Homework Week 2

Brandon Shurick

HW2.0

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

A race condition is what can result from two parallel tasks which originate from the same application but are not properly synchronized, so that the order in which the tasks finis in determines the final result. For example, if you write an application which divides the result of some improperly synchronized process by itself plus another value and loads it back to the original variable, the number that you divide may be different depending on if the threads finish near the same time or if one thread reads the result of another in sequence.

MapReduce is a framework that can process large datasets in parallel across multiple nodes.

The Hadoop core framework is a combination of the MapReduce framework plus a distributed file system called Hadoop File System (HDFS).

Hadoop is based on the Functional Programming paradigm, where a function (map) is first applied to each value in the dataset, the data is sorted by key, and then another function (reduce / fold) is applied to all of the values for each key.

Code example below:

In [1]:
```
%%writefile test.py
#!/usr/bin/env python
import re
from itertools import groupby
def mapper(line):
    ''' Map function outputs
        all words for each spam document
    '''
    line = re.sub(r'[^\t\sa-z]+',' ',line.lower())
    words = re.findall(r'[a-z]+',line)
    return('\n'.join('{}\t{}'.format(w,1) for w in words))

def reducer(g):
    ''' Aggregate grouped words
        into counts
    '''
    sums = 0
    for kv in g:
        k,v = kv.split('\t')
        sums += int(v)
    return('{}\t{}'.format(k,sums))

def run():
    ''' Run functional programming steps (mapper, sort, reducer)
        Print all word counts from spam file
    '''
    lines = '\n'.join(mapper(l) for l in open('enronemail_1h.txt','r').readlines() if mapper(l))
    words = sorted(lines.split('\n'))
    print('\n'.join(reducer(g) for k,g in groupby(words)))

if __name__=='__main__':
    run()
```

```
Overwriting test.py
```

In [2]:
```
!python test.py | head -n 10
```

```
a       543
ab      5
abidjan 2
ability 2
able    14
abn     1
about   52
above   11
absent  1
absenteeism     1
close failed in file object destructor:
sys.excepthook is missing
lost sys.stderr
```

HW2.1 Sort in Hadoop MapReduce

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

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

In [3]:
```
!mkdir input
!rmdir output
```

```
mkdir: input: File exists
rmdir: output: Directory not empty
```

In [4]:
```
from __future__ import print_function
import random
def generate_file(N,fname='input/randomrecords.txt'):
    ''' Function to generate random integers '''
    gen_number = lambda n: random.randint(0,n-1)
    nums = '\n'.join('<{},"{}">'.format(gen_number(N),'NA') for n in range(N))
    w = open(fname,'w')
    print(nums,file=w)
    w.close()
generate_file(10000)
```

In [5]:
```
!head -n2 input/randomrecords.txt
```

```
<7168,"NA">
<4898,"NA">
```

In [6]:
```
%%writefile mapper.py
#!/usr/bin/env python
import re, sys
def mapper(line):
    ''' Mapper function for Hadoop '''
    line = re.sub(r'[<>\"]','',line.strip())
    num,word = line.split(',')
    print '{}\t{}'.format(word,num)

for line in sys.stdin:
    mapper(line)
```

```
Overwriting mapper.py
```

In [7]:
```
%%writefile reducer.py
#!/usr/bin/env python
import sys
priorkey = None
values = []

def printkey(values):
    print 'Key: '+key
    print '\nBottom 10:'
    print '\n'.join(str(s) for s in sorted(values)[:10])
    print '\nTop 10:'
    print '\n'.join(str(s) for s in sorted(values,reverse=True)[:10])

for line in sys.stdin:
    ''' Reducer '''
    key,val = line.strip().split('\t')

    if key==priorkey or priorkey is None:
        values.append(int(val))
    else:
        printkey(values)
        values = []
    priorkey = key

# Last line
printkey(values)
```

```
Overwriting reducer.py
```

In [8]:
```
!chmod +x mapper.py && chmod +x reducer.py
```

In [9]:
```
!rm -Rf output
!hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar -input ./input/rand
records.txt -mapper ./mapper.py -reducer ./reducer.py -output ./output
```

```
16/01/26 14:16:09 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
16/01/26 14:16:10 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
16/01/26 14:16:10 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/01/26 14:16:10 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already i
nitialized
16/01/26 14:16:10 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/26 14:16:10 INFO mapreduce.JobSubmitter: number of splits:1
16/01/26 14:16:10 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local937590934_0001
16/01/26 14:16:10 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/26 14:16:10 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/26 14:16:10 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
16/01/26 14:16:10 INFO mapreduce.Job: Running job: job_local937590934_0001
16/01/26 14:16:10 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:16:10 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/26 14:16:10 INFO mapred.LocalJobRunner: Starting task: attempt_local937590934_0001_m_000000_0
16/01/26 14:16:10 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:16:10 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/26 14:16:10 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 14:16:10 INFO mapred.MapTask: Processing split: file:/Users/brandonshurick/School/ML at Scale/HW2/input/randomr
ecords.txt:0+118939
16/01/26 14:16:10 INFO mapred.MapTask: numReduceTasks: 1
16/01/26 14:16:10 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/26 14:16:10 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/26 14:16:10 INFO mapred.MapTask: soft limit at 83886080
16/01/26 14:16:10 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/26 14:16:10 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/26 14:16:10 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/26 14:16:10 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/brandonshurick/School/ML at Scale/HW2/./mapper.py
]
16/01/26 14:16:10 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
16/01/26 14:16:10 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.d
ir
16/01/26 14:16:10 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
16/01/26 14:16:10 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
16/01/26 14:16:10 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.lengt
h
16/01/26 14:16:10 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.outp
ut.dir
16/01/26 14:16:10 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
16/01/26 14:16:10 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
16/01/26 14:16:10 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
16/01/26 14:16:10 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
16/01/26 14:16:10 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
16/01/26 14:16:10 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.parti
tion
16/01/26 14:16:10 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:10 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:10 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:10 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
```

```
16/01/26 14:16:11 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:11 INFO streaming.PipeMapRed: Records R/W=10000/1
16/01/26 14:16:11 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 14:16:11 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 14:16:11 INFO mapred.LocalJobRunner:
16/01/26 14:16:11 INFO mapred.MapTask: Starting flush of map output
16/01/26 14:16:11 INFO mapred.MapTask: Spilling map output
16/01/26 14:16:11 INFO mapred.MapTask: bufstart = 0; bufend = 78939; bufvoid = 104857600
16/01/26 14:16:11 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26174400(104697600); length = 39997/655360
0
16/01/26 14:16:11 INFO mapred.MapTask: Finished spill 0
16/01/26 14:16:11 INFO mapred.Task: Task:attempt_local937590934_0001_m_000000_0 is done. And is in the process of commit
ting
16/01/26 14:16:11 INFO mapred.LocalJobRunner: Records R/W=10000/1
16/01/26 14:16:11 INFO mapred.Task: Task 'attempt_local937590934_0001_m_000000_0' done.
16/01/26 14:16:11 INFO mapred.LocalJobRunner: Finishing task: attempt_local937590934_0001_m_000000_0
16/01/26 14:16:11 INFO mapred.LocalJobRunner: map task executor complete.
16/01/26 14:16:11 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/26 14:16:11 INFO mapred.LocalJobRunner: Starting task: attempt_local937590934_0001_r_000000_0
16/01/26 14:16:11 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:16:11 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/26 14:16:11 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 14:16:11 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@2
a436ebd
16/01/26 14:16:11 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLimit=83584616, me
rgeThreshold=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/01/26 14:16:11 INFO reduce.EventFetcher: attempt_local937590934_0001_r_000000_0 Thread started: EventFetcher for fetc
hing Map Completion Events
16/01/26 14:16:11 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local937590934_0001_m_
000000_0 decomp: 98941 len: 98945 to MEMORY
16/01/26 14:16:11 INFO reduce.InMemoryMapOutput: Read 98941 bytes from map-output for attempt_local937590934_0001_m_0000
00_0
16/01/26 14:16:11 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 98941, inMemoryMapOutputs.size(
) -> 1, commitMemory -> 0, usedMemory ->98941
16/01/26 14:16:11 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/26 14:16:11 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:16:11 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
16/01/26 14:16:11 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 14:16:11 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 98936 bytes
16/01/26 14:16:11 INFO reduce.MergeManagerImpl: Merged 1 segments, 98941 bytes to disk to satisfy reduce memory limit
16/01/26 14:16:11 INFO reduce.MergeManagerImpl: Merging 1 files, 98945 bytes from disk
16/01/26 14:16:11 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/01/26 14:16:11 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 14:16:11 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 98936 bytes
16/01/26 14:16:11 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:16:11 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/brandonshurick/School/ML at Scale/HW2/././reducer.p
y]
16/01/26 14:16:11 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.ad
dress
16/01/26 14:16:11 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
16/01/26 14:16:11 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:11 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:11 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:11 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:11 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:11 INFO streaming.PipeMapRed: Records R/W=10000/1
16/01/26 14:16:11 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 14:16:11 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 14:16:11 INFO mapred.Task: Task:attempt_local937590934_0001_r_000000_0 is done. And is in the process of commit
ting
16/01/26 14:16:11 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:16:11 INFO mapred.Task: Task attempt_local937590934_0001_r_000000_0 is allowed to commit now
16/01/26 14:16:11 INFO output.FileOutputCommitter: Saved output of task 'attempt_local937590934_0001_r_000000_0' to file
:/Users/brandonshurick/School/ML at Scale/HW2/output/_temporary/0/task_local937590934_0001_r_000000
16/01/26 14:16:11 INFO mapred.LocalJobRunner: Records R/W=10000/1 > reduce
16/01/26 14:16:11 INFO mapred.Task: Task 'attempt_local937590934_0001_r_000000_0' done.
16/01/26 14:16:11 INFO mapred.LocalJobRunner: Finishing task: attempt_local937590934_0001_r_000000_0
16/01/26 14:16:11 INFO mapred.LocalJobRunner: reduce task executor complete.
16/01/26 14:16:11 INFO mapreduce.Job: Job job_local937590934_0001 running in uber mode : false
16/01/26 14:16:11 INFO mapreduce.Job:  map 100% reduce 100%
16/01/26 14:16:11 INFO mapreduce.Job: Job job_local937590934_0001 completed successfully
16/01/26 14:16:11 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=647912
                FILE: Number of bytes written=1093542
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=10000
                Map output records=10000
                Map output bytes=78939
                Map output materialized bytes=98945
                Input split bytes=125
                Combine input records=0
                Combine output records=0
```

```
                Reduce input groups=1
                Reduce shuffle bytes=98945
                Reduce input records=10000
                Reduce output records=25
                Spilled Records=20000
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=21
                Total committed heap usage (bytes)=468713472
        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=118939
        File Output Format Counters
                Bytes Written=139
16/01/26 14:16:11 INFO streaming.StreamJob: Output directory: ./output
```

In [10]:
```
!cat output/*
```

```
Key: NA

Bottom 10:
3
4
7
7
8
9
9
11
12
12

Top 10:
9999
9999
9997
9996
9995
9993
9992
9992
9990
9987
```

HW2.2. WORDCOUNT

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

CROSSCHECK: >grep assistance enronemail_1h.txt|cut -d$'\t' -f4| grep assistance|wc -l
8
NOTE "assistance" occurs on 8 lines but how many times does the token occur? 10 times! This is the number we ?are looking for!

In [11]:
```
%%writefile mapper.py
#!/usr/bin/env python
import re, sys
def mapper(line):
    ''' Map function for wordcount in Hadoop
    '''
    line = re.sub(r'[^\t\sa-z]+',' ',line.lower())
    words = re.findall(r'[a-z]+',line)
    print '\n'.join('{}\t{}'.format(w,1) for w in words)

for line in sys.stdin:
    mapper(line)
```

```
Overwriting mapper.py
```

In [12]:
```
%%writefile reducer.py
#!/usr/bin/env python
import sys

sums = 0
prev_sums = 0
prev_k = None
for line in sys.stdin:
    k,v = line.strip().split('\t')
    if k==prev_k or prev_k is None:
        sums += int(v)
    else:
        print '{}\t{}'.format(prev_k,prev_sums)
        sums = 0
        sums += int(v)
    prev_k = k
    prev_sums = sums

# Last line
print '{}\t{}'.format(k,sums)
```

```
Overwriting reducer.py
```

In [13]:
```
!chmod +x mapper.py && chmod +x reducer.py
```

In [14]:
```
!rm -Rf ./output
!hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar -input ./enronemail
h.txt -mapper ./mapper.py -reducer ./reducer.py -output ./output
```

```
16/01/26 14:16:13 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
16/01/26 14:16:13 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
16/01/26 14:16:13 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/01/26 14:16:13 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already i
nitialized
16/01/26 14:16:14 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/26 14:16:14 INFO mapreduce.JobSubmitter: number of splits:1
16/01/26 14:16:14 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local662459164_0001
16/01/26 14:16:14 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/26 14:16:14 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/26 14:16:14 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
16/01/26 14:16:14 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:16:14 INFO mapreduce.Job: Running job: job_local662459164_0001
16/01/26 14:16:14 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/26 14:16:14 INFO mapred.LocalJobRunner: Starting task: attempt_local662459164_0001_m_000000_0
16/01/26 14:16:14 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:16:14 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/26 14:16:14 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 14:16:14 INFO mapred.MapTask: Processing split: file:/Users/brandonshurick/School/ML at Scale/HW2/enronemail_1h
.txt:0+203979
16/01/26 14:16:14 INFO mapred.MapTask: numReduceTasks: 1
16/01/26 14:16:14 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/26 14:16:14 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/26 14:16:14 INFO mapred.MapTask: soft limit at 83886080
16/01/26 14:16:14 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/26 14:16:14 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/26 14:16:14 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/26 14:16:14 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/brandonshurick/School/ML at Scale/HW2/./mapper.py
]
16/01/26 14:16:14 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
16/01/26 14:16:14 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.d
ir
16/01/26 14:16:14 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
16/01/26 14:16:14 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
16/01/26 14:16:14 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.lengt
h
16/01/26 14:16:14 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.outp
ut.dir
16/01/26 14:16:14 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
16/01/26 14:16:14 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
16/01/26 14:16:14 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
16/01/26 14:16:14 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
16/01/26 14:16:14 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
16/01/26 14:16:14 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.parti
tion
16/01/26 14:16:14 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:14 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:14 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:14 INFO streaming.PipeMapRed: Records R/W=100/1
16/01/26 14:16:14 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 14:16:14 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 14:16:14 INFO mapred.LocalJobRunner:
16/01/26 14:16:14 INFO mapred.MapTask: Starting flush of map output
16/01/26 14:16:14 INFO mapred.MapTask: Spilling map output
16/01/26 14:16:14 INFO mapred.MapTask: bufstart = 0; bufend = 243994; bufvoid = 104857600
```

```
16/01/26 14:16:14 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26087896(104351584); length = 126501/65536
00
16/01/26 14:16:15 INFO mapred.MapTask: Finished spill 0
16/01/26 14:16:15 INFO mapred.Task: Task:attempt_local662459164_0001_m_000000_0 is done. And is in the process of commit
ting
16/01/26 14:16:15 INFO mapred.LocalJobRunner: Records R/W=100/1
16/01/26 14:16:15 INFO mapred.Task: Task 'attempt_local662459164_0001_m_000000_0' done.
16/01/26 14:16:15 INFO mapred.LocalJobRunner: Finishing task: attempt_local662459164_0001_m_000000_0
16/01/26 14:16:15 INFO mapred.LocalJobRunner: map task executor complete.
16/01/26 14:16:15 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/26 14:16:15 INFO mapred.LocalJobRunner: Starting task: attempt_local662459164_0001_r_000000_0
16/01/26 14:16:15 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:16:15 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/26 14:16:15 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 14:16:15 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@4
f068d7f
16/01/26 14:16:15 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLimit=83584616, me
rgeThreshold=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/01/26 14:16:15 INFO reduce.EventFetcher: attempt_local662459164_0001_r_000000_0 Thread started: EventFetcher for fetc
hing Map Completion Events
16/01/26 14:16:15 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local662459164_0001_m_
000000_0 decomp: 307248 len: 307252 to MEMORY
16/01/26 14:16:15 INFO reduce.InMemoryMapOutput: Read 307248 bytes from map-output for attempt_local662459164_0001_m_000
000_0
16/01/26 14:16:15 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 307248, inMemoryMapOutputs.size
() -> 1, commitMemory -> 0, usedMemory ->307248
16/01/26 14:16:15 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/26 14:16:15 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:16:15 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
16/01/26 14:16:15 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 14:16:15 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 307244 bytes
16/01/26 14:16:15 INFO reduce.MergeManagerImpl: Merged 1 segments, 307248 bytes to disk to satisfy reduce memory limit
16/01/26 14:16:15 INFO reduce.MergeManagerImpl: Merging 1 files, 307252 bytes from disk
16/01/26 14:16:15 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/01/26 14:16:15 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 14:16:15 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 307244 bytes
16/01/26 14:16:15 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:16:15 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/brandonshurick/School/ML at Scale/HW2/./reducer.p
y]
16/01/26 14:16:15 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.ad
dress
16/01/26 14:16:15 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
16/01/26 14:16:15 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:15 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:15 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:15 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:15 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:15 INFO streaming.PipeMapRed: Records R/W=21001/1
16/01/26 14:16:15 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 14:16:15 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 14:16:15 INFO mapred.Task: Task:attempt_local662459164_0001_r_000000_0 is done. And is in the process of commit
ting
16/01/26 14:16:15 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:16:15 INFO mapred.Task: Task attempt_local662459164_0001_r_000000_0 is allowed to commit now
16/01/26 14:16:15 INFO output.FileOutputCommitter: Saved output of task 'attempt_local662459164_0001_r_000000_0' to file
:/Users/brandonshurick/School/ML at Scale/HW2/output/_temporary/0/task_local662459164_0001_r_000000
16/01/26 14:16:15 INFO mapred.LocalJobRunner: Records R/W=21001/1 > reduce
16/01/26 14:16:15 INFO mapred.Task: Task 'attempt_local662459164_0001_r_000000_0' done.
16/01/26 14:16:15 INFO mapred.LocalJobRunner: Finishing task: attempt_local662459164_0001_r_000000_0
16/01/26 14:16:15 INFO mapred.LocalJobRunner: reduce task executor complete.
16/01/26 14:16:15 INFO mapreduce.Job: Job job_local662459164_0001 running in uber mode : false
16/01/26 14:16:15 INFO mapreduce.Job:  map 100% reduce 100%
16/01/26 14:16:15 INFO mapreduce.Job: Job job_local662459164_0001 completed successfully
16/01/26 14:16:15 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=1234594
                FILE: Number of bytes written=1768433
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=100
                Map output records=31626
                Map output bytes=243994
                Map output materialized bytes=307252
                Input split bytes=119
                Combine input records=0
                Combine output records=0
                Reduce input groups=5068
                Reduce shuffle bytes=307252
                Reduce input records=31626
                Reduce output records=5068
                Spilled Records=63252
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
```

```
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=536870912
        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=203979
        File Output Format Counters
                Bytes Written=50145
16/01/26 14:16:15 INFO streaming.StreamJob: Output directory: ./output
```

In [15]:
```
!cat ./output/* | grep assistance
```

```
assistance      10
```

HW2.2.1

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

In [16]:
```
%%writefile reducer.py
#!/usr/bin/env python
import sys

i = 0

for line in sys.stdin:
    if i<10:
        print line.strip()
        i+=1
```

```
Overwriting reducer.py
```

In [17]:
```
!chmod +x reducer.py
```

In [18]:
```
!rm -Rf output2
!hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar -D mapreduce.job.ou
ut.key.comparator.class=org.apache.hadoop.mapreduce.lib.partition.KeyFieldBasedComparator -D stream.map.output.field.sep
ator='\t' -D stream.num.map.output.key.fields=2 -D mapreduce.partition.keycomparator.options=-k2,2nr -D mapreduce.job.re
ces=1 -mapper /bin/cat -reducer reducer.py -input ./output -output ./output2
```

```
16/01/26 14:16:17 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
16/01/26 14:16:17 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
16/01/26 14:16:17 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/01/26 14:16:17 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already i
nitialized
16/01/26 14:16:18 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/26 14:16:18 INFO mapreduce.JobSubmitter: number of splits:1
16/01/26 14:16:18 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1299087259_0001
16/01/26 14:16:18 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/26 14:16:18 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/26 14:16:18 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
16/01/26 14:16:18 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:16:18 INFO mapreduce.Job: Running job: job_local1299087259_0001
16/01/26 14:16:18 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/26 14:16:18 INFO mapred.LocalJobRunner: Starting task: attempt_local1299087259_0001_m_000000_0
16/01/26 14:16:18 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:16:18 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/26 14:16:18 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 14:16:18 INFO mapred.MapTask: Processing split: file:/Users/brandonshurick/School/ML at Scale/HW2/output/part-0
0000:0+49745
16/01/26 14:16:18 INFO mapred.MapTask: numReduceTasks: 1
16/01/26 14:16:18 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/26 14:16:18 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/26 14:16:18 INFO mapred.MapTask: soft limit at 83886080
16/01/26 14:16:18 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/26 14:16:18 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/26 14:16:18 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/26 14:16:18 INFO streaming.PipeMapRed: PipeMapRed exec [/bin/cat]
16/01/26 14:16:18 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.outp
ut.dir
16/01/26 14:16:18 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
16/01/26 14:16:18 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
16/01/26 14:16:18 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
16/01/26 14:16:18 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
16/01/26 14:16:18 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.d
ir
```

```
16/01/26 14:16:18 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
16/01/26 14:16:18 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
16/01/26 14:16:18 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.lengt
h
16/01/26 14:16:18 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
16/01/26 14:16:18 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
16/01/26 14:16:18 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.parti
tion
16/01/26 14:16:18 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:18 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:18 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:18 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:18 INFO streaming.PipeMapRed: Records R/W=5068/1
16/01/26 14:16:18 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 14:16:18 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 14:16:18 INFO mapred.LocalJobRunner:
16/01/26 14:16:18 INFO mapred.MapTask: Starting flush of map output
16/01/26 14:16:18 INFO mapred.MapTask: Spilling map output
16/01/26 14:16:18 INFO mapred.MapTask: bufstart = 0; bufend = 54813; bufvoid = 104857600
16/01/26 14:16:18 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26194128(104776512); length = 20269/655360
0
16/01/26 14:16:18 INFO mapred.MapTask: Finished spill 0
16/01/26 14:16:18 INFO mapred.Task: Task:attempt_local1299087259_0001_m_000000_0 is done. And is in the process of commi
tting
16/01/26 14:16:18 INFO mapred.LocalJobRunner: Records R/W=5068/1
16/01/26 14:16:18 INFO mapred.Task: Task 'attempt_local1299087259_0001_m_000000_0' done.
16/01/26 14:16:18 INFO mapred.LocalJobRunner: Finishing task: attempt_local1299087259_0001_m_000000_0
16/01/26 14:16:18 INFO mapred.LocalJobRunner: map task executor complete.
16/01/26 14:16:18 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/26 14:16:18 INFO mapred.LocalJobRunner: Starting task: attempt_local1299087259_0001_r_000000_0
16/01/26 14:16:18 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:16:18 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/26 14:16:18 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 14:16:18 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@2
d2d2e4b
16/01/26 14:16:18 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLimit=83584616, me
rgeThreshold=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/01/26 14:16:18 INFO reduce.EventFetcher: attempt_local1299087259_0001_r_000000_0 Thread started: EventFetcher for fet
ching Map Completion Events
16/01/26 14:16:19 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1299087259_0001_m
_000000_0 decomp: 64951 len: 64955 to MEMORY
16/01/26 14:16:19 INFO reduce.InMemoryMapOutput: Read 64951 bytes from map-output for attempt_local1299087259_0001_m_000
000_0
16/01/26 14:16:19 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 64951, inMemoryMapOutputs.size(
) -> 1, commitMemory -> 0, usedMemory ->64951
16/01/26 14:16:19 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/26 14:16:19 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:16:19 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
16/01/26 14:16:19 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 14:16:19 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 64940 bytes
16/01/26 14:16:19 INFO reduce.MergeManagerImpl: Merged 1 segments, 64951 bytes to disk to satisfy reduce memory limit
16/01/26 14:16:19 INFO reduce.MergeManagerImpl: Merging 1 files, 64955 bytes from disk
16/01/26 14:16:19 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/01/26 14:16:19 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 14:16:19 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 64940 bytes
16/01/26 14:16:19 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:16:19 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/brandonshurick/School/ML at Scale/HW2/./reducer.py]
16/01/26 14:16:19 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.ad
dress
16/01/26 14:16:19 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
16/01/26 14:16:19 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:19 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:19 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:19 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:19 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 14:16:19 INFO streaming.PipeMapRed: Records R/W=5068/1
16/01/26 14:16:19 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 14:16:19 INFO mapred.Task: Task:attempt_local1299087259_0001_r_000000_0 is done. And is in the process of commi
tting
16/01/26 14:16:19 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:16:19 INFO mapred.Task: Task attempt_local1299087259_0001_r_000000_0 is allowed to commit now
16/01/26 14:16:19 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1299087259_0001_r_000000_0' to fil
e:/Users/brandonshurick/School/ML at Scale/HW2/output2/_temporary/0/task_local1299087259_0001_r_000000
16/01/26 14:16:19 INFO mapred.LocalJobRunner: Records R/W=5068/1 > reduce
16/01/26 14:16:19 INFO mapred.Task: Task 'attempt_local1299087259_0001_r_000000_0' done.
16/01/26 14:16:19 INFO mapred.LocalJobRunner: Finishing task: attempt_local1299087259_0001_r_000000_0
16/01/26 14:16:19 INFO mapred.LocalJobRunner: reduce task executor complete.
16/01/26 14:16:19 INFO mapreduce.Job: Job job_local1299087259_0001 running in uber mode : false
16/01/26 14:16:19 INFO mapreduce.Job:  map 100% reduce 100%
16/01/26 14:16:19 INFO mapreduce.Job: Job job_local1299087259_0001 completed successfully
16/01/26 14:16:19 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=442346
                FILE: Number of bytes written=997720
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
```

```
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=5068
                Map output records=5068
                Map output bytes=54813
                Map output materialized bytes=64955
                Input split bytes=119
                Combine input records=0
                Combine output records=0
                Reduce input groups=5068
                Reduce shuffle bytes=64955
                Reduce input records=5068
                Reduce output records=10
                Spilled Records=10136
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=536870912
        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=50153
        File Output Format Counters
                Bytes Written=89
16/01/26 14:16:19 INFO streaming.StreamJob: Output directory: ./output2
```

In [19]:
```
!cat output2/*
```

```
the     1247
to      964
and     686
of      566
a       543
you     445
in      418
your    395
ect     382
for     374
```

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 [20]:
```
%%writefile agg_files.sh
#!/usr/bin/env bash
## Compile all data into one training file ##
hams=`ls enron1-training-data-raw/ham/*`
spams=`ls enron1-training-data-raw/spam/*`
rm train_data.txt
for h in ${hams[@]}
do
    echo -e "$h\t0\t`cat $h | tr '\n' ' ' | tr '\r' ' '`" >> train_data.txt
done
for s in ${spams[@]}
do
    echo -e "$s\t1\t`cat $s | tr '\n' ' ' | tr '\r' ' '`" >> train_data.txt
done
```

```
Overwriting agg_files.sh
```

In [21]:
```
!chmod +x agg_files.sh
!./agg_files.sh
```

```
tr: Illegal byte sequence
tr: Illegal byte sequence
tr: Illegal byte sequence
tr: Illegal byte sequence
tr: Illegal byte sequence
tr: Illegal byte sequence
tr: Illegal byte sequence
tr: Illegal byte sequence
tr: Illegal byte sequence
tr: Illegal byte sequence
tr: Illegal byte sequence
tr: Illegal byte sequence
tr: Illegal byte sequence
tr: Illegal byte sequence
tr: Illegal byte sequence
tr: Illegal byte sequence
```

In [22]:
```
%%writefile mapper.py
#!/usr/bin/env python
import re, sys
WORDS = re.compile(r'[\w]+')
for line in sys.stdin:
    ## Read lines from data chunk ##
    # Remove non-word, non-whitespace characters
    components = line.strip().split('\t')
    if components[0]=='*':
        print line
        continue
    spamdoc = int(components[1])
    words = ' '.join(components[2:])
    words = re.sub(r'[^a-z\s\t]+',' ',words.lower())

    ## Compile list of words ##
    wordslist = WORDS.findall(words)
    vocab = set(wordslist)

    ## Count total words in document ##
    totalwords = len(wordslist)

    ## Find words based on user input ##
    for word in vocab:
        word_cnt = len([ w for w in wordslist if w==word ])
        # Send results for each inputted word to reducer
        print('K{}\t{}\t{}\t{}'.format(components[0],word,word_cnt,spamdoc))
```

```
Overwriting mapper.py
```

In [23]:
```
%%writefile reducer.py
#!/usr/bin/env python
import sys
import math

docs = 0
typecnt = {}
prior = {}
findwords = {}
totalwords = {}

prev_cid = None
for line in sys.stdin:
    ## Read in lines from Mapper ##
    line = line.strip()
    components = line.split('\t')
    cid,word,word_cnt,spamdoc = components

    if spamdoc not in findwords:
        wordcnt = {}
        wordcnt[word] = int(word_cnt)
        findwords[spamdoc] = wordcnt
    else:
        wordcnt = findwords[spamdoc]
        if word not in wordcnt:
            wordcnt[word] = int(word_cnt)
        else:
            wordcnt[word] += int(word_cnt)
        findwords[spamdoc] = wordcnt

    if prev_cid!=cid:
        docs += 1

        # Count of class
        if spamdoc not in typecnt: typecnt[spamdoc] = 1
        else: typecnt[spamdoc] += 1

    prev_cid = cid

findwords_spam = '~'.join('{}:{}'.format(w,findwords['1'][w]) for w in findwords['1'])
findwords_ham = '~'.join('{}:{}'.format(w,findwords['0'][w]) for w in findwords['0'])
out = '*\tdocs^{}'.format(docs)
out += '%spamdocs^{}'.format(typecnt['1'])
out += '%hamdocs^{}'.format(typecnt['0'])
out += '%findwords_spam^{}'.format(findwords_spam)
out += '%findwords_ham^{}'.format(findwords_ham)
print out
```

```
Overwriting reducer.py
```

In [24]:
```
%%writefile reducer2.py
#!/usr/bin/env python
import sys, math
''' Make a second pass through the data
    Read in lines from Mapper
    (i.e. 'cat' from previous output)
'''
docs = 0
spamdocs = 0
hamdocs = 0
totalwords_spam = 0
totalwords_ham = 0
prior_spam = 0
prior_ham = 0
spam_prob = 0
ham_prob = 0

findwords_spam = {}
findwords_ham = {}
doccnts = {}

prev_cid = None

for line in sys.stdin:
    line = line.strip()
    components = line.split('\t')
    cid = components[0]
    if cid == '*':
        d,s,h,fs,fh = components[1].split('%')
        docs = int(d.split('^')[1])
        spamdocs = int(s.split('^')[1])
        hamdocs = int(h.split('^')[1])

        findwords_spam = { x.split(':')[0]:int(x.split(':')[1]) \
                        for x in fs.split('^')[1].split('~') }
        findwords_ham = { x.split(':')[0]:int(x.split(':')[1])  \
                        for x in fh.split('^')[1].split('~') }

        totalwords_spam = sum(findwords_spam[w] for w in findwords_spam)
        totalwords_ham = sum(findwords_ham[w] for w in findwords_ham)

        prior_spam = (spamdocs*1.0) / docs
        prior_ham = (hamdocs*1.0) / docs
    elif cid == '':
        continue
    else:
        cid, word, word_cnt, spam = components
        if prev_cid!=cid and prev_cid is not None:
            spam_prob = math.log(prior_spam)
            ham_prob = math.log(prior_ham)

            for w in doccnts:
                if findwords_spam.get(w,0)>0:
                    spam_prob += math.log((findwords_spam.get(w,0)*1.0) / (totalwords_spam))*math.log(doccnts[w])
                if findwords_ham.get(w,0)>0:
                    ham_prob += math.log((findwords_ham.get(w,0)*1.0) / (totalwords_ham))*math.log(doccnts[w])

            doccnts = {}

            if spam_prob>ham_prob:
                print '{}\t{}\t{}\t{}\t{}'.format(cid,1,spam,math.e**spam_prob,math.e**ham_prob)
            else:
                print '{}\t{}\t{}\t{}\t{}'.format(cid,0,spam,math.e**spam_prob,math.e**ham_prob)

        if word not in doccnts:
            doccnts[word] = int(word_cnt)
        else:
            doccnts[word] += int(word_cnt)
        prev_cid = cid

## Last line
spam_prob = math.log(prior_spam)
ham_prob = math.log(prior_ham)

for w in doccnts:
    if findwords_spam.get(w,0)>0:
        spam_prob += math.log((findwords_spam.get(w,0)*1.0) / (totalwords_spam))*math.log(doccnts[w])
    if findwords_ham.get(w,0)>0:
        ham_prob += math.log((findwords_ham.get(w,0)*1.0) / (totalwords_ham))*math.log(doccnts[w])

if spam_prob>ham_prob:
    print '{}\t{}\t{}\t{}\t{}'.format(cid,1,spam,math.e**spam_prob,math.e**ham_prob)
else:
    print '{}\t{}\t{}\t{}\t{}'.format(cid,0,spam,math.e**spam_prob,math.e**ham_prob)
```

```
Overwriting reducer2.py
```

In [25]:
```
!chmod +x mapper.py && chmod +x reducer.py && chmod +x reducer2.py
```

In [26]:
```
!rm -Rf ./output && rm -Rf ./output2
!hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar -input ./train_data
xt -mapper ./mapper.py -reducer ./reducer.py -output ./output
!cat ./output/* > test_data.txt
!cat ./enronemail_1h.txt >> test_data.txt
!hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar -input ./test_data.
t -mapper ./mapper.py -reducer ./reducer2.py -output ./output2
```

```
16/01/26 14:16:54 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
16/01/26 14:16:54 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
16/01/26 14:16:54 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/01/26 14:16:54 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already i
nitialized
16/01/26 14:16:54 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/26 14:16:54 INFO mapreduce.JobSubmitter: number of splits:1
16/01/26 14:16:54 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local2061235449_0001
16/01/26 14:16:55 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/26 14:16:55 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/26 14:16:55 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
16/01/26 14:16:55 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:16:55 INFO mapreduce.Job: Running job: job_local2061235449_0001
16/01/26 14:16:55 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/26 14:16:55 INFO mapred.LocalJobRunner: Starting task: attempt_local2061235449_0001_m_000000_0
16/01/26 14:16:55 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:16:55 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/26 14:16:55 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 14:16:55 INFO mapred.MapTask: Processing split: file:/Users/brandonshurick/School/ML at Scale/HW2/train_data.tx
t:0+5719195
16/01/26 14:16:55 INFO mapred.MapTask: numReduceTasks: 1
16/01/26 14:16:55 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/26 14:16:55 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/26 14:16:55 INFO mapred.MapTask: soft limit at 83886080
16/01/26 14:16:55 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/26 14:16:55 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/26 14:16:55 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/26 14:16:55 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/brandonshurick/School/ML at Scale/HW2/./mapper.py
]
16/01/26 14:16:55 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
16/01/26 14:16:55 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.d
ir
16/01/26 14:16:55 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
16/01/26 14:16:55 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
16/01/26 14:16:55 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.lengt
h
16/01/26 14:16:55 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.outp
ut.dir
16/01/26 14:16:55 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
16/01/26 14:16:55 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
16/01/26 14:16:55 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
16/01/26 14:16:55 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
16/01/26 14:16:55 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
16/01/26 14:16:55 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.parti
tion
16/01/26 14:16:55 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:55 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:55 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:16:55 INFO streaming.PipeMapRed: Records R/W=243/1
16/01/26 14:16:56 INFO mapreduce.Job: Job job_local2061235449_0001 running in uber mode : false
16/01/26 14:16:56 INFO mapreduce.Job:  map 0% reduce 0%
16/01/26 14:16:58 INFO streaming.PipeMapRed: R/W/S=1000/63926/0 in:333=1000/3 [rec/s] out:21308=63926/3 [rec/s]
16/01/26 14:17:01 INFO mapred.LocalJobRunner: Records R/W=243/1 > map
16/01/26 14:17:02 INFO mapreduce.Job:  map 24% reduce 0%
16/01/26 14:17:04 INFO mapred.LocalJobRunner: Records R/W=243/1 > map
16/01/26 14:17:05 INFO mapreduce.Job:  map 39% reduce 0%
16/01/26 14:17:05 INFO streaming.PipeMapRed: Records R/W=3548/244712
16/01/26 14:17:07 INFO mapred.LocalJobRunner: Records R/W=3548/244712 > map
16/01/26 14:17:08 INFO mapreduce.Job:  map 48% reduce 0%
16/01/26 14:17:10 INFO mapred.LocalJobRunner: Records R/W=3548/244712 > map
16/01/26 14:17:11 INFO mapreduce.Job:  map 57% reduce 0%
16/01/26 14:17:13 INFO mapred.LocalJobRunner: Records R/W=3548/244712 > map
16/01/26 14:17:14 INFO mapreduce.Job:  map 65% reduce 0%
16/01/26 14:17:14 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 14:17:14 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 14:17:14 INFO mapred.LocalJobRunner: Records R/W=3548/244712 > map
16/01/26 14:17:14 INFO mapred.MapTask: Starting flush of map output
16/01/26 14:17:14 INFO mapred.MapTask: Spilling map output
16/01/26 14:17:14 INFO mapred.MapTask: bufstart = 0; bufend = 30350301; bufvoid = 104857600
16/01/26 14:17:14 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 24493148(97972592); length = 1721249/65536
00
16/01/26 14:17:15 INFO mapred.MapTask: Finished spill 0
```

```
16/01/26 14:17:15 INFO mapred.Task: Task:attempt_local2061235449_0001_m_000000_0 is done. And is in the process of commi
tting
16/01/26 14:17:15 INFO mapred.LocalJobRunner: Records R/W=3548/244712
16/01/26 14:17:15 INFO mapred.Task: Task 'attempt_local2061235449_0001_m_000000_0' done.
16/01/26 14:17:15 INFO mapred.LocalJobRunner: Finishing task: attempt_local2061235449_0001_m_000000_0
16/01/26 14:17:15 INFO mapred.LocalJobRunner: map task executor complete.
16/01/26 14:17:15 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/26 14:17:15 INFO mapred.LocalJobRunner: Starting task: attempt_local2061235449_0001_r_000000_0
16/01/26 14:17:15 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:17:15 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/26 14:17:15 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 14:17:15 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@6
a9a1e56
16/01/26 14:17:15 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLimit=83584616, me
rgeThreshold=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/01/26 14:17:15 INFO reduce.EventFetcher: attempt_local2061235449_0001_r_000000_0 Thread started: EventFetcher for fet
ching Map Completion Events
16/01/26 14:17:15 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local2061235449_0001_m
_000000_0 decomp: 31210929 len: 31210933 to MEMORY
16/01/26 14:17:15 INFO reduce.InMemoryMapOutput: Read 31210929 bytes from map-output for attempt_local2061235449_0001_m_
000000_0
16/01/26 14:17:15 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 31210929, inMemoryMapOutputs.si
ze() -> 1, commitMemory -> 0, usedMemory ->31210929
16/01/26 14:17:15 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/26 14:17:15 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:17:15 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
16/01/26 14:17:15 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 14:17:15 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 31210866 bytes
16/01/26 14:17:16 INFO mapreduce.Job:  map 100% reduce 0%
16/01/26 14:17:16 INFO reduce.MergeManagerImpl: Merged 1 segments, 31210929 bytes to disk to satisfy reduce memory limit
16/01/26 14:17:16 INFO reduce.MergeManagerImpl: Merging 1 files, 31210933 bytes from disk
16/01/26 14:17:16 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/01/26 14:17:16 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 14:17:16 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 31210866 bytes
16/01/26 14:17:16 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:17:16 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/brandonshurick/School/ML at Scale/HW2/././reducer.p
y]
16/01/26 14:17:16 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.ad
dress
16/01/26 14:17:16 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
16/01/26 14:17:16 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:16 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:16 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:16 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:16 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:16 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:16 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:17 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:17 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:400000=400000/1 [rec/s] out:0=0/1 [rec/s]
16/01/26 14:17:17 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 14:17:17 INFO streaming.PipeMapRed: Records R/W=430313/1
16/01/26 14:17:17 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 14:17:17 INFO mapred.Task: Task:attempt_local2061235449_0001_r_000000_0 is done. And is in the process of commi
tting
16/01/26 14:17:17 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:17:17 INFO mapred.Task: Task attempt_local2061235449_0001_r_000000_0 is allowed to commit now
16/01/26 14:17:17 INFO output.FileOutputCommitter: Saved output of task 'attempt_local2061235449_0001_r_000000_0' to fil
e:/Users/brandonshurick/School/ML at Scale/HW2/output/_temporary/0/task_local2061235449_0001_r_000000
16/01/26 14:17:17 INFO mapred.LocalJobRunner: Records R/W=430313/1 > reduce
16/01/26 14:17:17 INFO mapred.Task: Task 'attempt_local2061235449_0001_r_000000_0' done.
16/01/26 14:17:17 INFO mapred.LocalJobRunner: Finishing task: attempt_local2061235449_0001_r_000000_0
16/01/26 14:17:17 INFO mapred.LocalJobRunner: reduce task executor complete.
16/01/26 14:17:18 INFO mapreduce.Job:  map 100% reduce 100%
16/01/26 14:17:18 INFO mapreduce.Job: Job job_local2061235449_0001 completed successfully
16/01/26 14:17:18 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=74072382
                FILE: Number of bytes written=94983808
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=5172
                Map output records=430313
                Map output bytes=30350301
                Map output materialized bytes=31210933
                Input split bytes=116
                Combine input records=0
                Combine output records=0
                Reduce input groups=5172
                Reduce shuffle bytes=31210933
                Reduce input records=430313
                Reduce output records=1
                Spilled Records=860626
                Shuffled Maps =1
                Failed Shuffles=0
```

```
                Merged Map outputs=1
                GC time elapsed (ms)=13
                Total committed heap usage (bytes)=536870912
        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=5719195
        File Output Format Counters
                Bytes Written=551491
16/01/26 14:17:18 INFO streaming.StreamJob: Output directory: ./output
16/01/26 14:17:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
16/01/26 14:17:19 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
16/01/26 14:17:19 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/01/26 14:17:19 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already i
nitialized
16/01/26 14:17:20 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/26 14:17:20 INFO mapreduce.JobSubmitter: number of splits:1
16/01/26 14:17:20 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1616742445_0001
16/01/26 14:17:20 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/26 14:17:20 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/26 14:17:20 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
16/01/26 14:17:20 INFO mapreduce.Job: Running job: job_local1616742445_0001
16/01/26 14:17:20 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:17:20 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/26 14:17:20 INFO mapred.LocalJobRunner: Starting task: attempt_local1616742445_0001_m_000000_0
16/01/26 14:17:20 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:17:20 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/26 14:17:20 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 14:17:20 INFO mapred.MapTask: Processing split: file:/Users/brandonshurick/School/ML at Scale/HW2/test_data.txt
:0+751186
16/01/26 14:17:20 INFO mapred.MapTask: numReduceTasks: 1
16/01/26 14:17:20 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/26 14:17:20 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/26 14:17:20 INFO mapred.MapTask: soft limit at 83886080
16/01/26 14:17:20 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/26 14:17:20 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/26 14:17:20 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/26 14:17:20 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/brandonshurick/School/ML at Scale/HW2/./mapper.py
]
16/01/26 14:17:20 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
16/01/26 14:17:20 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.d
ir
16/01/26 14:17:20 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
16/01/26 14:17:20 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
16/01/26 14:17:20 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.lengt
h
16/01/26 14:17:20 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.outp
ut.dir
16/01/26 14:17:20 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
16/01/26 14:17:20 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
16/01/26 14:17:20 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
16/01/26 14:17:20 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
16/01/26 14:17:20 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
16/01/26 14:17:20 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.parti
tion
16/01/26 14:17:20 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:20 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:20 INFO streaming.PipeMapRed: Records R/W=73/1
16/01/26 14:17:20 INFO streaming.PipeMapRed: R/W/S=100/775/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:21 INFO mapreduce.Job: Job job_local1616742445_0001 running in uber mode : false
16/01/26 14:17:21 INFO mapreduce.Job:  map 0% reduce 0%
16/01/26 14:17:22 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 14:17:22 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 14:17:22 INFO mapred.LocalJobRunner:
16/01/26 14:17:22 INFO mapred.MapTask: Starting flush of map output
16/01/26 14:17:22 INFO mapred.MapTask: Spilling map output
16/01/26 14:17:22 INFO mapred.MapTask: bufstart = 0; bufend = 1053874; bufvoid = 104857600
16/01/26 14:17:22 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26156144(104624576); length = 58253/655360
0
16/01/26 14:17:22 INFO mapred.MapTask: Finished spill 0
16/01/26 14:17:22 INFO mapred.Task: Task:attempt_local1616742445_0001_m_000000_0 is done. And is in the process of commi
tting
16/01/26 14:17:22 INFO mapred.LocalJobRunner: Records R/W=73/1
16/01/26 14:17:22 INFO mapred.Task: Task 'attempt_local1616742445_0001_m_000000_0' done.
16/01/26 14:17:22 INFO mapred.LocalJobRunner: Finishing task: attempt_local1616742445_0001_m_000000_0
16/01/26 14:17:22 INFO mapred.LocalJobRunner: map task executor complete.
16/01/26 14:17:22 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/26 14:17:22 INFO mapred.LocalJobRunner: Starting task: attempt_local1616742445_0001_r_000000_0
16/01/26 14:17:22 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:17:22 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
```

```
16/01/26 14:17:22 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 14:17:22 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@2
371e6a0
16/01/26 14:17:22 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLimit=83584616, me
rgeThreshold=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/01/26 14:17:22 INFO reduce.EventFetcher: attempt_local1616742445_0001_r_000000_0 Thread started: EventFetcher for fet
ching Map Completion Events
16/01/26 14:17:22 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1616742445_0001_m
_000000_0 decomp: 1083007 len: 1083011 to MEMORY
16/01/26 14:17:22 INFO reduce.InMemoryMapOutput: Read 1083007 bytes from map-output for attempt_local1616742445_0001_m_0
00000_0
16/01/26 14:17:22 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1083007, inMemoryMapOutputs.siz
e() -> 1, commitMemory -> 0, usedMemory ->1083007
16/01/26 14:17:22 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/26 14:17:22 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:17:22 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
16/01/26 14:17:22 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 14:17:22 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 1083004 bytes
16/01/26 14:17:22 INFO mapreduce.Job:  map 100% reduce 0%
16/01/26 14:17:22 INFO reduce.MergeManagerImpl: Merged 1 segments, 1083007 bytes to disk to satisfy reduce memory limit
16/01/26 14:17:22 INFO reduce.MergeManagerImpl: Merging 1 files, 1083011 bytes from disk
16/01/26 14:17:22 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/01/26 14:17:22 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 14:17:22 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 1083004 bytes
16/01/26 14:17:22 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:17:22 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/brandonshurick/School/ML at Scale/HW2/./reducer2.
py]
16/01/26 14:17:22 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.ad
dress
16/01/26 14:17:22 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
16/01/26 14:17:22 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:22 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:22 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:22 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:22 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:22 INFO streaming.PipeMapRed: Records R/W=14564/1
16/01/26 14:17:22 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 14:17:22 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 14:17:22 INFO mapred.Task: Task:attempt_local1616742445_0001_r_000000_0 is done. And is in the process of commi
tting
16/01/26 14:17:22 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:17:22 INFO mapred.Task: Task attempt_local1616742445_0001_r_000000_0 is allowed to commit now
16/01/26 14:17:22 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1616742445_0001_r_000000_0' to fil
e:/Users/brandonshurick/School/ML at Scale/HW2/output2/_temporary/0/task_local1616742445_0001_r_000000
16/01/26 14:17:22 INFO mapred.LocalJobRunner: Records R/W=14564/1 > reduce
16/01/26 14:17:22 INFO mapred.Task: Task 'attempt_local1616742445_0001_r_000000_0' done.
16/01/26 14:17:22 INFO mapred.LocalJobRunner: Finishing task: attempt_local1616742445_0001_r_000000_0
16/01/26 14:17:22 INFO mapred.LocalJobRunner: reduce task executor complete.
16/01/26 14:17:23 INFO mapreduce.Job:  map 100% reduce 100%
16/01/26 14:17:23 INFO mapreduce.Job: Job job_local1616742445_0001 completed successfully
16/01/26 14:17:23 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=3880518
                FILE: Number of bytes written=4054692
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=101
                Map output records=14564
                Map output bytes=1053874
                Map output materialized bytes=1083011
                Input split bytes=115
                Combine input records=0
                Combine output records=0
                Reduce input groups=102
                Reduce shuffle bytes=1083011
                Reduce input records=14564
                Reduce output records=100
                Spilled Records=29128
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=536870912
        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=751186
        File Output Format Counters
                Bytes Written=6139
```

```
16/01/26 14:17:23 INFO streaming.StreamJob: Output directory: ./output2
```

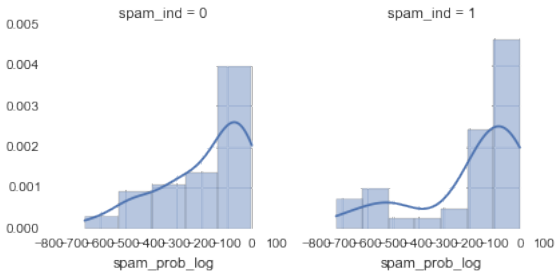
In [27]:
```
!cat ./output2/*
!cat ./output2/* | awk -F'\t' '{if ($2!=$3) {errors+=1;} total+=1; \
                            }END{ print "Error: ",(errors)*100/total,"%" }'
!cat output2/* > results_23.txt
```

```
K0001.1999-12-10.kaminski       0       0       0.290023201856  0.709976798144
K0001.2000-01-17.beck   0       0       0.290023201856  0.709976798144
K0001.2000-06-06.lokay  0       0       6.70140492137e-286      3.81695346508e-274
K0001.2001-02-07.kitchen        1       0       0.0     0.0
K0001.2001-04-02.williams       0       0       1.03833694159e-15       1.92734467667e-12
K0002.1999-12-13.farmer 0       0       6.86683238048e-121      8.51438636562e-118
K0002.2001-02-07.kitchen        0       0       1.67149506758e-193      1.22116737778e-169
K0002.2001-05-25.SA_and_HP      0       1       5.59791991941e-15       1.18625461063e-14
K0002.2003-12-18.GP     0       1       2.40175947244e-19       6.97009430299e-19
K0002.2004-08-01.BG     0       1       2.80355941464e-78       6.73574688998e-73
K0003.1999-12-10.kaminski       0       0       1.76197062059e-62       1.10719670739e-60
K0003.1999-12-14.farmer 0       0       1.19827566766e-25       2.64225719501e-22
K0003.2000-01-17.beck   0       0       5.06911035349e-10       5.81689858483e-09
K0003.2001-02-08.kitchen        0       0       4.78395234639e-102      2.27254812502e-84
K0003.2003-12-18.GP     0       1       2.37378770426e-105      2.09412874381e-93
K0003.2004-08-01.BG     1       1       8.17810889618e-51       8.48387779403e-53
K0004.1999-12-10.kaminski       1       0       8.38926503405e-25       4.58208222415e-25
K0004.1999-12-14.farmer 1       0       1.94928223078e-45       7.66295688258e-55
K0004.2001-04-02.williams       0       0       1.57015358622e-113      3.4830212331e-85
K0004.2001-06-12.SA_and_HP      1       1       5.52761764473e-42       2.03976600519e-43
K0004.2004-08-01.BG     1       1       2.33267150496e-59       1.71270954915e-59
K0005.1999-12-12.kaminski       1       0       1.84853677433e-21       4.26412690212e-22
K0005.1999-12-14.farmer 0       0       6.91429067607e-33       1.43306793348e-28
K0005.2000-06-06.lokay  0       0       8.43138959633e-116      2.86230149932e-89
K0005.2001-02-08.kitchen        0       0       4.06017795313e-21       2.88422472125e-19
K0005.2001-06-23.SA_and_HP      1       1       1.63148320233e-26       8.42553271417e-35
K0005.2003-12-18.GP     0       1       0.290023201856  0.709976798144
K0006.1999-12-13.kaminski       0       0       0.0     0.0
K0006.2001-02-08.kitchen        0       0       1.33342364703e-17       4.11868285537e-17
K0006.2001-04-03.williams       0       0       0.0     0.0
K0006.2001-06-25.SA_and_HP      0       1       6.02598970727e-20       5.28541670964e-18
K0006.2003-12-18.GP     1       1       1.64582319331e-07       1.12869835288e-07
K0006.2004-08-01.BG     0       1       9.0833574284e-56        2.22732004997e-45
K0007.1999-12-13.kaminski       1       0       1.81626934313e-32       7.57270603955e-33
K0007.1999-12-14.farmer 0       0       5.64572254213e-146      1.24281578471e-132
K0007.2000-01-17.beck   1       0       1.76308480813e-38       4.85835521556e-49
K0007.2001-02-09.kitchen        1       0       1.29186007121e-209      8.40113991159e-212
K0007.2003-12-18.GP     0       1       2.85899020562e-99       8.66527141319e-97
K0007.2004-08-01.BG     0       1       3.78955841953e-43       7.40748649327e-40
K0008.2001-02-09.kitchen        0       0       6.42343187865e-105      4.82274463907e-95
K0008.2001-06-12.SA_and_HP      1       1       2.88615653923e-199      1.21029113263e-216
K0008.2001-06-25.SA_and_HP      1       1       2.33267150496e-59       1.71270954915e-59
K0008.2003-12-18.GP     0       1       4.44774750523e-314      1.99141183105e-307
K0008.2004-08-01.BG     1       1       7.36241798157e-55       3.71737695037e-58
K0009.1999-12-13.kaminski       0       0       0.0     0.0
K0009.1999-12-14.farmer 0       0       0.0     0.0
K0009.2000-06-07.lokay  1       0       4.17821295254e-35       3.31096513006e-35
K0009.2001-02-09.kitchen        1       0       1.1211769217e-185       2.24390885243e-188
K0009.2001-06-26.SA_and_HP      0       1       0.0     0.0
K0009.2003-12-18.GP     0       1       1.21636737036e-72       7.13979753921e-65
K0010.1999-12-14.farmer 0       0       3.85717557242e-09       1.73520614314e-08
K0010.1999-12-14.kaminski       0       0       3.11693325254e-74       1.00667054592e-50
K0010.2001-02-09.kitchen        0       0       0.0116501592338 0.0398920831073
K0010.2001-06-28.SA_and_HP      0       1       1.04478504238e-229      4.428079518e-214
K0010.2003-12-18.GP     1       1       3.78309773766e-226      8.61244331496e-244
K0010.2004-08-01.BG     0       1       0.290023201856  0.709976798144
K0011.1999-12-14.farmer 0       0       8.95552043619e-140      8.53953829887e-125
K0011.2001-06-28.SA_and_HP      0       1       3.67428165035e-178      9.44946426219e-171
K0011.2001-06-29.SA_and_HP      1       1       3.00178540095e-225      9.28047696914e-243
K0011.2003-12-18.GP     1       1       0.0     0.0
K0011.2004-08-01.BG     0       1       4.31516699215e-23       7.98434379239e-20
K0012.1999-12-14.farmer 0       0       3.95463206376e-18       9.2925533772e-17
K0012.1999-12-14.kaminski       0       0       4.77159201256e-253      1.3244443039e-243
K0012.2000-01-17.beck   0       0       3.68478572585e-105      1.62276670632e-80
K0012.2000-06-08.lokay  1       0       5.75607466144e-213      1.38295614144e-213
K0012.2001-02-09.kitchen        1       0       5.93578098201e-57       7.54118974164e-58
K0012.2003-12-19.GP     0       1       2.19368956294e-19       2.98148031705e-16
K0013.1999-12-14.farmer 0       0       0.290023201856  0.709976798144
K0013.1999-12-14.kaminski       1       0       1.01772053936e-136      1.02849862882e-140
K0013.2001-04-03.williams       0       0       9.60433631811e-138      3.51523022317e-114
K0013.2001-06-30.SA_and_HP      0       1       1.08210540997e-30       1.10926583389e-28
K0013.2004-08-01.BG     1       1       0.0     0.0
K0014.1999-12-14.kaminski       1       0       4.2329859703e-69        3.04720039093e-74
K0014.1999-12-15.farmer 0       0       1.04379142795e-199      9.00720575997e-169
K0014.2001-02-12.kitchen        0       0       6.85112165523e-71       5.46862323746e-67
K0014.2001-07-04.SA_and_HP      0       1       8.64522322611e-86       9.63674760054e-82
K0014.2003-12-19.GP     1       1       6.91821192506e-295      1.11080367101e-299
K0014.2004-08-01.BG     0       1       0.290023201856  0.709976798144
K0015.1999-12-14.kaminski       0       0       6.22777240655e-16       5.48350812472e-15
```

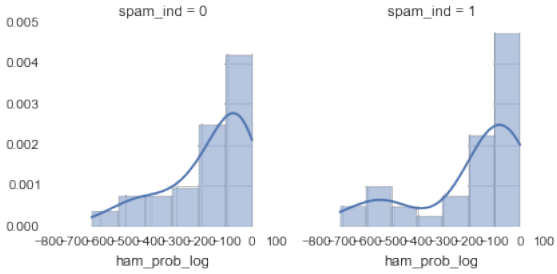
```
K0015.1999-12-15.farmer 0       0       2.41537636371e-36       2.73653084472e-31
K0015.2000-06-09.lokay  0       0       4.28188351459e-47       9.75582823704e-47
K0015.2001-02-12.kitchen        0       0       0.290023201856  0.709976798144
K0015.2001-07-05.SA_and_HP      1       1       0.0     0.0
K0015.2003-12-19.GP     0       1       2.70377936262e-40       1.58497553247e-37
K0016.1999-12-15.farmer 0       0       9.07242397482e-82       2.39730662917e-79
K0016.2001-02-12.kitchen        0       0       8.82210275954e-49       3.36773958848e-42
K0016.2001-07-05.SA_and_HP      1       1       1.78824665904e-83       2.01215691876e-106
K0016.2001-07-06.SA_and_HP      0       1       2.70377936262e-40       1.58497553247e-37
K0016.2003-12-19.GP     1       1       0.0     0.0
K0016.2004-08-01.BG     0       1       5.80310687462e-08       9.94068117865e-08
K0017.1999-12-14.kaminski       0       0       9.40075436488e-44       1.28088778163e-40
K0017.2000-01-17.beck   0       0       7.67216701044e-16       1.71466848901e-14
K0017.2001-04-03.williams       1       0       5.75607466144e-213      1.38295614144e-213
K0017.2003-12-18.GP     0       1       1.93789347381e-27       3.68016930972e-21
K0017.2004-08-01.BG     0       1       9.23373121892e-08       1.16542646598e-07
K0017.2004-08-02.BG     0       1       0.00722166045951        0.0329163968879
K0018.1999-12-14.kaminski       1       0       2.58552541779e-158      1.84465825979e-168
K0018.2001-07-13.SA_and_HP      0       1       5.00500068579e-68       2.67870930578e-65
K0018.2003-12-18.GP     0       1       3.24850199952e-270      5.8833801939e-254
K0018.2003-12-18.GP     1       1       2.70805887098e-231      5.47340895236e-239
Error:  43 %
```

In [28]:
```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import numpy as np
import math
%matplotlib inline
data = pd.read_csv('results_23.txt',sep='\t',header=None)
data.columns = ['id','prediction','spam_ind','spam_prob','ham_prob']
data['spam_prob_log'] = np.log(data['spam_prob'])
data['ham_prob_log'] = np.log(data['ham_prob'])
data = data.replace([np.inf, -np.inf], np.nan)
```

In [29]:
```
g = sns.FacetGrid(data[['spam_prob_log','spam_ind']].dropna(), col="spam_ind")
g = g.map(sns.distplot, 'spam_prob_log')
```



In [30]:
```
g = sns.FacetGrid(data[['ham_prob_log','spam_ind']].dropna(), col="spam_ind")
g = g.map(sns.distplot, 'ham_prob_log')
```



In [31]:
```
print('Zero probability count: {}'.format(np.sum(np.isnan(data['spam_prob_log']))))
```

```
Zero probability count: 10
```

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 [32]:
```
%%writefile reducer2.py
#!/usr/bin/env python
import sys, math
''' Make a second pass through the data
    Read in lines from Mapper
    (i.e. 'cat' from previous output)
'''
docs = 0
spamdocs = 0
hamdocs = 0
totalwords_spam = 0
totalwords_ham = 0
prior_spam = 0
prior_ham = 0
spam_prob = 0
ham_prob = 0

findwords_spam = {}
findwords_ham = {}
doccnts = {}

prev_cid = None

for line in sys.stdin:
    line = line.strip()
    components = line.split('\t')
    cid = components[0]
    if cid == '*':
        d,s,h,fs,fh = components[1].split('%')
        docs = int(d.split('^')[1])
        spamdocs = int(s.split('^')[1])
        hamdocs = int(h.split('^')[1])

        findwords_spam = { x.split(':')[0]:int(x.split(':')[1]) \
                        for x in fs.split('^')[1].split('~') }
        findwords_ham = { x.split(':')[0]:int(x.split(':')[1])  \
                        for x in fh.split('^')[1].split('~') }

        totalwords_spam = sum(findwords_spam[w] for w in findwords_spam)
        totalwords_ham = sum(findwords_ham[w] for w in findwords_ham)

        prior_spam = (spamdocs*1.0) / docs
        prior_ham = (hamdocs*1.0) / docs
    elif cid == '':
        continue
    else:
        cid, word, word_cnt, spam = components
        if prev_cid!=cid and prev_cid is not None:
            spam_prob = math.log(prior_spam)
            ham_prob = math.log(prior_ham)

            for w in doccnts:
                spam_prob += math.log((1+findwords_spam.get(w,0)*1.0) / (totalwords_spam+len(doccnts)))*math.log(doccnts
])
                ham_prob += math.log((1+findwords_ham.get(w,0)*1.0) / (totalwords_ham+len(doccnts)))*math.log(doccnts[w]

            doccnts = {}

            if spam_prob>ham_prob:
                print '{}\t{}\t{}\t{}\t{}'.format(cid,1,spam,math.e**spam_prob,math.e**ham_prob)
            else:
                print '{}\t{}\t{}\t{}\t{}'.format(cid,0,spam,math.e**spam_prob,math.e**ham_prob)

        if word not in doccnts:
            doccnts[word] = int(word_cnt)
        else:
            doccnts[word] += int(word_cnt)
        prev_cid = cid

## Last line
spam_prob = math.log(prior_spam)
ham_prob = math.log(prior_ham)

for w in doccnts:
    spam_prob += math.log((1+findwords_spam.get(w,0)*1.0) / (totalwords_spam+len(doccnts)))*math.log(doccnts[w])
    ham_prob += math.log((1+findwords_ham.get(w,0)*1.0) / (totalwords_ham+len(doccnts)))*math.log(doccnts[w])

if spam_prob>ham_prob:
    print '{}\t{}\t{}\t{}\t{}'.format(cid,1,spam,math.e**spam_prob,math.e**ham_prob)
else:
    print '{}\t{}\t{}\t{}\t{}'.format(cid,0,spam,math.e**spam_prob,math.e**ham_prob)
```

```
Overwriting reducer2.py
```

In [33]:
```
!rm -Rf ./output && rm -Rf ./output2
!hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar -input ./train_data
xt -mapper ./mapper.py -reducer ./reducer.py -output ./output
!cat ./output/* > test_data.txt
!cat ./enronemail_1h.txt >> test_data.txt
!hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar -input ./test_data.
t -mapper ./mapper.py -reducer ./reducer2.py -output ./output2
```

```
16/01/26 14:17:27 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
16/01/26 14:17:28 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
16/01/26 14:17:28 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/01/26 14:17:28 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already i
nitialized
16/01/26 14:17:28 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/26 14:17:28 INFO mapreduce.JobSubmitter: number of splits:1
16/01/26 14:17:28 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1982899152_0001
16/01/26 14:17:28 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/26 14:17:28 INFO mapreduce.Job: Running job: job_local1982899152_0001
16/01/26 14:17:28 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/26 14:17:28 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
16/01/26 14:17:28 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:17:28 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/26 14:17:28 INFO mapred.LocalJobRunner: Starting task: attempt_local1982899152_0001_m_000000_0
16/01/26 14:17:28 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:17:28 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/26 14:17:28 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 14:17:28 INFO mapred.MapTask: Processing split: file:/Users/brandonshurick/School/ML at Scale/HW2/train_data.tx
t:0+5719195
16/01/26 14:17:28 INFO mapred.MapTask: numReduceTasks: 1
16/01/26 14:17:28 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/26 14:17:28 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/26 14:17:28 INFO mapred.MapTask: soft limit at 83886080
16/01/26 14:17:28 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/26 14:17:28 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/26 14:17:28 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/26 14:17:28 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/brandonshurick/School/ML at Scale/HW2/./mapper.py
]
16/01/26 14:17:28 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
16/01/26 14:17:28 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.d
ir
16/01/26 14:17:28 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
16/01/26 14:17:28 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
16/01/26 14:17:28 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.lengt
h
16/01/26 14:17:28 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.outp
ut.dir
16/01/26 14:17:28 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
16/01/26 14:17:28 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
16/01/26 14:17:28 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
16/01/26 14:17:28 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
16/01/26 14:17:28 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
16/01/26 14:17:28 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.parti
tion
16/01/26 14:17:28 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:28 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:28 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:29 INFO streaming.PipeMapRed: Records R/W=243/1
16/01/26 14:17:29 INFO mapreduce.Job: Job job_local1982899152_0001 running in uber mode : false
16/01/26 14:17:29 INFO mapreduce.Job:  map 0% reduce 0%
16/01/26 14:17:31 INFO streaming.PipeMapRed: R/W/S=1000/63926/0 in:500=1000/2 [rec/s] out:31963=63926/2 [rec/s]
16/01/26 14:17:34 INFO mapred.LocalJobRunner: Records R/W=243/1 > map
16/01/26 14:17:35 INFO mapreduce.Job:  map 30% reduce 0%
16/01/26 14:17:37 INFO mapred.LocalJobRunner: Records R/W=243/1 > map
16/01/26 14:17:38 INFO mapreduce.Job:  map 42% reduce 0%
16/01/26 14:17:39 INFO streaming.PipeMapRed: Records R/W=3804/269343
16/01/26 14:17:40 INFO mapred.LocalJobRunner: Records R/W=3804/269343 > map
16/01/26 14:17:41 INFO mapreduce.Job:  map 51% reduce 0%
16/01/26 14:17:43 INFO mapred.LocalJobRunner: Records R/W=3804/269343 > map
16/01/26 14:17:44 INFO mapreduce.Job:  map 62% reduce 0%
16/01/26 14:17:46 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 14:17:46 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 14:17:46 INFO mapred.LocalJobRunner: Records R/W=3804/269343 > map
16/01/26 14:17:46 INFO mapred.MapTask: Starting flush of map output
16/01/26 14:17:46 INFO mapred.MapTask: Spilling map output
16/01/26 14:17:46 INFO mapred.MapTask: bufstart = 0; bufend = 30350301; bufvoid = 104857600
16/01/26 14:17:46 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 24493148(97972592); length = 1721249/65536
00
16/01/26 14:17:46 INFO mapred.LocalJobRunner: Records R/W=3804/269343 > sort
16/01/26 14:17:46 INFO mapred.MapTask: Finished spill 0
16/01/26 14:17:47 INFO mapred.Task: Task:attempt_local1982899152_0001_m_000000_0 is done. And is in the process of commi
tting
16/01/26 14:17:47 INFO mapred.LocalJobRunner: Records R/W=3804/269343
16/01/26 14:17:47 INFO mapred.Task: Task 'attempt_local1982899152_0001_m_000000_0' done.
16/01/26 14:17:47 INFO mapred.LocalJobRunner: Finishing task: attempt_local1982899152_0001_m_000000_0
```

```
16/01/26 14:17:47 INFO mapred.LocalJobRunner: map task executor complete.
16/01/26 14:17:47 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/26 14:17:47 INFO mapred.LocalJobRunner: Starting task: attempt_local1982899152_0001_r_000000_0
16/01/26 14:17:47 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:17:47 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/26 14:17:47 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 14:17:47 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@5
898e19a
16/01/26 14:17:47 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLimit=83584616, me
rgeThreshold=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/01/26 14:17:47 INFO reduce.EventFetcher: attempt_local1982899152_0001_r_000000_0 Thread started: EventFetcher for fet
ching Map Completion Events
16/01/26 14:17:47 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1982899152_0001_m
_000000_0 decomp: 31210929 len: 31210933 to MEMORY
16/01/26 14:17:47 INFO reduce.InMemoryMapOutput: Read 31210929 bytes from map-output for attempt_local1982899152_0001_m_
000000_0
16/01/26 14:17:47 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 31210929, inMemoryMapOutputs.si
ze() -> 1, commitMemory -> 0, usedMemory ->31210929
16/01/26 14:17:47 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/26 14:17:47 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:17:47 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
16/01/26 14:17:47 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 14:17:47 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 31210866 bytes
16/01/26 14:17:47 INFO reduce.MergeManagerImpl: Merged 1 segments, 31210929 bytes to disk to satisfy reduce memory limit
16/01/26 14:17:47 INFO reduce.MergeManagerImpl: Merging 1 files, 31210933 bytes from disk
16/01/26 14:17:47 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/01/26 14:17:47 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 14:17:47 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 31210866 bytes
16/01/26 14:17:47 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:17:47 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/brandonshurick/School/ML at Scale/HW2/././reducer.p
y]
16/01/26 14:17:47 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.ad
dress
16/01/26 14:17:47 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
16/01/26 14:17:47 INFO mapreduce.Job:  map 100% reduce 0%
16/01/26 14:17:47 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:47 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:47 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:47 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:47 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:48 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:48 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:48 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:48 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:400000=400000/1 [rec/s] out:0=0/1 [rec/s]
16/01/26 14:17:48 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 14:17:48 INFO streaming.PipeMapRed: Records R/W=430313/1
16/01/26 14:17:48 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 14:17:48 INFO mapred.Task: Task:attempt_local1982899152_0001_r_000000_0 is done. And is in the process of commi
tting
16/01/26 14:17:48 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:17:48 INFO mapred.Task: Task attempt_local1982899152_0001_r_000000_0 is allowed to commit now
16/01/26 14:17:48 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1982899152_0001_r_000000_0' to fil
e:/Users/brandonshurick/School/ML at Scale/HW2/output/_temporary/0/task_local1982899152_0001_r_000000
16/01/26 14:17:48 INFO mapred.LocalJobRunner: Records R/W=430313/1 > reduce
16/01/26 14:17:48 INFO mapred.Task: Task 'attempt_local1982899152_0001_r_000000_0' done.
16/01/26 14:17:48 INFO mapred.LocalJobRunner: Finishing task: attempt_local1982899152_0001_r_000000_0
16/01/26 14:17:48 INFO mapred.LocalJobRunner: reduce task executor complete.
16/01/26 14:17:49 INFO mapreduce.Job:  map 100% reduce 100%
16/01/26 14:17:49 INFO mapreduce.Job: Job job_local1982899152_0001 completed successfully
16/01/26 14:17:49 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=74072382
                FILE: Number of bytes written=94983808
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=5172
                Map output records=430313
                Map output bytes=30350301
                Map output materialized bytes=31210933
                Input split bytes=116
                Combine input records=0
                Combine output records=0
                Reduce input groups=5172
                Reduce shuffle bytes=31210933
                Reduce input records=430313
                Reduce output records=1
                Spilled Records=860626
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=8
                Total committed heap usage (bytes)=536870912
        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=5719195
        File Output Format Counters
                Bytes Written=551491
16/01/26 14:17:49 INFO streaming.StreamJob: Output directory: ./output
16/01/26 14:17:51 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
16/01/26 14:17:51 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
16/01/26 14:17:51 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/01/26 14:17:51 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already i
nitialized
16/01/26 14:17:51 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/26 14:17:51 INFO mapreduce.JobSubmitter: number of splits:1
16/01/26 14:17:51 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1846963993_0001
16/01/26 14:17:51 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/26 14:17:51 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/26 14:17:51 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
16/01/26 14:17:51 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:17:51 INFO mapreduce.Job: Running job: job_local1846963993_0001
16/01/26 14:17:51 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/26 14:17:51 INFO mapred.LocalJobRunner: Starting task: attempt_local1846963993_0001_m_000000_0
16/01/26 14:17:51 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:17:51 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/26 14:17:51 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 14:17:51 INFO mapred.MapTask: Processing split: file:/Users/brandonshurick/School/ML at Scale/HW2/test_data.txt
:0+751186
16/01/26 14:17:51 INFO mapred.MapTask: numReduceTasks: 1
16/01/26 14:17:51 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/26 14:17:51 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/26 14:17:51 INFO mapred.MapTask: soft limit at 83886080
16/01/26 14:17:51 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/26 14:17:51 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/26 14:17:51 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/26 14:17:51 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/brandonshurick/School/ML at Scale/HW2/./mapper.py
]
16/01/26 14:17:51 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
16/01/26 14:17:51 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.d
ir
16/01/26 14:17:51 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
16/01/26 14:17:51 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
16/01/26 14:17:51 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.lengt
h
16/01/26 14:17:51 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.outp
ut.dir
16/01/26 14:17:51 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
16/01/26 14:17:51 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
16/01/26 14:17:51 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
16/01/26 14:17:51 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
16/01/26 14:17:51 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
16/01/26 14:17:51 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.parti
tion
16/01/26 14:17:52 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:52 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:52 INFO streaming.PipeMapRed: Records R/W=73/1
16/01/26 14:17:52 INFO streaming.PipeMapRed: R/W/S=100/775/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:52 INFO mapreduce.Job: Job job_local1846963993_0001 running in uber mode : false
16/01/26 14:17:52 INFO mapreduce.Job:  map 0% reduce 0%
16/01/26 14:17:53 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 14:17:53 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 14:17:53 INFO mapred.LocalJobRunner:
16/01/26 14:17:53 INFO mapred.MapTask: Starting flush of map output
16/01/26 14:17:53 INFO mapred.MapTask: Spilling map output
16/01/26 14:17:53 INFO mapred.MapTask: bufstart = 0; bufend = 1053874; bufvoid = 104857600
16/01/26 14:17:53 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26156144(104624576); length = 58253/655360
0
16/01/26 14:17:53 INFO mapred.MapTask: Finished spill 0
16/01/26 14:17:53 INFO mapred.Task: Task:attempt_local1846963993_0001_m_000000_0 is done. And is in the process of commi
tting
16/01/26 14:17:53 INFO mapred.LocalJobRunner: Records R/W=73/1
16/01/26 14:17:53 INFO mapred.Task: Task 'attempt_local1846963993_0001_m_000000_0' done.
16/01/26 14:17:53 INFO mapred.LocalJobRunner: Finishing task: attempt_local1846963993_0001_m_000000_0
16/01/26 14:17:53 INFO mapred.LocalJobRunner: map task executor complete.
16/01/26 14:17:53 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/26 14:17:53 INFO mapred.LocalJobRunner: Starting task: attempt_local1846963993_0001_r_000000_0
16/01/26 14:17:53 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:17:53 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/26 14:17:53 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 14:17:53 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@2
b445205
16/01/26 14:17:53 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLimit=83584616, me
rgeThreshold=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
```

```
16/01/26 14:17:53 INFO reduce.EventFetcher: attempt_local1846963993_0001_r_000000_0 Thread started: EventFetcher for fet
ching Map Completion Events
16/01/26 14:17:53 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1846963993_0001_m
_000000_0 decomp: 1083007 len: 1083011 to MEMORY
16/01/26 14:17:53 INFO reduce.InMemoryMapOutput: Read 1083007 bytes from map-output for attempt_local1846963993_0001_m_0
00000_0
16/01/26 14:17:53 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1083007, inMemoryMapOutputs.siz
e() -> 1, commitMemory -> 0, usedMemory ->1083007
16/01/26 14:17:53 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/26 14:17:53 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:17:53 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
16/01/26 14:17:53 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 14:17:53 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 1083004 bytes
16/01/26 14:17:53 INFO reduce.MergeManagerImpl: Merged 1 segments, 1083007 bytes to disk to satisfy reduce memory limit
16/01/26 14:17:53 INFO reduce.MergeManagerImpl: Merging 1 files, 1083011 bytes from disk
16/01/26 14:17:53 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/01/26 14:17:53 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 14:17:53 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 1083004 bytes
16/01/26 14:17:53 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:17:53 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/brandonshurick/School/ML at Scale/HW2/././reducer2.
py]
16/01/26 14:17:53 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.ad
dress
16/01/26 14:17:53 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
16/01/26 14:17:53 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:53 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:53 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:53 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:53 INFO mapreduce.Job:  map 100% reduce 0%
16/01/26 14:17:53 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:53 INFO streaming.PipeMapRed: Records R/W=14564/1
16/01/26 14:17:53 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 14:17:53 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 14:17:53 INFO mapred.Task: Task:attempt_local1846963993_0001_r_000000_0 is done. And is in the process of commi
tting
16/01/26 14:17:53 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:17:53 INFO mapred.Task: Task attempt_local1846963993_0001_r_000000_0 is allowed to commit now
16/01/26 14:17:53 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1846963993_0001_r_000000_0' to fil
e:/Users/brandonshurick/School/ML at Scale/HW2/output2/_temporary/0/task_local1846963993_0001_r_000000
16/01/26 14:17:53 INFO mapred.LocalJobRunner: Records R/W=14564/1 > reduce
16/01/26 14:17:53 INFO mapred.Task: Task 'attempt_local1846963993_0001_r_000000_0' done.
16/01/26 14:17:53 INFO mapred.LocalJobRunner: Finishing task: attempt_local1846963993_0001_r_000000_0
16/01/26 14:17:53 INFO mapred.LocalJobRunner: reduce task executor complete.
16/01/26 14:17:54 INFO mapreduce.Job:  map 100% reduce 100%
16/01/26 14:17:54 INFO mapreduce.Job: Job job_local1846963993_0001 completed successfully
16/01/26 14:17:54 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=3880518
                FILE: Number of bytes written=4054657
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=101
                Map output records=14564
                Map output bytes=1053874
                Map output materialized bytes=1083011
                Input split bytes=115
                Combine input records=0
                Combine output records=0
                Reduce input groups=102
                Reduce shuffle bytes=1083011
                Reduce input records=14564
                Reduce output records=100
                Spilled Records=29128
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=536870912
        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=751186
        File Output Format Counters
                Bytes Written=6104
16/01/26 14:17:54 INFO streaming.StreamJob: Output directory: ./output2
```

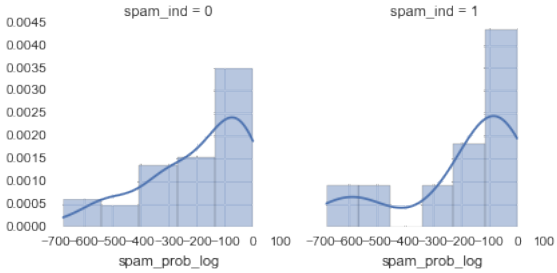
In [34]:
```
!cat ./output2/*
!cat ./output2/* | awk -F'\t' '{if ($2==$3) {corrects+=1;} total+=1; \
                          }END{ print "Error: ",(total-corrects)*100/total,"%" }'
!cat ./output2/* > results_24.txt
```

```
K0001.1999-12-10.kaminski       0       0       0.290023201856  0.709976798144
K0001.2000-01-17.beck   0       0       0.290023201856  0.709976798144
K0001.2000-06-06.lokay  0       0       0.0     5.95350251244e-282
K0001.2001-02-07.kitchen        0       0       0.0     0.0
K0001.2001-04-02.williams       0       0       1.34455009042e-15       1.95195269715e-12
K0002.1999-12-13.farmer 0       0       1.61497523571e-120      2.31313896811e-117
K0002.2001-02-07.kitchen        0       0       2.95576487918e-219      2.11642183132e-169
K0002.2001-05-25.SA_and_HP      0       1       5.87483233206e-15       1.2435447208e-14
K0002.2003-12-18.GP     0       1       2.42323000384e-19       7.00097696062e-19
K0002.2004-08-01.BG     1       1       7.86445675512e-78       1.56824500713e-82
K0003.1999-12-10.kaminski       0       0       6.59896792433e-70       3.01638872205e-68
K0003.1999-12-14.farmer 0       0       6.44720577357e-29       4.80609808381e-26
K0003.2000-01-17.beck   0       0       1.50391009642e-13       5.85838170676e-09
K0003.2001-02-08.kitchen        0       0       5.86355735091e-105      2.54616606545e-84
K0003.2003-12-18.GP     0       1       9.66135811724e-118      3.92756598052e-93
K0003.2004-08-01.BG     1       1       1.08581275097e-50       1.1105506715e-52
K0004.1999-12-10.kaminski       1       0       9.34756695118e-25       4.85191974999e-25
K0004.1999-12-14.farmer 0       0       1.98992573348e-59       1.33925175785e-54
K0004.2001-04-02.williams       0       0       5.79345493451e-127      3.7798492706e-85
K0004.2001-06-12.SA_and_HP      0       1       2.29157740568e-49       2.17328166418e-43
K0004.2004-08-01.BG     1       1       4.93014780996e-59       2.58479841487e-59
K0005.1999-12-12.kaminski       1       0       1.87757476773e-21       4.32445527145e-22
K0005.1999-12-14.farmer 0       0       1.92360069666e-44       1.28221597282e-41
K0005.2000-06-06.lokay  0       0       8.20161484776e-133      3.20504115893e-89
K0005.2001-02-08.kitchen        0       0       8.62696366384e-21       3.42257376455e-19
K0005.2001-06-23.SA_and_HP      0       1       1.95141106342e-43       1.30527051496e-38
K0005.2003-12-18.GP     0       1       0.290023201856  0.709976798144
K0006.1999-12-13.kaminski       1       0       0.0     0.0
K0006.2001-02-08.kitchen        0       0       1.51956459596e-17       4.8800536125e-17
K0006.2001-04-03.williams       0       0       0.0     0.0
K0006.2001-06-25.SA_and_HP      0       1       7.62141840117e-20       5.57950743566e-18
K0006.2003-12-18.GP     1       1       1.66344245187e-07       1.15350922827e-07
K0006.2004-08-01.BG     1       1       1.30953602992e-55       3.30746742545e-62
K0007.1999-12-13.kaminski       1       0       2.96230086268e-32       1.25258743342e-32
K0007.1999-12-14.farmer 0       0       2.33608481527e-156      2.64054259532e-136
K0007.2000-01-17.beck   0       0       9.05511288851e-61       6.2512407045e-49
K0007.2001-02-09.kitchen        0       0       8.20819453728e-236      1.74974965565e-211
K0007.2003-12-18.GP     0       1       7.61299097537e-112      1.51524312095e-106
K0007.2004-08-01.BG     1       1       4.50359966062e-43       2.39014579259e-47
K0008.2001-02-09.kitchen        1       0       9.37273021967e-105      1.81568119626e-143
K0008.2001-06-12.SA_and_HP      0       1       1.60447101944e-290      6.53184335957e-263
K0008.2001-06-25.SA_and_HP      1       1       4.93014780996e-59       2.58479841487e-59
K0008.2003-12-18.GP     1       1       0.0     0.0
K0008.2004-08-01.BG     1       1       8.90349639841e-55       1.93268732223e-61
K0009.1999-12-13.kaminski       1       0       0.0     0.0
K0009.1999-12-14.farmer 0       0       0.0     0.0
K0009.2000-06-07.lokay  0       0       1.04891399095e-38       4.67694065988e-35
K0009.2001-02-09.kitchen        0       0       5.96484272683e-204      1.85444792084e-187
K0009.2001-06-26.SA_and_HP      0       1       0.0     0.0
K0009.2003-12-18.GP     1       1       2.69638979585e-72       4.41243411521e-81
K0010.1999-12-14.farmer 0       0       3.88109146096e-09       1.73962126536e-08
K0010.1999-12-14.kaminski       0       0       4.19318911314e-77       1.12823303144e-50
K0010.2001-02-09.kitchen        0       0       2.06773884746e-06       4.42380599148e-06
K0010.2001-06-28.SA_and_HP      0       1       8.82804130557e-252      1.06040626143e-235
K0010.2003-12-18.GP     1       1       5.69361850477e-245      1.46324979105e-256
K0010.2004-08-01.BG     0       1       0.290023201856  0.709976798144
K0011.1999-12-14.farmer 1       0       1.19974178372e-146      3.90339151647e-173
K0011.2001-06-28.SA_and_HP      0       1       1.40352992421e-206      1.76730681875e-170
K0011.2001-06-29.SA_and_HP      1       1       4.47783853263e-244      1.54902441832e-255
K0011.2003-12-18.GP     1       1       0.0     0.0
K0011.2004-08-01.BG     1       1       4.49955125496e-23       1.49065386858e-31
K0012.1999-12-14.farmer 0       0       3.97482136338e-18       9.31034030991e-17
K0012.1999-12-14.kaminski       0       0       2.2572592113e-293       2.43806136026e-243
K0012.2000-01-17.beck   0       0       1.68716368305e-114      4.42726081248e-92
K0012.2000-06-08.lokay  0       0       1.45041909286e-238      2.88447232327e-213
K0012.2001-02-09.kitchen        0       0       1.32315454198e-62       1.17958082359e-57
K0012.2003-12-19.GP     0       1       2.51484584043e-19       3.02047497602e-16
K0013.1999-12-14.farmer 0       0       0.290023201856  0.709976798144
K0013.1999-12-14.kaminski       0       0       3.20845611634e-157      1.51813660089e-140
K0013.2001-04-03.williams       0       0       2.07603891251e-145      6.09992958531e-124
K0013.2001-06-30.SA_and_HP      0       1       1.12903357039e-30       1.12746488254e-28
K0013.2004-08-01.BG     1       1       0.0     0.0
K0014.1999-12-14.kaminski       0       0       6.48562122156e-75       8.80199292674e-74
K0014.1999-12-15.farmer 0       0       1.43624141563e-206      1.67749569918e-178
K0014.2001-02-12.kitchen        0       0       9.94199222086e-93       6.24123885196e-67
K0014.2001-07-04.SA_and_HP      0       1       4.89508242788e-102      2.45851838821e-99
K0014.2003-12-19.GP     1       1       2.42870996018e-294      7.72787978022e-307
K0014.2004-08-01.BG     0       1       0.290023201856  0.709976798144
K0015.1999-12-14.kaminski       0       0       6.277139505e-16 5.49737775678e-15
K0015.1999-12-15.farmer 0       0       4.84813892678e-36       3.25784769724e-31
```

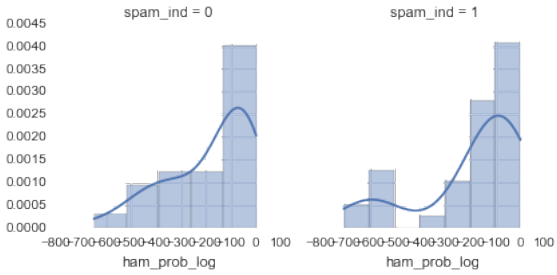
```
K0015.2000-06-09.lokay  0       0       3.82400987565e-69       1.06275065401e-46
K0015.2001-02-12.kitchen        0       0       0.290023201856  0.709976798144
K0015.2001-07-05.SA_and_HP      0       1       0.0     0.0
K0015.2003-12-19.GP     1       1       5.44314812905e-44       2.12474334373e-45
K0016.1999-12-15.farmer 1       0       7.87222633308e-125      3.42294316413e-131
K0016.2001-02-12.kitchen        0       0       2.22284487231e-48       3.63411474264e-42
K0016.2001-07-05.SA_and_HP      0       1       1.3097426289e-120       1.43696088231e-105
K0016.2001-07-06.SA_and_HP      1       1       5.44314812905e-44       2.12474334373e-45
K0016.2003-12-19.GP     1       1       0.0     0.0
K0016.2004-08-01.BG     0       1       5.83276321413e-08       9.99991803064e-08
K0017.1999-12-14.kaminski       1       0       2.78437983066e-53       9.47639545224e-55
K0017.2000-01-17.beck   0       0       8.14993923704e-16       1.74692956218e-14
K0017.2001-04-03.williams       0       0       1.45001296929e-238      2.88406228121e-213
K0017.2003-12-18.GP     0       1       2.72996642867e-27       4.15498365095e-25
K0017.2004-08-01.BG     0       1       9.77772654916e-08       1.22340634761e-07
K0017.2004-08-02.BG     0       1       0.00722377470566        0.0329158135929
K0018.1999-12-14.kaminski       1       0       4.55943900735e-158      1.14998835893e-167
K0018.2001-07-13.SA_and_HP      0       1       1.18448212114e-67       4.64268024088e-65
K0018.2003-12-18.GP     1       1       8.14317240004e-285      2.78617454275e-294
K0018.2003-12-18.GP     1       1       9.91213665527e-231      3.3436789456e-238
Error:  34 %
```

In [35]:
```
data = pd.read_csv('results_24.txt',sep='\t',header=None)
data.columns = ['id','prediction','spam_ind','spam_prob','ham_prob']
data['spam_prob_log'] = np.log(data['spam_prob'])
data['ham_prob_log'] = np.log(data['ham_prob'])
data = data.replace([np.inf, -np.inf], np.nan)
```

In [36]:
```
g = sns.FacetGrid(data[['spam_prob_log','spam_ind']].dropna(), col="spam_ind")
g = g.map(sns.distplot, 'spam_prob_log')
```



In [37]:
```
g = sns.FacetGrid(data[['ham_prob_log','spam_ind']].dropna(), col="spam_ind")
g = g.map(sns.distplot, 'ham_prob_log')
```



In [38]:
```
print('Zero probability count: {}'.format(np.sum(np.isnan(data['spam_prob_log']))))
```

```
Zero probability count: 12
```

Laplace smoothing reduces my error rate because now words that are missing from the vocabulary contribute a small, near-zero probability, rather than completely ignoring the word.

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 [39]:
```
%%writefile reducer2.py
#!/usr/bin/env python
import sys, math
''' Make a second pass through the data
    Read in lines from Mapper
    (i.e. 'cat' from previous output)
```

```
'''
docs = 0
spamdocs = 0
hamdocs = 0
totalwords_spam = 0
totalwords_ham = 0
prior_spam = 0
prior_ham = 0
spam_prob = 0
ham_prob = 0

findwords_spam = {}
findwords_ham = {}
doccnts = {}

prev_cid = None

for line in sys.stdin:
    line = line.strip()
    components = line.split('\t')
    cid = components[0]
    if cid == '*':
        d,s,h,fs,fh = components[1].split('%')
        docs = int(d.split('^')[1])
        spamdocs = int(s.split('^')[1])
        hamdocs = int(h.split('^')[1])

        findwords_spam = { x.split(':')[0]:int(x.split(':')[1]) \
                        for x in fs.split('^')[1].split('~')
                          }
        findwords_ham = { x.split(':')[0]:int(x.split(':')[1])  \
                        for x in fh.split('^')[1].split('~')
                          }

        totalwords_spam = sum(findwords_spam[w] for w in findwords_spam)
        totalwords_ham = sum(findwords_ham[w] for w in findwords_ham)

        prior_spam = (spamdocs*1.0) / docs
        prior_ham = (hamdocs*1.0) / docs
    elif cid == '':
        continue
    else:
        cid, word, word_cnt, spam = components
        if prev_cid!=cid and prev_cid is not None:
            spam_prob = math.log(prior_spam)
            ham_prob = math.log(prior_ham)

            for w in doccnts:
                if (findwords_spam.get(w,0)+findwords_ham.get(w,0))>=3:
                    spam_prob += math.log((1+findwords_spam.get(w,0)*1.0) / (totalwords_spam+len(doccnts)))*math.log(doc
ts[w])
                    ham_prob += math.log((1+findwords_ham.get(w,0)*1.0) / (totalwords_ham+len(doccnts)))*math.log(doccnt
w])

            doccnts = {}

            if spam_prob>ham_prob:
                print '{}\t{}\t{}\t{}\t{}'.format(cid,1,spam,math.e**spam_prob,math.e**ham_prob)
            else:
                print '{}\t{}\t{}\t{}\t{}'.format(cid,0,spam,math.e**spam_prob,math.e**ham_prob)

        if word not in doccnts:
            doccnts[word] = int(word_cnt)
        else:
            doccnts[word] += int(word_cnt)
        prev_cid = cid

## Last line
spam_prob = math.log(prior_spam)
ham_prob = math.log(prior_ham)

for w in doccnts:
    if (findwords_spam.get(w,0)+findwords_ham.get(w,0))>=3:
        spam_prob += math.log((1+findwords_spam.get(w,0)*1.0) / (totalwords_spam+len(doccnts)))*math.log(doccnts[w])
        ham_prob += math.log((1+findwords_ham.get(w,0)*1.0) / (totalwords_ham+len(doccnts)))*math.log(doccnts[w])

if spam_prob>ham_prob:
    print '{}\t{}\t{}\t{}\t{}'.format(cid,1,spam,math.e**spam_prob,math.e**ham_prob)
else:
    print '{}\t{}\t{}\t{}\t{}'.format(cid,0,spam,math.e**spam_prob,math.e**ham_prob)
```

```
Overwriting reducer2.py
```

In [40]:
```
!rm -Rf ./output && rm -Rf ./output2
!hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar -input ./train_data
xt -mapper ./mapper.py -reducer ./reducer.py -output ./output
!cat ./output/* > test_data.txt
!cat ./enronemail_1h.txt >> test_data.txt
!hadoop jar /usr/local/Cellar/hadoop/2.7.1/libexec/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar -input ./test_data.
t -mapper ./mapper.py -reducer ./reducer2.py -output ./output2
```

```
16/01/26 14:17:57 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
16/01/26 14:17:57 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
16/01/26 14:17:57 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/01/26 14:17:57 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already i
nitialized
16/01/26 14:17:57 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/26 14:17:57 INFO mapreduce.JobSubmitter: number of splits:1
16/01/26 14:17:57 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1933670724_0001
16/01/26 14:17:57 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/26 14:17:57 INFO mapreduce.Job: Running job: job_local1933670724_0001
16/01/26 14:17:57 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/26 14:17:57 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
16/01/26 14:17:57 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:17:57 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/26 14:17:57 INFO mapred.LocalJobRunner: Starting task: attempt_local1933670724_0001_m_000000_0
16/01/26 14:17:57 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:17:57 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/26 14:17:57 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 14:17:57 INFO mapred.MapTask: Processing split: file:/Users/brandonshurick/School/ML at Scale/HW2/train_data.tx
t:0+5719195
16/01/26 14:17:57 INFO mapred.MapTask: numReduceTasks: 1
16/01/26 14:17:57 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/26 14:17:57 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/26 14:17:57 INFO mapred.MapTask: soft limit at 83886080
16/01/26 14:17:57 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/26 14:17:57 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/26 14:17:57 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/26 14:17:57 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/brandonshurick/School/ML at Scale/HW2/./mapper.py
]
16/01/26 14:17:57 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
16/01/26 14:17:57 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.d
ir
16/01/26 14:17:57 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
16/01/26 14:17:57 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
16/01/26 14:17:57 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.lengt
h
16/01/26 14:17:57 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.outp
ut.dir
16/01/26 14:17:57 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
16/01/26 14:17:57 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
16/01/26 14:17:57 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
16/01/26 14:17:57 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
16/01/26 14:17:57 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
16/01/26 14:17:57 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.parti
tion
16/01/26 14:17:57 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:57 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:57 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:17:58 INFO streaming.PipeMapRed: Records R/W=243/1
16/01/26 14:17:58 INFO mapreduce.Job: Job job_local1933670724_0001 running in uber mode : false
16/01/26 14:17:58 INFO mapreduce.Job:  map 0% reduce 0%
16/01/26 14:18:00 INFO streaming.PipeMapRed: R/W/S=1000/63926/0 in:500=1000/2 [rec/s] out:31963=63926/2 [rec/s]
16/01/26 14:18:03 INFO mapred.LocalJobRunner: Records R/W=243/1 > map
16/01/26 14:18:04 INFO mapreduce.Job:  map 30% reduce 0%
16/01/26 14:18:06 INFO mapred.LocalJobRunner: Records R/W=243/1 > map
16/01/26 14:18:07 INFO mapreduce.Job:  map 44% reduce 0%
16/01/26 14:18:08 INFO streaming.PipeMapRed: Records R/W=3804/271917
16/01/26 14:18:09 INFO mapred.LocalJobRunner: Records R/W=3804/271917 > map
16/01/26 14:18:10 INFO mapreduce.Job:  map 53% reduce 0%
16/01/26 14:18:12 INFO mapred.LocalJobRunner: Records R/W=3804/271917 > map
16/01/26 14:18:13 INFO mapreduce.Job:  map 63% reduce 0%
16/01/26 14:18:15 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 14:18:15 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 14:18:15 INFO mapred.LocalJobRunner: Records R/W=3804/271917 > map
16/01/26 14:18:15 INFO mapred.MapTask: Starting flush of map output
16/01/26 14:18:15 INFO mapred.MapTask: Spilling map output
16/01/26 14:18:15 INFO mapred.MapTask: bufstart = 0; bufend = 30350301; bufvoid = 104857600
16/01/26 14:18:15 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 24493148(97972592); length = 1721249/65536
00
16/01/26 14:18:15 INFO mapred.MapTask: Finished spill 0
16/01/26 14:18:15 INFO mapred.Task: Task:attempt_local1933670724_0001_m_000000_0 is done. And is in the process of commi
tting
16/01/26 14:18:15 INFO mapred.LocalJobRunner: Records R/W=3804/271917
16/01/26 14:18:15 INFO mapred.Task: Task 'attempt_local1933670724_0001_m_000000_0' done.
16/01/26 14:18:15 INFO mapred.LocalJobRunner: Finishing task: attempt_local1933670724_0001_m_000000_0
16/01/26 14:18:15 INFO mapred.LocalJobRunner: map task executor complete.
```

```
16/01/26 14:18:15 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/26 14:18:15 INFO mapred.LocalJobRunner: Starting task: attempt_local1933670724_0001_r_000000_0
16/01/26 14:18:15 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:18:15 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/26 14:18:15 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 14:18:15 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@2
e8703a1
16/01/26 14:18:15 INFO mapreduce.Job:  map 100% reduce 0%
16/01/26 14:18:16 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLimit=83584616, me
rgeThreshold=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/01/26 14:18:16 INFO reduce.EventFetcher: attempt_local1933670724_0001_r_000000_0 Thread started: EventFetcher for fet
ching Map Completion Events
16/01/26 14:18:16 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1933670724_0001_m
_000000_0 decomp: 31210929 len: 31210933 to MEMORY
16/01/26 14:18:16 INFO reduce.InMemoryMapOutput: Read 31210929 bytes from map-output for attempt_local1933670724_0001_m_
000000_0
16/01/26 14:18:16 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 31210929, inMemoryMapOutputs.si
ze() -> 1, commitMemory -> 0, usedMemory ->31210929
16/01/26 14:18:16 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/26 14:18:16 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:18:16 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
16/01/26 14:18:16 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 14:18:16 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 31210866 bytes
16/01/26 14:18:16 INFO reduce.MergeManagerImpl: Merged 1 segments, 31210929 bytes to disk to satisfy reduce memory limit
16/01/26 14:18:16 INFO reduce.MergeManagerImpl: Merging 1 files, 31210933 bytes from disk
16/01/26 14:18:16 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/01/26 14:18:16 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 14:18:16 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 31210866 bytes
16/01/26 14:18:16 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:18:16 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/brandonshurick/School/ML at Scale/HW2/././reducer.p
y]
16/01/26 14:18:16 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.ad
dress
16/01/26 14:18:16 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
16/01/26 14:18:16 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:18:16 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:18:16 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:18:16 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:18:16 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:18:16 INFO streaming.PipeMapRed: R/W/S=100000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:18:17 INFO streaming.PipeMapRed: R/W/S=200000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:18:17 INFO streaming.PipeMapRed: R/W/S=300000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:18:17 INFO streaming.PipeMapRed: R/W/S=400000/0/0 in:400000=400000/1 [rec/s] out:0=0/1 [rec/s]
16/01/26 14:18:17 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 14:18:17 INFO streaming.PipeMapRed: Records R/W=430313/1
16/01/26 14:18:17 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 14:18:17 INFO mapred.Task: Task:attempt_local1933670724_0001_r_000000_0 is done. And is in the process of commi
tting
16/01/26 14:18:17 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:18:17 INFO mapred.Task: Task attempt_local1933670724_0001_r_000000_0 is allowed to commit now
16/01/26 14:18:17 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1933670724_0001_r_000000_0' to fil
e:/Users/brandonshurick/School/ML at Scale/HW2/output/_temporary/0/task_local1933670724_0001_r_000000
16/01/26 14:18:17 INFO mapred.LocalJobRunner: Records R/W=430313/1 > reduce
16/01/26 14:18:17 INFO mapred.Task: Task 'attempt_local1933670724_0001_r_000000_0' done.
16/01/26 14:18:17 INFO mapred.LocalJobRunner: Finishing task: attempt_local1933670724_0001_r_000000_0
16/01/26 14:18:17 INFO mapred.LocalJobRunner: reduce task executor complete.
16/01/26 14:18:17 INFO mapreduce.Job:  map 100% reduce 100%
16/01/26 14:18:17 INFO mapreduce.Job: Job job_local1933670724_0001 completed successfully
16/01/26 14:18:17 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=74072382
                FILE: Number of bytes written=94983808
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=5172
                Map output records=430313
                Map output bytes=30350301
                Map output materialized bytes=31210933
                Input split bytes=116
                Combine input records=0
                Combine output records=0
                Reduce input groups=5172
                Reduce shuffle bytes=31210933
                Reduce input records=430313
                Reduce output records=1
                Spilled Records=860626
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=9
                Total committed heap usage (bytes)=605028352
        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=5719195
        File Output Format Counters
                Bytes Written=551491
16/01/26 14:18:17 INFO streaming.StreamJob: Output directory: ./output
16/01/26 14:18:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
16/01/26 14:18:19 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
16/01/26 14:18:19 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/01/26 14:18:19 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already i
nitialized
16/01/26 14:18:19 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/26 14:18:19 INFO mapreduce.JobSubmitter: number of splits:1
16/01/26 14:18:19 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local842739678_0001
16/01/26 14:18:19 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/26 14:18:19 INFO mapreduce.Job: Running job: job_local842739678_0001
16/01/26 14:18:19 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/26 14:18:19 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
16/01/26 14:18:19 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:18:19 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/26 14:18:19 INFO mapred.LocalJobRunner: Starting task: attempt_local842739678_0001_m_000000_0
16/01/26 14:18:19 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:18:19 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/26 14:18:19 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 14:18:19 INFO mapred.MapTask: Processing split: file:/Users/brandonshurick/School/ML at Scale/HW2/test_data.txt
:0+751186
16/01/26 14:18:19 INFO mapred.MapTask: numReduceTasks: 1
16/01/26 14:18:20 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/26 14:18:20 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/26 14:18:20 INFO mapred.MapTask: soft limit at 83886080
16/01/26 14:18:20 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/26 14:18:20 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/26 14:18:20 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/26 14:18:20 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/brandonshurick/School/ML at Scale/HW2/./mapper.py
]
16/01/26 14:18:20 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
16/01/26 14:18:20 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.d
ir
16/01/26 14:18:20 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
16/01/26 14:18:20 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
16/01/26 14:18:20 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.lengt
h
16/01/26 14:18:20 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.outp
ut.dir
16/01/26 14:18:20 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
16/01/26 14:18:20 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
16/01/26 14:18:20 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
16/01/26 14:18:20 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
16/01/26 14:18:20 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
16/01/26 14:18:20 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.parti
tion
16/01/26 14:18:20 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:18:20 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:18:20 INFO streaming.PipeMapRed: Records R/W=73/1
16/01/26 14:18:20 INFO streaming.PipeMapRed: R/W/S=100/775/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:18:20 INFO mapreduce.Job: Job job_local842739678_0001 running in uber mode : false
16/01/26 14:18:20 INFO mapreduce.Job:  map 0% reduce 0%
16/01/26 14:18:21 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 14:18:21 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 14:18:21 INFO mapred.LocalJobRunner:
16/01/26 14:18:21 INFO mapred.MapTask: Starting flush of map output
16/01/26 14:18:21 INFO mapred.MapTask: Spilling map output
16/01/26 14:18:21 INFO mapred.MapTask: bufstart = 0; bufend = 1053874; bufvoid = 104857600
16/01/26 14:18:21 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26156144(104624576); length = 58253/655360
0
16/01/26 14:18:21 INFO mapred.MapTask: Finished spill 0
16/01/26 14:18:21 INFO mapred.Task: Task:attempt_local842739678_0001_m_000000_0 is done. And is in the process of commit
ting
16/01/26 14:18:21 INFO mapred.LocalJobRunner: Records R/W=73/1
16/01/26 14:18:21 INFO mapred.Task: Task 'attempt_local842739678_0001_m_000000_0' done.
16/01/26 14:18:21 INFO mapred.LocalJobRunner: Finishing task: attempt_local842739678_0001_m_000000_0
16/01/26 14:18:21 INFO mapred.LocalJobRunner: map task executor complete.
16/01/26 14:18:21 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/26 14:18:21 INFO mapred.LocalJobRunner: Starting task: attempt_local842739678_0001_r_000000_0
16/01/26 14:18:21 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/26 14:18:21 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
16/01/26 14:18:21 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/26 14:18:21 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@2
371e6a0
16/01/26 14:18:21 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLimit=83584616, me
rgeThreshold=220663392, ioSortFactor=10, memToMemMergeOutputsThreshold=10
16/01/26 14:18:21 INFO reduce.EventFetcher: attempt_local842739678_0001_r_000000_0 Thread started: EventFetcher for fetc
```

```
hing Map Completion Events
16/01/26 14:18:21 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local842739678_0001_m_
000000_0 decomp: 1083007 len: 1083011 to MEMORY
16/01/26 14:18:21 INFO reduce.InMemoryMapOutput: Read 1083007 bytes from map-output for attempt_local842739678_0001_m_00
0000_0
16/01/26 14:18:21 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1083007, inMemoryMapOutputs.siz
e() -> 1, commitMemory -> 0, usedMemory ->1083007
16/01/26 14:18:21 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/26 14:18:21 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:18:21 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
16/01/26 14:18:21 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 14:18:21 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 1083004 bytes
16/01/26 14:18:21 INFO reduce.MergeManagerImpl: Merged 1 segments, 1083007 bytes to disk to satisfy reduce memory limit
16/01/26 14:18:21 INFO reduce.MergeManagerImpl: Merging 1 files, 1083011 bytes from disk
16/01/26 14:18:21 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/01/26 14:18:21 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 14:18:21 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 1083004 bytes
16/01/26 14:18:21 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:18:21 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/brandonshurick/School/ML at Scale/HW2/././reducer2.
py]
16/01/26 14:18:21 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.ad
dress
16/01/26 14:18:21 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
16/01/26 14:18:21 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:18:21 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:18:21 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:18:21 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:18:21 INFO mapreduce.Job:  map 100% reduce 0%
16/01/26 14:18:21 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/26 14:18:22 INFO streaming.PipeMapRed: Records R/W=14564/1
16/01/26 14:18:22 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 14:18:22 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 14:18:22 INFO mapred.Task: Task:attempt_local842739678_0001_r_000000_0 is done. And is in the process of commit
ting
16/01/26 14:18:22 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 14:18:22 INFO mapred.Task: Task attempt_local842739678_0001_r_000000_0 is allowed to commit now
16/01/26 14:18:22 INFO output.FileOutputCommitter: Saved output of task 'attempt_local842739678_0001_r_000000_0' to file
:/Users/brandonshurick/School/ML at Scale/HW2/output2/_temporary/0/task_local842739678_0001_r_000000
16/01/26 14:18:22 INFO mapred.LocalJobRunner: Records R/W=14564/1 > reduce
16/01/26 14:18:22 INFO mapred.Task: Task 'attempt_local842739678_0001_r_000000_0' done.
16/01/26 14:18:22 INFO mapred.LocalJobRunner: Finishing task: attempt_local842739678_0001_r_000000_0
16/01/26 14:18:22 INFO mapred.LocalJobRunner: reduce task executor complete.
16/01/26 14:18:22 INFO mapreduce.Job:  map 100% reduce 100%
16/01/26 14:18:22 INFO mapreduce.Job: Job job_local842739678_0001 completed successfully
16/01/26 14:18:22 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=3880518
                FILE: Number of bytes written=4051682
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=101
                Map output records=14564
                Map output bytes=1053874
                Map output materialized bytes=1083011
                Input split bytes=115
                Combine input records=0
                Combine output records=0
                Reduce input groups=102
                Reduce shuffle bytes=1083011
                Reduce input records=14564
                Reduce output records=100
                Spilled Records=29128
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=536870912
        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=751186
        File Output Format Counters
                Bytes Written=6133
16/01/26 14:18:22 INFO streaming.StreamJob: Output directory: ./output2
```

In [41]:
```
!cat ./output2/*
!cat ./output2/* | awk -F'\t' '{if ($2==$3) {corrects+=1;} total+=1; \
                          }END{ print "Error: ",(total-corrects)*100/total,"%" }'
!cat ./output2/* > results_25.txt
```

```
K0001.1999-12-10.kaminski       0       0       0.290023201856  0.709976798144
K0001.2000-01-17.beck   0       0       0.290023201856  0.709976798144
K0001.2000-06-06.lokay  0       0       4.06332926912e-318      3.82899713106e-269
K0001.2001-02-07.kitchen        0       0       0.0     0.0
K0001.2001-04-02.williams       0       0       1.34455009042e-15       1.95195269715e-12
K0002.1999-12-13.farmer 0       0       1.61497523571e-120      2.31313896811e-117
K0002.2001-02-07.kitchen        0       0       2.95576487918e-219      2.11642183132e-169
K0002.2001-05-25.SA_and_HP      0       1       5.87483233206e-15       1.2435447208e-14
K0002.2003-12-18.GP     0       1       2.42323000384e-19       7.00097696062e-19
K0002.2004-08-01.BG     1       1       2.07003051378e-74       1.4144123104e-78
K0003.1999-12-10.kaminski       0       0       2.09628442971e-62       1.14567005624e-60
K0003.1999-12-14.farmer 0       0       3.63336985238e-25       4.33422552295e-22
K0003.2000-01-17.beck   0       0       1.50391009642e-13       5.85838170676e-09
K0003.2001-02-08.kitchen        0       0       5.86355735091e-105      2.54616606545e-84
K0003.2003-12-18.GP     0       1       9.66135811724e-118      3.92756598052e-93
K0003.2004-08-01.BG     1       1       1.08581275097e-50       1.1105506715e-52
K0004.1999-12-10.kaminski       1       0       9.34756695118e-25       4.85191974999e-25
K0004.1999-12-14.farmer 0       0       1.98992573348e-59       1.33925175785e-54
K0004.2001-04-02.williams       0       0       5.79345493451e-127      3.7798492706e-85
K0004.2001-06-12.SA_and_HP      0       1       2.29157740568e-49       2.17328166418e-43
K0004.2004-08-01.BG     1       1       4.93014780996e-59       2.58479841487e-59
K0005.1999-12-12.kaminski       1       0       1.87757476773e-21       4.32445527145e-22
K0005.1999-12-14.farmer 0       0       5.55418658747e-32       1.76466539095e-28
K0005.2000-06-06.lokay  0       0       8.20161484776e-133      3.20504115893e-89
K0005.2001-02-08.kitchen        0       0       8.62696366384e-21       3.42257376455e-19
K0005.2001-06-23.SA_and_HP      0       1       1.09986789828e-39       1.17719023309e-34
K0005.2003-12-18.GP     0       1       0.290023201856  0.709976798144
K0006.1999-12-13.kaminski       1       0       0.0     0.0
K0006.2001-02-08.kitchen        0       0       1.51956459596e-17       4.8800536125e-17
K0006.2001-04-03.williams       0       0       0.0     0.0
K0006.2001-06-25.SA_and_HP      0       1       7.62141840117e-20       5.57950743566e-18
K0006.2003-12-18.GP     1       1       1.66344245187e-07       1.15350922827e-07
K0006.2004-08-01.BG     1       1       1.30953602992e-55       3.30746742545e-62
K0007.1999-12-13.kaminski       1       0       1.03274517954e-28       6.98733361373e-29
K0007.1999-12-14.farmer 0       0       1.31678653891e-152      2.38153554831e-132
K0007.2000-01-17.beck   0       0       9.05511288851e-61       6.2512407045e-49
K0007.2001-02-09.kitchen        0       0       8.20819453728e-236      1.74974965565e-211
K0007.2003-12-18.GP     0       1       5.61541469285e-95       5.96349893343e-89
K0007.2004-08-01.BG     1       1       4.50359966062e-43       2.39014579259e-47
K0008.2001-02-09.kitchen        1       0       9.37273021967e-105      1.81568119626e-143
K0008.2001-06-12.SA_and_HP      0       1       6.5560141939e-240       7.05239513799e-210
K0008.2001-06-25.SA_and_HP      1       1       4.93014780996e-59       2.58479841487e-59
K0008.2003-12-18.GP     1       1       6.18253014825e-300      1.41470630163e-310
K0008.2004-08-01.BG     1       1       8.90349639841e-55       1.93268732223e-61
K0009.1999-12-13.kaminski       1       0       0.0     0.0
K0009.1999-12-14.farmer 0       0       0.0     0.0
K0009.2000-06-07.lokay  0       0       5.91113600182e-35       1.96952695333e-31
K0009.2001-02-09.kitchen        0       0       1.89625846876e-196      5.77232266928e-180
K0009.2001-06-26.SA_and_HP      0       1       0.0     0.0
K0009.2003-12-18.GP     1       1       2.69638979585e-72       4.41243411521e-81
K0010.1999-12-14.farmer 0       0       3.88109146096e-09       1.73962126536e-08
K0010.1999-12-14.kaminski       0       0       4.19318911314e-77       1.12823303144e-50
K0010.2001-02-09.kitchen        0       0       0.0116525173133 0.0398940235488
K0010.2001-06-28.SA_and_HP      0       1       7.39103699935e-241      7.78404910356e-224
K0010.2003-12-18.GP     1       1       1.59717300909e-231      2.21194585112e-242
K0010.2004-08-01.BG     0       1       0.290023201856  0.709976798144
K0011.1999-12-14.farmer 1       0       3.4992546068e-132       2.5833060047e-157
K0011.2001-06-28.SA_and_HP      0       1       7.91133209165e-203      7.44322562205e-167
K0011.2001-06-29.SA_and_HP      1       1       1.25612259364e-230      2.34160849114e-241
K0011.2003-12-18.GP     1       1       0.0     0.0
K0011.2004-08-01.BG     1       1       4.49955125496e-23       1.49065386858e-31
K0012.1999-12-14.farmer 0       0       3.97482136338e-18       9.31034030991e-17
K0012.1999-12-14.kaminski       0       0       2.2572592113e-293       2.43806136026e-243
K0012.2000-01-17.beck   0       0       3.02028886264e-103      3.24736292566e-80
K0012.2000-06-08.lokay  0       0       1.45041909286e-238      2.88447232327e-213
K0012.2001-02-09.kitchen        0       0       1.32315454198e-62       1.17958082359e-57
K0012.2003-12-19.GP     0       1       2.51484584043e-19       3.02047497602e-16
K0013.1999-12-14.farmer 0       0       0.290023201856  0.709976798144
K0013.1999-12-14.kaminski       0       0       3.20845611634e-157      1.51813660089e-140
K0013.2001-04-03.williams       0       0       1.03127705473e-135      1.02143889506e-113
K0013.2001-06-30.SA_and_HP      0       1       1.12903357039e-30       1.12746488254e-28
K0013.2004-08-01.BG     1       1       0.0     0.0
K0014.1999-12-14.kaminski       0       0       5.71792277875e-69       7.63268885023e-68
K0014.1999-12-15.farmer 0       0       7.13614790717e-197      2.80929909001e-168
K0014.2001-02-12.kitchen        0       0       9.94199222086e-93       6.24123885196e-67
K0014.2001-07-04.SA_and_HP      0       1       4.778960058e-85 3.34894593122e-81
K0014.2003-12-19.GP     1       1       2.42870996018e-294      7.72787978022e-307
K0014.2004-08-01.BG     0       1       0.290023201856  0.709976798144
K0015.1999-12-14.kaminski       0       0       6.277139505e-16 5.49737775678e-15
K0015.1999-12-15.farmer 0       0       4.84813892678e-36       3.25784769724e-31
```

```
K0015.2000-06-09.lokay  0       0       3.82400987565e-69       1.06275065401e-46
K0015.2001-02-12.kitchen        0       0       0.290023201856  0.709976798144
K0015.2001-07-05.SA_and_HP      0       1       0.0     0.0
K0015.2003-12-19.GP     1       1       3.06800399429e-40       1.91628326681e-41
K0016.1999-12-15.farmer 1       0       5.92157235782e-76       4.20237202077e-79
K0016.2001-02-12.kitchen        0       0       2.22284487231e-48       3.63411474264e-42
K0016.2001-07-05.SA_and_HP      0       1       6.5070802384e-111       3.36106267078e-96
K0016.2001-07-06.SA_and_HP      1       1       3.06800399429e-40       1.91628326681e-41
K0016.2003-12-19.GP     1       1       0.0     0.0
K0016.2004-08-01.BG     0       1       5.83276321413e-08       9.99991803064e-08
K0017.1999-12-14.kaminski       1       0       3.63997622748e-40       1.43104524436e-40
K0017.2000-01-17.beck   0       0       8.14993923704e-16       1.74692956218e-14
K0017.2001-04-03.williams       0       0       1.45001296929e-238      2.88406228121e-213
K0017.2003-12-18.GP     0       1       7.18436355251e-24       3.74707461475e-21
K0017.2004-08-01.BG     0       1       9.77772654916e-08       1.22340634761e-07
K0017.2004-08-02.BG     0       1       0.00722377470566        0.0329158135929
K0018.1999-12-14.kaminski       1       0       4.55943900735e-158      1.14998835893e-167
K0018.2001-07-13.SA_and_HP      0       1       1.18448212114e-67       4.64268024088e-65
K0018.2003-12-18.GP     1       1       6.33415598662e-273      9.61158973704e-282
K0018.2003-12-18.GP     1       1       9.91213665527e-231      3.3436789456e-238
Error:  34 %
```

The error rate does not substantially change when ignoring words with low counts.

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

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

In [42]:
```
from sklearn.naive_bayes import MultinomialNB
from sklearn.feature_extraction.text import CountVectorizer
import pandas as pd

# Read data
train_data = pd.read_csv('train_data.txt',sep='\t',header=None)
train_data.columns = ['id','spam','header_message']

test_data = pd.read_csv('enronemail_1h.txt',sep='\t',header=None)
test_data.columns = ['id','spam','header','message']
test_data['header_message'] = test_data['header'].fillna('') +' '+ test_data['message'].fillna('')

# Build sparse vector matrix
cv = CountVectorizer()
X = cv.fit_transform(train_data['header_message'])
Y = np.array(train_data['spam'].ravel())

# Fit NB model
nb = MultinomialNB()
nb.fit(X, Y)

# Test accuracy
X = cv.transform(test_data['header_message'])
Y = np.array(test_data['spam'].ravel())
p = nb.predict(X)
accuracy_rate = np.mean(p == Y)
print('Accuracy: {}%'.format(accuracy_rate*100))
```

```
Accuracy: 98.0%
```

```
/Library/Python/2.7/site-packages/numpy/core/fromnumeric.py:2641: VisibleDeprecationWarning: `rank` is deprecated; use t
he `ndim` attribute or function instead. To find the rank of a matrix see `numpy.linalg.matrix_rank`.
  VisibleDeprecationWarning)
```

Approach	Error Rate
Map-Reduce	34%
Sklearn	2%

I would recommend the Sklearn approach, which beats the error rate from the map-reduce classifier and is significantly easier to implement; however, it is not scalable.