/25 21:30:13 INFO Configuration.deprecation: mapred.tip.id is
deprecated. Instead, use mapreduce.task.id
16/01/25 21:30:13 INFO Configuration.deprecation: mapred.local.dir
```

```
is deprecated. Instead, use mapreduce.cluster.local.dir
16/01/25 21:30:13 INFO Configuration.deprecation: map.input.file i
s deprecated. Instead, use mapreduce.map.input.file
16/01/25 21:30:13 INFO Configuration.deprecation: mapred.skip.on i
s deprecated. Instead, use mapreduce.job.skiprecords
16/01/25 21:30:13 INFO Configuration.deprecation: map.input.length
is deprecated. Instead, use mapreduce.map.input.length
16/01/25 21:30:13 INFO Configuration.deprecation: mapred.work.outp
ut.dir is deprecated. Instead, use mapreduce.task.output.dir
16/01/25 21:30:13 INFO Configuration.deprecation: map.input.start
is deprecated. Instead, use mapreduce.map.input.start
16/01/25 21:30:13 INFO Configuration.deprecation: mapred.job.id is
deprecated. Instead, use mapreduce.job.id
16/01/25 21:30:13 INFO Configuration.deprecation: user.name is dep
recated. Instead, use mapreduce.job.user.name
16/01/25 21:30:13 INFO Configuration.deprecation: mapred.task.is.m
ap is deprecated. Instead, use mapreduce.task.ismap
16/01/25 21:30:13 INFO Configuration.deprecation: mapred.task.id i
s deprecated. Instead, use mapreduce.task.attempt.id
16/01/25 21:30:13 INFO Configuration.deprecation: mapred.task.part
ition is deprecated. Instead, use mapreduce.task.partition
16/01/25 21:30:13 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/25 21:30:13 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/25 21:30:13 INFO mapreduce.Job: Job job_local1804261376_0001
running in uber mode : false
16/01/25 21:30:13 INFO mapreduce.Job:  map 0% reduce 0%
16/01/25 21:30:14 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:10
0=100/1 [rec/s] out:0=0/1 [rec/s]
16/01/25 21:30:17 INFO streaming.PipeMapRed: MRErrorThread done
16/01/25 21:30:17 INFO streaming.PipeMapRed: Records R/W=100/1
16/01/25 21:30:17 INFO streaming.PipeMapRed: mapRedFinished
16/01/25 21:30:17 INFO mapred.LocalJobRunner:
16/01/25 21:30:17 INFO mapred.MapTask: Starting flush of map outpu
t
16/01/25 21:30:17 INFO mapred.MapTask: Spilling map output
16/01/25 21:30:17 INFO mapred.MapTask: bufstart = 0; bufend = 428
4; bufvoid = 104857600
16/01/25 21:30:17 INFO mapred.MapTask: kvstart = 26214396(10485758
4); kvend = 26214000(104856000); length = 397/6553600
16/01/25 21:30:17 INFO mapred.MapTask: Finished spill 0
16/01/25 21:30:17 INFO mapred.Task: Task:attempt_local1804261376_0
001_m_000000_0 is done. And is in the process of committing
16/01/25 21:30:17 INFO mapred.LocalJobRunner: Records R/W=100/1
16/01/25 21:30:17 INFO mapred.Task: Task 'attempt_local180426137
6_0001_m_000000_0' done.
16/01/25 21:30:17 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local1804261376_0001_m_000000_0
16/01/25 21:30:17 INFO mapred.LocalJobRunner: map task executor co
mplete.
16/01/25 21:30:17 INFO mapred.LocalJobRunner: Waiting for reduce t
asks
16/01/25 21:30:17 INFO mapred.LocalJobRunner: Starting task: attem
```

```
pt_local1804261376_0001_r_000000_0
16/01/25 21:30:17 INFO output.FileOutputCommitter: File Output Com
mitter Algorithm version is 1
16/01/25 21:30:17 INFO util.ProcfsBasedProcessTree: ProcfsBasedPro
cessTree currently is supported only on Linux.
16/01/25 21:30:17 INFO mapred.Task:  Using ResourceCalculatorProce
ssTree : null
16/01/25 21:30:17 INFO mapred.ReduceTask: Using ShuffleConsumerPlu
gin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@60045d35
16/01/25 21:30:17 INFO reduce.MergeManagerImpl: MergerManager: mem
oryLimit=334338464, maxSingleShuffleLimit=83584616, mergeThreshol
d=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/01/25 21:30:17 INFO reduce.EventFetcher: attempt_local180426137
6_0001_r_000000_0 Thread started: EventFetcher for fetching Map Co
mpletion Events
16/01/25 21:30:17 INFO reduce.LocalFetcher: localfetcher#1 about t
o shuffle output of map attempt_local1804261376_0001_m_000000_0 de
comp: 4486 len: 4490 to MEMORY
16/01/25 21:30:17 INFO reduce.InMemoryMapOutput: Read 4486 bytes f
rom map-output for attempt_local1804261376_0001_m_000000_0
16/01/25 21:30:17 INFO reduce.MergeManagerImpl: closeInMemoryFile
-> map-output of size: 4486, inMemoryMapOutputs.size() -> 1, commi
tMemory -> 0, usedMemory ->4486
16/01/25 21:30:17 INFO reduce.EventFetcher: EventFetcher is interr
upted.. Returning
16/01/25 21:30:17 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 21:30:17 INFO reduce.MergeManagerImpl: finalMerge called
with 1 in-memory map-outputs and 0 on-disk map-outputs
16/01/25 21:30:17 INFO mapred.Merger: Merging 1 sorted segments
16/01/25 21:30:17 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 4461 bytes
16/01/25 21:30:17 INFO reduce.MergeManagerImpl: Merged 1 segments,
4486 bytes to disk to satisfy reduce memory limit
16/01/25 21:30:17 INFO reduce.MergeManagerImpl: Merging 1 files, 4
490 bytes from disk
16/01/25 21:30:17 INFO reduce.MergeManagerImpl: Merging 0 segment
s, 0 bytes from memory into reduce
16/01/25 21:30:17 INFO mapred.Merger: Merging 1 sorted segments
16/01/25 21:30:17 INFO mapred.Merger: Down to the last merge-pass,
with 1 segments left of total size: 4461 bytes
16/01/25 21:30:17 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 21:30:17 INFO streaming.PipeMapRed: PipeMapRed exec [/Use
rs/dunmireg/Documents/261HW/HW2/./reducer.py]
16/01/25 21:30:17 INFO Configuration.deprecation: mapred.job.track
er is deprecated. Instead, use mapreduce.jobtracker.address
16/01/25 21:30:17 INFO Configuration.deprecation: mapred.map.tasks
is deprecated. Instead, use mapreduce.job.maps
16/01/25 21:30:17 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [re
c/s] out:NA [rec/s]
16/01/25 21:30:17 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [r
ec/s] out:NA [rec/s]
16/01/25 21:30:17 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA
[rec/s] out:NA [rec/s]
16/01/25 21:30:17 INFO streaming.PipeMapRed: Records R/W=100/1
```

```
16/01/25 21:30:17 INFO streaming.PipeMapRed: MRErrorThread done
16/01/25 21:30:17 INFO streaming.PipeMapRed: mapRedFinished
16/01/25 21:30:17 INFO mapred.Task: Task:attempt_local1804261376_0
001_r_000000_0 is done. And is in the process of committing
16/01/25 21:30:17 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/25 21:30:17 INFO mapred.Task: Task attempt_local1804261376_0
001_r_000000_0 is allowed to commit now
16/01/25 21:30:17 INFO output.FileOutputCommitter: Saved output of
task 'attempt_local1804261376_0001_r_000000_0' to hdfs://localhos
t:9000/user/dunmireg/enroneEmailClass3/_temporary/0/task_local1804
261376_0001_r_000000
16/01/25 21:30:17 INFO mapred.LocalJobRunner: Records R/W=100/1 >
reduce
16/01/25 21:30:17 INFO mapred.Task: Task 'attempt_local180426137
6_0001_r_000000_0' done.
16/01/25 21:30:17 INFO mapred.LocalJobRunner: Finishing task: atte
mpt_local1804261376_0001_r_000000_0
16/01/25 21:30:17 INFO mapred.LocalJobRunner: reduce task executor
complete.
16/01/25 21:30:17 INFO mapreduce.Job:  map 100% reduce 100%
16/01/25 21:30:17 INFO mapreduce.Job: Job job_local1804261376_0001
completed successfully
16/01/25 21:30:17 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=221096
                FILE: Number of bytes written=815286
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=409316
                HDFS: Number of bytes written=4343
                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=100
                Map output records=100
                Map output bytes=4284
                Map output materialized bytes=4490
                Input split bytes=105
                Combine input records=0
                Combine output records=0
                Reduce input groups=100
                Reduce shuffle bytes=4490
                Reduce input records=100
                Reduce output records=101
                Spilled Records=200
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=7
                Total committed heap usage (bytes)=510656512
        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=204658
        File Output Format Counters
                Bytes Written=4343
16/01/25 21:30:17 INFO streaming.StreamJob: Output directory: enro
neEmailClass3
```

In [49]:
```
!hadoop fs -copyToLocal /user/dunmireg/enroneEmailClass3
```

```
16/01/25 21:30:25 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:30:26 WARN hdfs.DFSClient: DFSInputStream has been clo
sed already
16/01/25 21:30:26 WARN hdfs.DFSClient: DFSInputStream has been clo
sed already
```

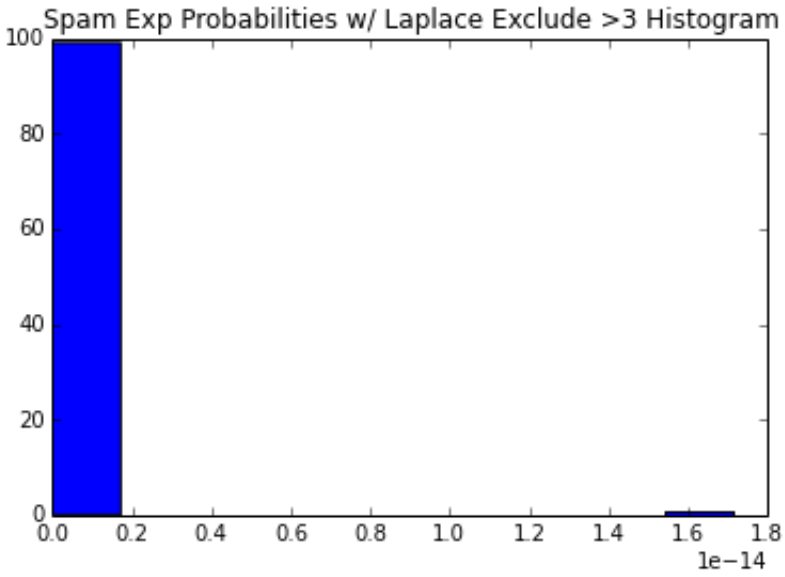
In [50]:
```
#Remove output directory and stop yarn and hdfs
!hadoop fs -rmr /user/dunmireg/enronemail_1h.txt
!hadoop fs -rmr /user/dunmireg/enroneEmailClass3
!/usr/local/Cellar/hadoop/2.7.1/sbin/stop-yarn.sh
!/usr/local/Cellar/hadoop/2.7.1/sbin/stop-dfs.sh
```

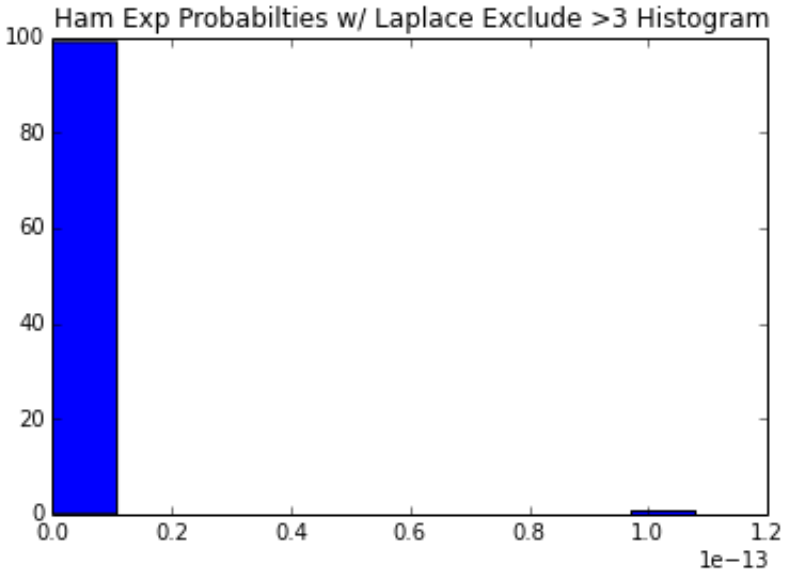
```
rmr: DEPRECATED: Please use 'rm -r' instead.
16/01/25 21:30:28 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:30:28 INFO fs.TrashPolicyDefault: Namenode trash confi
guration: Deletion interval = 0 minutes, Emptier interval = 0 minu
tes.
Deleted /user/dunmireg/enronemail_1h.txt
rmr: DEPRECATED: Please use 'rm -r' instead.
16/01/25 21:30:29 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
16/01/25 21:30:30 INFO fs.TrashPolicyDefault: Namenode trash confi
guration: Deletion interval = 0 minutes, Emptier interval = 0 minu
tes.
Deleted /user/dunmireg/enroneEmailClass3
stopping yarn daemons
stopping resourcemanager
localhost: stopping nodemanager
no proxyserver to stop
16/01/25 21:30:41 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here 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 21:31:00 WARN util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [51]:
```
#Display outputs from results file
import os

with open(os.path.join('./enroneEmailClass3', 'part-00000'), 'r') a
s myfile:
    lines = myfile.readlines() #read file
    lines = lines[-1:] #get last line with results
    for line in lines:
        print line #print results
```

```
Misclassified: 0 which means this has an accuracy of 100%
```

In [9]:
```
%matplotlib inline
#Make histogram
import os
import matplotlib.pyplot as plt

spam_probs = [] #list of spam probabilities
ham_probs = [] #list of ham probabilities
with open(os.path.join('./enroneEmailClass3', 'part-00000'), 'r') a
s myfile: #read file
    lines = myfile.readlines()
    for line in lines[:-1]: #exclude last line which has results
        components = line.split('\t')
        spam_probs.append(float(components[3]))
        ham_probs.append(float(components[4]))

s = plt.figure(1)
plt.hist(spam_probs)
plt.xlabel = "Probability"
plt.ylabel = "Frequency"
plt.title("Spam Exp Probabilities w/ Laplace Exclude >3 Histogram")

h = plt.figure(2)
plt.hist(ham_probs)
plt.xlabel = "Probability"
plt.ylabel = "Frequency"
plt.title("Ham Exp Probabilties w/ Laplace Exclude >3 Histogram")
plt.show()
```





Summary

This has the exact performance of the previous laplace smoother with all tokens included, for a misclassification rate of 0. It should be noted that this classifier also has the same behavior as the laplace only classifier with the spam 'classifier' words having a larger impact on score than the ham 'classifier' words. I am seeing the same thing as before with the vast majority of words contributing small amounts to the score with a handful of words having outsized impact.

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 [6]:
```
#Credit to master solution from week 1 for structure
import numpy as np
from sklearn.naive_bayes import MultinomialNB
from sklearn.feature_extraction.text import CountVectorizer

#make lists to hold email text and classes of correct classificatio
n
emails = []
classes = []
with open('enronemail_1h.txt', 'r') as myfile:
    lines = myfile.readlines()
    for line in lines:
        line = line.strip()
        line = line.rstrip()
        components = line.split('\t') #split text
        text = " ".join(components[-2:]).strip() #join subject and
content into one text field
        emails.append(text)
        classes.append(components[1])

classes = np.array(classes) #convert to array

#initialize count vectorizer to create matrix of token counts
vectorizer = CountVectorizer(min_df=3)
trainingData = vectorizer.fit_transform(emails)


clf = MultinomialNB()
clf.fit(trainingData, classes)
print "  SkLearn Multinomial NB:\t", 1-clf.score(trainingData,class
es)
```

```
  SkLearn Multinomial NB:	0.04
```

Classifier	Misclassification	Accuracy
MapReduce Naive Bayes	0%	100%
Scikit Learn NB	4%	96%

Interestingly, this results in a 4% classification error, meaning an accuracy of 96%. This would suggest that my model above is overfitting the data because I am getting a misclassification rate of 0. I suspect this has to do with how I am incrementing my scores when performing classification, as I am seeing this error rate in all of my classifiers.

In [ ]:
```

```