In [ ]:

```
# DATASCI W261: Machine Learning at Scale
```

Name: Prabhakar Gundugola
Email: prabhakar@berkeley.edu
Time of Initial Submission: Jan 26, 2016
W261-3: Spring 2016
Week 2: Homework 2
Date: January 26, 2016

HW2.0

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

A race condition is a special condition that can occur when two or more threads can access shared data and they try to change it at the same time.

Race condition screenshot

As shown in the above diagram, consider process A and process B access the same variable filename in slot 7. Initially process B updates the filename variable with "ProgB.c" in slot 7. If process A writes its filename to slot 7 which erases process B's filename, then process B will thrown an error.

What is MapReduce?

MapReduce is an **embarrassingly parallel** framework for processing and generating large datasets with a parallel, distributed algorithm on a cluster.

A MapReduce program is composed of 2 procedures:

- Map: Procedure that processes to generate a set of intermediate key/value pairs
- Reduce: Procedure that performs a summary operation on all intermediate values associated with the same intermediate key.

MapReduce programs written in the above functional style are automatically parallelized and executed on a large number of commodity machines.

How does it differ from Hadoop?

Hadoop is a software platform that implements the MapReduce programming paradigm. It also provides the run-time system that takes care of the details of partitioning of input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication.

Which programming paradigm is Hadoop based on?

Hadoop is based on MapReduce programming model, which is embarrassingly parallel. A problem is considered to be embarrassingly parallel if there is little or no effort required to split the problem into a number of parallel tasks.

Explain and give a simple example in code and show the code running?

- Hadoop MapReduce divides single large data processing job into a number of parallel small map tasks. Once the tasks have been created, they are spread across multiple nodes and are run in parallel. Each of these map tasks generate a set of key/value pairs.
- It then implements the barrier and sorts all the key/value pairs by key.
- The framework then distributes the keys to reduce jobs based on a hash function.
- The reduce jobs processes the input data from barrier and generate the output, which is the output of the MapReduce program.

A typical example of MapReduce algorithm is find the word count in a given list of files and then print the top 10 frequent words.

MapReduce Example - Map

- This map procedure reads files from local disk line by line.
- It then breaks each line into words and prints the key, value pair (word, 1) for each word in a different file

In [29]:

```
%%writefile mapper_example.py
#!/usr/bin/python
import os
import re
WORD_RE = re.compile(r"[\w]+")

files = [f for f in os.listdir("/root/hw2") if "wordcount" in f]

for f in files:
    with open(f, "r") as fp:
        for line in fp.readlines():
            for word in WORD_RE.findall(line):
                print '{0}\t{1}'.format(word.lower(), str(1))
```

```
Overwriting mapper_example.py
```

MapReduce Example - Map

- This reduce procedure takes input from mapper in the form of key, value pair .
- It then accumulates the count for each key.

In [16]:

```
%%writefile reducer_example.py
#!/usr/bin/python
import sys

def wcount(prev_word, count):
    if prev_word is not None:
        print prev_word.ljust(20) + "\t" + str(count)

prev_word = None
count = 0
for line in sys.stdin:
    word, value = line.split('\t', 1)
    if word != prev_word:
        wcount(prev_word, count)
        prev_word = word
        count = 0
    count += eval(value)
wcount(prev_word, count)
```

```
Overwriting reducer_example.py
```

MapReduce Example - Printing the top 10 frequent words

```
In [30]: !python mapper_example.py | sort -k1,1 | python reducer_example.py| sort -r
         k2,2 | head -10
```

```
a                        8
state                    5
output                   4
for                      4
of                       4
system                   3
inputs                   3
may                      3
to                       3
hi                       3
```

HW2.1. Sort in Hadoop MapReduce

Given as input: Records of the form , where integer is any integer, and “NA” is just the empty string.
Output: sorted key value pairs of the form in decreasing order.

What happens if you have multiple reducers? Do you need additional steps? Explain.

In Hadoop MapReduce, MapReduce programming model relies on the sorting feature of Hadoop framework between mappers and reducers.

If there are multiple reducers, each reducer will generate outputs/keys that are sorted within each reducer, but not sorted across all reducers.

Additional steps are required to sort the consolidated key value pairs. They are:

- Sort the multiple reducers output of key, value pairs and then pass this as input to a single reducer.

Another way to achieve this is having a combiner in between the Map and Reduce procedures. A combiner, also known as a semi-reducer, accepts the inputs from the Map procedure and thereafter passes the output of key,value pairs to the Reduce procedure

Combiner - Multiple reducers

Write code to generate N random records of the form . 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 [46]:

```
%%writefile randomgenerator.py
#!/usr/bin/python
import sys
import random

for i in range(10000):
    sys.stdout.write("{0},{1}\n".format(random.randint(1, 10000), "NA"))
```

```
Overwriting randomgenerator.py
```

In [50]:

```
# Change permissions on randomgenerator.py
!chmod a+x randomgenerator.py

# Call randomgenerator.py to generate and output the random numbers to a file
!./randomgenerator.py > randomrecords.txt

# Copy it to HDFS
!hdfs dfs -mkdir -p /user/root/wk2/hw21/input
!hdfs dfs -put randomrecords.txt /user/root/wk2/hw21/input
```

```
16/01/25 23:58:01 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/25 23:58:03 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
```

Mapper for Hadoop streaming MapReduce job to sort the numbers

In [40]:

```
%%writefile mapreducer.py
#!/usr/bin/python
import sys
for line in sys.stdin:
    key, value = line.strip().split(',')
    print '{0},{1}'.format(key, value)
```

```
Overwriting mapreducer.py
```

In [33]:

```
# Change permissions on mapreducer.py
!chmod a+x mapreducer.py
```

Running the MapReduce in Hadoop

In [53]:

```
# Check whether the output folder already exists or not
!hdfs dfs -rm -r /user/root/wk2/hw21/output

# Run Hadoop job
!hadoop jar hadoop-streaming-2.7.1.jar \
-D mapreduce.job.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
-D mapreduce.partition.keycomparator.options=-nr \
-D mapred.combine.tasks=2 \
-mapper /root/hw2/mapreducer.py \
-combiner /root/hw2/mapreducer.py \
-reducer /root/hw2/mapreducer.py \
-input /user/root/wk2/hw21/input/randomrecords.txt \
-output /user/root/wk2/hw21/output
```

```
16/01/26 00:05:24 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 00:05:25 INFO fs.TrashPolicyDefault: Namenode trash configuratio
n: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/root/wk2/hw21/output
16/01/26 00:05:27 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 00:05:28 INFO Configuration.deprecation: session.id is deprecate
d. Instead, use dfs.metrics.session-id
16/01/26 00:05:28 INFO jvm.JvmMetrics: Initializing JVM Metrics with proce
ssName=JobTracker, sessionId=
16/01/26 00:05:28 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with
processName=JobTracker, sessionId= - already initialized
16/01/26 00:05:28 INFO mapred.FileInputFormat: Total input paths to proces
s : 1
16/01/26 00:05:28 INFO mapreduce.JobSubmitter: number of splits:1
16/01/26 00:05:28 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_local1386910411_0001
16/01/26 00:05:29 INFO mapreduce.Job: The url to track the job: http://loc
alhost:8080/
16/01/26 00:05:29 INFO mapreduce.Job: Running job: job_local1386910411_000
1
16/01/26 00:05:29 INFO mapred.LocalJobRunner: OutputCommitter set in confi
g null
16/01/26 00:05:29 INFO mapred.LocalJobRunner: OutputCommitter is org.apach
e.hadoop.mapred.FileOutputCommitter
16/01/26 00:05:29 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 00:05:29 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/26 00:05:29 INFO mapred.LocalJobRunner: Starting task: attempt_local
1386910411_0001_m_000000_0
16/01/26 00:05:29 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 00:05:29 INFO mapred.Task:  Using ResourceCalculatorProcessTree :
[ ]
16/01/26 00:05:29 INFO mapred.MapTask: Processing split: hdfs://localhos
t:54310/user/root/wk2/hw21/input/randomrecords.txt:0+78881
16/01/26 00:05:29 INFO mapred.MapTask: numReduceTasks: 1
16/01/26 00:05:29 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/26 00:05:29 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/26 00:05:29 INFO mapred.MapTask: soft limit at 83886080
16/01/26 00:05:29 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/26 00:05:29 INFO mapred.MapTask: kvstart = 26214396; length = 655360
0
16/01/26 00:05:29 INFO mapred.MapTask: Map output collector class = org.ap
ache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/26 00:05:29 INFO streaming.PipeMapRed: PipeMapRed exec [/root/hw2/ma
preducer.py]
16/01/26 00:05:29 INFO Configuration.deprecation: mapred.task.id is deprec
ated. Instead, use mapreduce.task.attempt.id
16/01/26 00:05:29 INFO Configuration.deprecation: mapred.task.is.map is de
precated. Instead, use mapreduce.task.ismap
16/01/26 00:05:29 INFO Configuration.deprecation: mapred.skip.on is deprec
ated. Instead, use mapreduce.job.skiprecords
16/01/26 00:05:29 INFO Configuration.deprecation: mapred.local.dir is depr
```

```
ecated. Instead, use mapreduce.cluster.local.dir
16/01/26 00:05:29 INFO Configuration.deprecation: map.input.file is deprec
ated. Instead, use mapreduce.map.input.file
16/01/26 00:05:29 INFO Configuration.deprecation: mapred.job.id is depreca
ted. Instead, use mapreduce.job.id
16/01/26 00:05:29 INFO Configuration.deprecation: user.name is deprecated.
Instead, use mapreduce.job.user.name
16/01/26 00:05:29 INFO Configuration.deprecation: map.input.start is depre
cated. Instead, use mapreduce.map.input.start
16/01/26 00:05:29 INFO Configuration.deprecation: mapred.tip.id is depreca
ted. Instead, use mapreduce.task.id
16/01/26 00:05:29 INFO Configuration.deprecation: mapred.task.partition is
deprecated. Instead, use mapreduce.task.partition
16/01/26 00:05:29 INFO Configuration.deprecation: map.input.length is depr
ecated. Instead, use mapreduce.map.input.length
16/01/26 00:05:29 INFO Configuration.deprecation: mapred.work.output.dir i
s deprecated. Instead, use mapreduce.task.output.dir
16/01/26 00:05:29 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 00:05:29 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 00:05:29 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] o
ut:NA [rec/s]
16/01/26 00:05:29 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s]
out:NA [rec/s]
16/01/26 00:05:29 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s]
out:NA [rec/s]
16/01/26 00:05:29 INFO streaming.PipeMapRed: Records R/W=10000/1
16/01/26 00:05:29 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 00:05:29 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 00:05:29 INFO mapred.LocalJobRunner:
16/01/26 00:05:29 INFO mapred.MapTask: Starting flush of map output
16/01/26 00:05:29 INFO mapred.MapTask: Spilling map output
16/01/26 00:05:29 INFO mapred.MapTask: bufstart = 0; bufend = 88881; bufvo
id = 104857600
16/01/26 00:05:29 INFO mapred.MapTask: kvstart = 26214396(104857584); kven
d = 26174400(104697600); length = 39997/6553600
16/01/26 00:05:29 INFO streaming.PipeMapRed: PipeMapRed exec [/root/hw2/ma
preducer.py]
16/01/26 00:05:29 INFO Configuration.deprecation: mapred.skip.map.auto.inc
r.proc.count is deprecated. Instead, use mapreduce.map.skip.proc-count.aut
o-incr
16/01/26 00:05:29 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 00:05:29 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 00:05:29 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] o
ut:NA [rec/s]
16/01/26 00:05:29 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s]
out:NA [rec/s]
16/01/26 00:05:30 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s]
out:NA [rec/s]
16/01/26 00:05:30 INFO streaming.PipeMapRed: Records R/W=10000/1
16/01/26 00:05:30 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 00:05:30 INFO mapreduce.Job: Job job_local1386910411_0001 running
in uber mode : false
```

```
16/01/26 00:05:30 INFO mapreduce.Job:  map 0% reduce 0%
16/01/26 00:05:30 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 00:05:30 INFO mapred.MapTask: Finished spill 0
16/01/26 00:05:30 INFO mapred.Task: Task:attempt_local1386910411_0001_m_00
0000_0 is done. And is in the process of committing
16/01/26 00:05:30 INFO mapred.LocalJobRunner: Records R/W=10000/1
16/01/26 00:05:30 INFO mapred.Task: Task 'attempt_local1386910411_0001_m_0
00000_0' done.
16/01/26 00:05:30 INFO mapred.LocalJobRunner: Finishing task: attempt_loca
l1386910411_0001_m_000000_0
16/01/26 00:05:30 INFO mapred.LocalJobRunner: map task executor complete.
16/01/26 00:05:30 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/26 00:05:30 INFO mapred.LocalJobRunner: Starting task: attempt_local
1386910411_0001_r_000000_0
16/01/26 00:05:30 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 00:05:30 INFO mapred.Task:  Using ResourceCalculatorProcessTree :
[ ]
16/01/26 00:05:30 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: or
g.apache.hadoop.mapreduce.task.reduce.Shuffle@23a83610
16/01/26 00:05:30 INFO reduce.MergeManagerImpl: MergerManager: memoryLimi
t=353422528, maxSingleShuffleLimit=88355632, mergeThreshold=233258880, ioS
ortFactor=10, memToMemMergeOutputsThreshold=10
16/01/26 00:05:30 INFO reduce.EventFetcher: attempt_local1386910411_000
1_r_000000_0 Thread started: EventFetcher for fetching Map Completion Even
ts
16/01/26 00:05:30 INFO reduce.LocalFetcher: localfetcher#1 about to shuffl
e output of map attempt_local1386910411_0001_m_000000_0 decomp: 108883 le
n: 108887 to MEMORY
16/01/26 00:05:30 INFO reduce.InMemoryMapOutput: Read 108883 bytes from ma
p-output for attempt_local1386910411_0001_m_000000_0
16/01/26 00:05:30 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-o
utput of size: 108883, inMemoryMapOutputs.size() -> 1, commitMemory -> 0,
usedMemory ->108883
16/01/26 00:05:30 INFO reduce.EventFetcher: EventFetcher is interrupted..
Returning
16/01/26 00:05:30 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 00:05:30 INFO reduce.MergeManagerImpl: finalMerge called with 1 i
n-memory map-outputs and 0 on-disk map-outputs
16/01/26 00:05:30 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 00:05:30 INFO mapred.Merger: Down to the last merge-pass, with 1
segments left of total size: 108873 bytes
16/01/26 00:05:30 INFO reduce.MergeManagerImpl: Merged 1 segments, 108883
bytes to disk to satisfy reduce memory limit
16/01/26 00:05:30 INFO reduce.MergeManagerImpl: Merging 1 files, 108887 by
tes from disk
16/01/26 00:05:30 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 byte
s from memory into reduce
16/01/26 00:05:30 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 00:05:30 INFO mapred.Merger: Down to the last merge-pass, with 1
segments left of total size: 108873 bytes
16/01/26 00:05:30 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 00:05:30 INFO streaming.PipeMapRed: PipeMapRed exec [/root/hw2/ma
preducer.py]
16/01/26 00:05:30 INFO Configuration.deprecation: mapred.job.tracker is de
precated. Instead, use mapreduce.jobtracker.address
```

```
16/01/26 00:05:30 INFO Configuration.deprecation: mapred.map.tasks is depr
ecated. Instead, use mapreduce.job.maps
16/01/26 00:05:30 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 00:05:30 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 00:05:30 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] o
ut:NA [rec/s]
16/01/26 00:05:30 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s]
out:NA [rec/s]
16/01/26 00:05:30 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s]
out:NA [rec/s]
16/01/26 00:05:30 INFO streaming.PipeMapRed: Records R/W=10000/1
16/01/26 00:05:30 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 00:05:30 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 00:05:30 INFO mapred.Task: Task:attempt_local1386910411_0001_r_00
0000_0 is done. And is in the process of committing
16/01/26 00:05:30 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 00:05:30 INFO mapred.Task: Task attempt_local1386910411_0001_r_00
0000_0 is allowed to commit now
16/01/26 00:05:30 INFO output.FileOutputCommitter: Saved output of task 'a
ttempt_local1386910411_0001_r_000000_0' to hdfs://localhost:54310/user/roo
t/wk2/hw21/output/_temporary/0/task_local1386910411_0001_r_000000
16/01/26 00:05:30 INFO mapred.LocalJobRunner: Records R/W=10000/1 > reduce
16/01/26 00:05:30 INFO mapred.Task: Task 'attempt_local1386910411_0001_r_0
00000_0' done.
16/01/26 00:05:30 INFO mapred.LocalJobRunner: Finishing task: attempt_loca
l1386910411_0001_r_000000_0
16/01/26 00:05:30 INFO mapred.LocalJobRunner: reduce task executor complet
e.
16/01/26 00:05:31 INFO mapreduce.Job:  map 100% reduce 100%
16/01/26 00:05:31 INFO mapreduce.Job: Job job_local1386910411_0001 complet
ed successfully
16/01/26 00:05:31 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=429654
                FILE: Number of bytes written=1099275
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=157762
                HDFS: Number of bytes written=88881
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=10000
                Map output records=10000
                Map output bytes=88881
                Map output materialized bytes=108887
                Input split bytes=117
                Combine input records=10000
                Combine output records=10000
                Reduce input groups=6295
                Reduce shuffle bytes=108887
                Reduce input records=10000
                Reduce output records=10000
                Spilled Records=20000
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=32
                Total committed heap usage (bytes)=1013448704
        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=78881
        File Output Format Counters
                Bytes Written=88881
16/01/26 00:05:31 INFO streaming.StreamJob: Output directory: /user/root/w
k2/hw21/output
```

In [60]:

```
!hdfs dfs -copyToLocal /user/root/wk2/hw21/output/part*
```

```
16/01/26 00:14:21 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 00:14:22 WARN hdfs.DFSClient: DFSInputStream has been closed alre
ady
```

Display the 10 largest numbers

```
In [68]: !echo "ls part-*"
         !echo "---------"
         !ls part*
         !echo " "
         !echo "Printing 10 largest numbers"
         !echo "---------------------------"
         !head -10 part-00000|cut -d "," -f 1
         !echo ""
         !echo ""
         !echo "Printing 10 smallest numbers"
         !echo "----------------------------"
         !tail -10 part-00000|cut -d "," -f 1
```

```
ls part-*
---------
part-00000

Printing 10 largest numbers
---------------------------
9999
9999
9998
9996
9996
9995
9994
9993
9990
9990


Printing 10 smallest numbers
----------------------------
11
10
9
9
8
7
7
5
4
2
```

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.

Data validation and cleansing By exploring the data in the input dataset, 2 problems with 3 records are identified:

- There are 2 records with only 3 fields instead of 4 fields.
- There is 1 record with extra new line character in the body field

Data cleansing algorithm

- Open enronemail_1new.txt with "w" permissions
- Initialize prev_line = ""
- For each line as line in enronemail_1h.txt file
    - Tokenize with delimiter "\t"
    - If number of tokens >= 3 then
    - If number of tokens == 3 then
        - Add "\t" as another token between 2nd and 3rd tokens. Now total number of tokens = 4.
        - Update line by concatenating all the 4 okens
        - If prev_line != "" then write prev_line in enronemail_1new.txt
    - prev_line = line
    - If number of tokens == 1 then
        - prev_line = prev_line + line

In [69]:

```
# Data cleansing algorithm

import os
import re

# Open enronemail_1new.txt with "w" permissions.
with open("enronemail_1new.txt", "w") as new:
    with open("enronemail_1h.txt", "rU") as old:
        # curr_line is the line to be written to new file. Initially it is
set to "".
        prev_line = ""

        # For every line in enronemail_1h.txt file
        for line in old:

            line = line.strip()
            # Split the line into tokens
            tokens = line.split('\t')

            if len(tokens) >= 3:

                # If subject field is missed out, add blank token and recon
struct the line
                if len(tokens) == 3:
                    line = tokens[0] + '\t' + tokens[1] + '\t' + '' + '\t'
+ tokens[2]

                # If len(tokens) == 4 then this line is valid. Keep it in b
uffer.
                # Now copy the previous line (if not blank).
                if prev_line != "":
                    prev_line += '\n'
                    new.write(prev_line)
                prev_line = line

            # If there is only one field, it must be because of an
            # extra new line character in the previous line body field.
            if len(tokens) == 1:
                # Add this line too to the previous line
                prev_line += line

        # Add the last line to the new file
        new.write(prev_line)

# Now rename enronemail_1new.txt to enronemail_1h.txt
os.rename('enronemail_1new.txt', 'enronemail_1h.txt')

print "Cleanup completed"
```

```
Cleanup completed
```

In [98]:

```
%%writefile mapper.py
#!/usr/bin/python
## mapper.py
## Author: Prabhakar Gundugola
## Description: mapper code for HW2.2

import sys
import re
import string

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

## collect user input
for line in sys.stdin:
    tokens = line.lower().split('\t')

    # Concatenate subject and body fields and store it in word_string
    word_string = tokens[2] + ' ' + tokens[3].strip()

    # Remove punctuation
    word_string = word_string.translate(string.maketrans("",""),
                                        string.punctuation)

    for word in WORD_RE.findall(word_string):
        print('{0}\t{1}'.format(word.lower(), 1))
```

```
Overwriting mapper.py
```

In [99]:

```
%%writefile reducer.py
#!/usr/bin/python
## reducer.py
## Author: Prabhakar Gundugola
## Description: reducer code for HW2.2

import sys

def wcount(prev_word, counts):
    if prev_word is not None:
        print prev_word + "\t" + str(counts)

prev_word = None
counts = 0
for line in sys.stdin:
    word, value = line.split('\t', 1)
    if word != prev_word:
        wcount(prev_word, counts)
        prev_word = word
        counts = 0

    counts += eval(value)
wcount(prev_word, counts)
```

```
Overwriting reducer.py
```

In [92]:

```
# Change permissions on mapper.py and reducer.py
!chmod a+x mapper.py
!chmod a+x reducer.py
```

Testing the code with command line

In [80]:

```
!cat enronemail_1h.txt| ./mapper.py "assistance" | ./reducer.py
```

```
assistance      10
```

Running the MapReduce in Hadoop

In [93]:

```
# Ensure hw22 folder doesn't exist
!hdfs dfs -rm -r /user/root/wk2/hw22

# Create HDFS input folder
!hdfs dfs -mkdir -p /user/root/wk2/hw22/input

# Copy Enron email file to HDFS input folder
!hdfs dfs -put enronemail_1h.txt /user/root/wk2/hw22/input
```

```
16/01/26 01:11:00 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 01:11:01 INFO fs.TrashPolicyDefault: Namenode trash configuratio
n: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/root/wk2/hw22
16/01/26 01:11:02 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 01:11:05 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
```

Run Hadoop Streaming job

In [103]:

```
#Delete output folder if exists
!hdfs dfs -rm -r /user/root/wk2/hw22/output

# Run Hadoop Streaming job
!hadoop jar hadoop-streaming-2.7.1.jar \
-mapper mapper.py \
-reducer reducer.py \
-input /user/root/wk2/hw22/input \
-output /user/root/wk2/hw22/output
```

```
16/01/26 01:16:15 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 01:16:16 INFO fs.TrashPolicyDefault: Namenode trash configuratio
n: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/root/wk2/hw22/output
16/01/26 01:16:17 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 01:16:18 INFO Configuration.deprecation: session.id is deprecate
d. Instead, use dfs.metrics.session-id
16/01/26 01:16:18 INFO jvm.JvmMetrics: Initializing JVM Metrics with proce
ssName=JobTracker, sessionId=
16/01/26 01:16:18 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with
processName=JobTracker, sessionId= - already initialized
16/01/26 01:16:19 INFO mapred.FileInputFormat: Total input paths to proces
s : 1
16/01/26 01:16:19 INFO mapreduce.JobSubmitter: number of splits:1
16/01/26 01:16:19 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_local126404396_0001
16/01/26 01:16:19 INFO mapreduce.Job: The url to track the job: http://loc
alhost:8080/
16/01/26 01:16:19 INFO mapred.LocalJobRunner: OutputCommitter set in confi
g null
16/01/26 01:16:19 INFO mapreduce.Job: Running job: job_local126404396_0001
16/01/26 01:16:19 INFO mapred.LocalJobRunner: OutputCommitter is org.apach
e.hadoop.mapred.FileOutputCommitter
16/01/26 01:16:19 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 01:16:19 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/26 01:16:19 INFO mapred.LocalJobRunner: Starting task: attempt_local
126404396_0001_m_000000_0
16/01/26 01:16:19 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 01:16:19 INFO mapred.Task:  Using ResourceCalculatorProcessTree :
[ ]
16/01/26 01:16:19 INFO mapred.MapTask: Processing split: hdfs://localhos
t:54310/user/root/wk2/hw22/input/enronemail_1h.txt:0+203954
16/01/26 01:16:19 INFO mapred.MapTask: numReduceTasks: 1
16/01/26 01:16:19 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/26 01:16:19 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/26 01:16:19 INFO mapred.MapTask: soft limit at 83886080
16/01/26 01:16:19 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/26 01:16:19 INFO mapred.MapTask: kvstart = 26214396; length = 655360
0
16/01/26 01:16:19 INFO mapred.MapTask: Map output collector class = org.ap
ache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/26 01:16:19 INFO streaming.PipeMapRed: PipeMapRed exec [/root/hw
2/./mapper.py]
16/01/26 01:16:19 INFO Configuration.deprecation: mapred.task.id is deprec
ated. Instead, use mapreduce.task.attempt.id
16/01/26 01:16:19 INFO Configuration.deprecation: user.name is deprecated.
Instead, use mapreduce.job.user.name
16/01/26 01:16:19 INFO Configuration.deprecation: map.input.start is depre
cated. Instead, use mapreduce.map.input.start
16/01/26 01:16:19 INFO Configuration.deprecation: mapred.task.is.map is de
precated. Instead, use mapreduce.task.ismap
```

```
16/01/26 01:16:19 INFO Configuration.deprecation: mapred.tip.id is depreca
ted. Instead, use mapreduce.task.id
16/01/26 01:16:19 INFO Configuration.deprecation: mapred.skip.on is deprec
ated. Instead, use mapreduce.job.skiprecords
16/01/26 01:16:19 INFO Configuration.deprecation: mapred.task.partition is
deprecated. Instead, use mapreduce.task.partition
16/01/26 01:16:19 INFO Configuration.deprecation: map.input.length is depr
ecated. Instead, use mapreduce.map.input.length
16/01/26 01:16:19 INFO Configuration.deprecation: mapred.local.dir is depr
ecated. Instead, use mapreduce.cluster.local.dir
16/01/26 01:16:19 INFO Configuration.deprecation: mapred.work.output.dir i
s deprecated. Instead, use mapreduce.task.output.dir
16/01/26 01:16:19 INFO Configuration.deprecation: map.input.file is deprec
ated. Instead, use mapreduce.map.input.file
16/01/26 01:16:19 INFO Configuration.deprecation: mapred.job.id is depreca
ted. Instead, use mapreduce.job.id
16/01/26 01:16:19 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 01:16:19 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 01:16:20 INFO streaming.PipeMapRed: Records R/W=72/1
16/01/26 01:16:20 INFO streaming.PipeMapRed: R/W/S=100/1035/0 in:NA [re
c/s] out:NA [rec/s]
16/01/26 01:16:20 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 01:16:20 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 01:16:20 INFO mapred.LocalJobRunner:
16/01/26 01:16:20 INFO mapred.MapTask: Starting flush of map output
16/01/26 01:16:20 INFO mapred.MapTask: Spilling map output
16/01/26 01:16:20 INFO mapred.MapTask: bufstart = 0; bufend = 246970; bufv
oid = 104857600
16/01/26 01:16:20 INFO mapred.MapTask: kvstart = 26214396(104857584); kven
d = 26088536(104354144); length = 125861/6553600
16/01/26 01:16:20 INFO mapred.MapTask: Finished spill 0
16/01/26 01:16:20 INFO mapred.Task: Task:attempt_local126404396_0001_m_000
000_0 is done. And is in the process of committing
16/01/26 01:16:20 INFO mapred.LocalJobRunner: Records R/W=72/1
16/01/26 01:16:20 INFO mapred.Task: Task 'attempt_local126404396_0001_m_00
0000_0' done.
16/01/26 01:16:20 INFO mapred.LocalJobRunner: Finishing task: attempt_loca
l126404396_0001_m_000000_0
16/01/26 01:16:20 INFO mapred.LocalJobRunner: map task executor complete.
16/01/26 01:16:20 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/26 01:16:20 INFO mapred.LocalJobRunner: Starting task: attempt_local
126404396_0001_r_000000_0
16/01/26 01:16:20 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 01:16:20 INFO mapred.Task:  Using ResourceCalculatorProcessTree :
[ ]
16/01/26 01:16:20 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: or
g.apache.hadoop.mapreduce.task.reduce.Shuffle@74327a4
16/01/26 01:16:20 INFO reduce.MergeManagerImpl: MergerManager: memoryLimi
t=353422528, maxSingleShuffleLimit=88355632, mergeThreshold=233258880, ioS
ortFactor=10, memToMemMergeOutputsThreshold=10
16/01/26 01:16:20 INFO reduce.EventFetcher: attempt_local126404396_000
1_r_000000_0 Thread started: EventFetcher for fetching Map Completion Even
ts
```

```
16/01/26 01:16:20 INFO reduce.LocalFetcher: localfetcher#1 about to shuffl
e output of map attempt_local126404396_0001_m_000000_0 decomp: 309904 len:
309908 to MEMORY
16/01/26 01:16:20 INFO mapreduce.Job: Job job_local126404396_0001 running
in uber mode : false
16/01/26 01:16:20 INFO reduce.InMemoryMapOutput: Read 309904 bytes from ma
p-output for attempt_local126404396_0001_m_000000_0
16/01/26 01:16:20 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-o
utput of size: 309904, inMemoryMapOutputs.size() -> 1, commitMemory -> 0,
usedMemory ->309904
16/01/26 01:16:20 INFO mapreduce.Job:  map 100% reduce 0%
16/01/26 01:16:20 INFO reduce.EventFetcher: EventFetcher is interrupted..
Returning
16/01/26 01:16:20 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 01:16:20 INFO reduce.MergeManagerImpl: finalMerge called with 1 i
n-memory map-outputs and 0 on-disk map-outputs
16/01/26 01:16:20 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 01:16:20 INFO mapred.Merger: Down to the last merge-pass, with 1
segments left of total size: 309900 bytes
16/01/26 01:16:20 INFO reduce.MergeManagerImpl: Merged 1 segments, 309904
bytes to disk to satisfy reduce memory limit
16/01/26 01:16:20 INFO reduce.MergeManagerImpl: Merging 1 files, 309908 by
tes from disk
16/01/26 01:16:20 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 byte
s from memory into reduce
16/01/26 01:16:20 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 01:16:20 INFO mapred.Merger: Down to the last merge-pass, with 1
segments left of total size: 309900 bytes
16/01/26 01:16:20 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 01:16:20 INFO streaming.PipeMapRed: PipeMapRed exec [/root/hw
2/./reducer.py]
16/01/26 01:16:20 INFO Configuration.deprecation: mapred.job.tracker is de
precated. Instead, use mapreduce.jobtracker.address
16/01/26 01:16:20 INFO Configuration.deprecation: mapred.map.tasks is depr
ecated. Instead, use mapreduce.job.maps
16/01/26 01:16:20 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 01:16:20 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 01:16:20 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] o
ut:NA [rec/s]
16/01/26 01:16:20 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s]
out:NA [rec/s]
16/01/26 01:16:20 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s]
out:NA [rec/s]
16/01/26 01:16:21 INFO streaming.PipeMapRed: Records R/W=16560/1
16/01/26 01:16:21 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 01:16:21 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 01:16:21 INFO mapred.Task: Task:attempt_local126404396_0001_r_000
000_0 is done. And is in the process of committing
16/01/26 01:16:21 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 01:16:21 INFO mapred.Task: Task attempt_local126404396_0001_r_000
000_0 is allowed to commit now
16/01/26 01:16:21 INFO output.FileOutputCommitter: Saved output of task 'a
ttempt_local126404396_0001_r_000000_0' to hdfs://localhost:54310/user/roo
t/wk2/hw22/output/_temporary/0/task_local126404396_0001_r_000000
```

```
16/01/26 01:16:21 INFO mapred.LocalJobRunner: Records R/W=16560/1 > reduce
16/01/26 01:16:21 INFO mapred.Task: Task 'attempt_local126404396_0001_r_00
0000_0' done.
16/01/26 01:16:21 INFO mapred.LocalJobRunner: Finishing task: attempt_loca
l126404396_0001_r_000000_0
16/01/26 01:16:21 INFO mapred.LocalJobRunner: reduce task executor complet
e.
16/01/26 01:16:21 INFO mapreduce.Job:  map 100% reduce 100%
16/01/26 01:16:21 INFO mapreduce.Job: Job job_local126404396_0001 complete
d successfully
16/01/26 01:16:21 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=831696
                FILE: Number of bytes written=1695196
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=407908
                HDFS: Number of bytes written=57078
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=100
                Map output records=31466
                Map output bytes=246970
                Map output materialized bytes=309908
                Input split bytes=117
                Combine input records=0
                Combine output records=0
                Reduce input groups=5740
                Reduce shuffle bytes=309908
                Reduce input records=31466
                Reduce output records=5740
                Spilled Records=62932
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=25
                Total committed heap usage (bytes)=1013448704
        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=203954
        File Output Format Counters
                Bytes Written=57078
16/01/26 01:16:21 INFO streaming.StreamJob: Output directory: /user/root/w
k2/hw22/output
```

In [104]:

```
!rm part*
!hdfs dfs -copyToLocal /user/root/wk2/hw22/output/part*
```

```
16/01/26 01:16:30 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 01:16:31 WARN hdfs.DFSClient: DFSInputStream has been closed alre
ady
```

Examine the word "assistance"

In [105]:

```
!grep assistance part*
```

```
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 [115]:

```
%%writefile mapper.py
#!/usr/bin/python
## mapper.py
## Author: Prabhakar Gundugola
## Description: mapper code for HW2.2.1
import sys

#Swap key and value
for line in sys.stdin:
    vals = line.strip().split('\t')
    print('{0}\t{1}'.format(vals[1], vals[0]))
```

```
Overwriting mapper.py
```

In [ ]:

```
%%writefile reduce.py
#!/usr/bin/python
## mapper.py
## Author: Prabhakar Gundugola
## Description: reducer code for HW2.2.1
import sys

#Swap key and value
for line in sys.stdin:
    vals = line.replace('\n', '').split('\t')
    print('{0}\t{1}'.format(vals[1], vals[0]))
```

In [116]:

```
# Change permissions on mapper
!chmod a+x mapreduce.py

# Delete output folder if exists
!hdfs dfs -rm -r /user/root/wk2/hw22/output_1

# Run Hadoop Streaming job
!hadoop jar hadoop-streaming-2.7.1.jar \
-D mapreduce.job.output.key.comparator.class=org.apache.hadoop.mapreduce.lib.partition.KeyFieldBasedComparator \
-D mapreduce.partition.keycomparator.options=-k1,1n \
-mapper mapreduce.py \
-reducer org.apache.hadoop.mapred.lib.IdentityReducer \
-input /user/root/wk2/hw22/output \
-output /user/root/wk2/hw22/output_1
```

```
16/01/26 01:34:12 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 01:34:13 INFO fs.TrashPolicyDefault: Namenode trash configuratio
n: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/root/wk2/hw22/output_1
16/01/26 01:34:14 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 01:34:15 INFO Configuration.deprecation: session.id is deprecate
d. Instead, use dfs.metrics.session-id
16/01/26 01:34:15 INFO jvm.JvmMetrics: Initializing JVM Metrics with proce
ssName=JobTracker, sessionId=
16/01/26 01:34:15 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with
processName=JobTracker, sessionId= - already initialized
16/01/26 01:34:15 INFO mapred.FileInputFormat: Total input paths to proces
s : 1
16/01/26 01:34:15 INFO mapreduce.JobSubmitter: number of splits:1
16/01/26 01:34:16 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_local2107275415_0001
16/01/26 01:34:16 INFO mapreduce.Job: The url to track the job: http://loc
alhost:8080/
16/01/26 01:34:16 INFO mapred.LocalJobRunner: OutputCommitter set in confi
g null
16/01/26 01:34:16 INFO mapreduce.Job: Running job: job_local2107275415_000
1
16/01/26 01:34:16 INFO mapred.LocalJobRunner: OutputCommitter is org.apach
e.hadoop.mapred.FileOutputCommitter
16/01/26 01:34:16 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 01:34:16 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/26 01:34:16 INFO mapred.LocalJobRunner: Starting task: attempt_local
2107275415_0001_m_000000_0
16/01/26 01:34:16 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 01:34:16 INFO mapred.Task:  Using ResourceCalculatorProcessTree :
[ ]
16/01/26 01:34:16 INFO mapred.MapTask: Processing split: hdfs://localhos
t:54310/user/root/wk2/hw22/output/part-00000:0+57078
16/01/26 01:34:16 INFO mapred.MapTask: numReduceTasks: 1
16/01/26 01:34:16 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/26 01:34:16 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/26 01:34:16 INFO mapred.MapTask: soft limit at 83886080
16/01/26 01:34:16 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/26 01:34:16 INFO mapred.MapTask: kvstart = 26214396; length = 655360
0
16/01/26 01:34:16 INFO mapred.MapTask: Map output collector class = org.ap
ache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/26 01:34:16 INFO streaming.PipeMapRed: PipeMapRed exec [/root/hw
2/./mapreduce.py]
16/01/26 01:34:16 INFO Configuration.deprecation: mapred.task.id is deprec
ated. Instead, use mapreduce.task.attempt.id
16/01/26 01:34:16 INFO Configuration.deprecation: user.name is deprecated.
Instead, use mapreduce.job.user.name
16/01/26 01:34:16 INFO Configuration.deprecation: map.input.start is depre
cated. Instead, use mapreduce.map.input.start
16/01/26 01:34:16 INFO Configuration.deprecation: mapred.task.is.map is de
```

```
precated. Instead, use mapreduce.task.ismap
16/01/26 01:34:16 INFO Configuration.deprecation: mapred.tip.id is depreca
ted. Instead, use mapreduce.task.id
16/01/26 01:34:16 INFO Configuration.deprecation: mapred.skip.on is deprec
ated. Instead, use mapreduce.job.skiprecords
16/01/26 01:34:16 INFO Configuration.deprecation: mapred.task.partition is
deprecated. Instead, use mapreduce.task.partition
16/01/26 01:34:16 INFO Configuration.deprecation: map.input.length is depr
ecated. Instead, use mapreduce.map.input.length
16/01/26 01:34:16 INFO Configuration.deprecation: mapred.local.dir is depr
ecated. Instead, use mapreduce.cluster.local.dir
16/01/26 01:34:16 INFO Configuration.deprecation: mapred.work.output.dir i
s deprecated. Instead, use mapreduce.task.output.dir
16/01/26 01:34:16 INFO Configuration.deprecation: map.input.file is deprec
ated. Instead, use mapreduce.map.input.file
16/01/26 01:34:16 INFO Configuration.deprecation: mapred.job.id is depreca
ted. Instead, use mapreduce.job.id
16/01/26 01:34:16 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 01:34:16 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 01:34:16 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] o
ut:NA [rec/s]
16/01/26 01:34:16 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s]
out:NA [rec/s]
16/01/26 01:34:16 INFO streaming.PipeMapRed: Records R/W=5740/1
16/01/26 01:34:16 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 01:34:17 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 01:34:17 INFO mapred.LocalJobRunner:
16/01/26 01:34:17 INFO mapred.MapTask: Starting flush of map output
16/01/26 01:34:17 INFO mapred.MapTask: Spilling map output
16/01/26 01:34:17 INFO mapred.MapTask: bufstart = 0; bufend = 57078; bufvo
id = 104857600
16/01/26 01:34:17 INFO mapred.MapTask: kvstart = 26214396(104857584); kven
d = 26191440(104765760); length = 22957/6553600
16/01/26 01:34:17 INFO mapred.MapTask: Finished spill 0
16/01/26 01:34:17 INFO mapred.Task: Task:attempt_local2107275415_0001_m_00
0000_0 is done. And is in the process of committing
16/01/26 01:34:17 INFO mapred.LocalJobRunner: Records R/W=5740/1
16/01/26 01:34:17 INFO mapred.Task: Task 'attempt_local2107275415_0001_m_0
00000_0' done.
16/01/26 01:34:17 INFO mapred.LocalJobRunner: Finishing task: attempt_loca
l2107275415_0001_m_000000_0
16/01/26 01:34:17 INFO mapred.LocalJobRunner: map task executor complete.
16/01/26 01:34:17 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/26 01:34:17 INFO mapred.LocalJobRunner: Starting task: attempt_local
2107275415_0001_r_000000_0
16/01/26 01:34:17 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 01:34:17 INFO mapred.Task:  Using ResourceCalculatorProcessTree :
[ ]
16/01/26 01:34:17 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: or
g.apache.hadoop.mapreduce.task.reduce.Shuffle@39741f43
16/01/26 01:34:17 INFO reduce.MergeManagerImpl: MergerManager: memoryLimi
t=353422528, maxSingleShuffleLimit=88355632, mergeThreshold=233258880, ioS
ortFactor=10, memToMemMergeOutputsThreshold=10
```

```
16/01/26 01:34:17 INFO reduce.EventFetcher: attempt_local2107275415_000
1_r_000000_0 Thread started: EventFetcher for fetching Map Completion Even
ts
16/01/26 01:34:17 INFO reduce.LocalFetcher: localfetcher#1 about to shuffl
e output of map attempt_local2107275415_0001_m_000000_0 decomp: 68560 len:
68564 to MEMORY
16/01/26 01:34:17 INFO reduce.InMemoryMapOutput: Read 68560 bytes from ma
p-output for attempt_local2107275415_0001_m_000000_0
16/01/26 01:34:17 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-o
utput of size: 68560, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, u
sedMemory ->68560
16/01/26 01:34:17 INFO reduce.EventFetcher: EventFetcher is interrupted..
Returning
16/01/26 01:34:17 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 01:34:17 INFO reduce.MergeManagerImpl: finalMerge called with 1 i
n-memory map-outputs and 0 on-disk map-outputs
16/01/26 01:34:17 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 01:34:17 INFO mapred.Merger: Down to the last merge-pass, with 1
segments left of total size: 68556 bytes
16/01/26 01:34:17 INFO reduce.MergeManagerImpl: Merged 1 segments, 68560 b
ytes to disk to satisfy reduce memory limit
16/01/26 01:34:17 INFO reduce.MergeManagerImpl: Merging 1 files, 68564 byt
es from disk
16/01/26 01:34:17 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 byte
s from memory into reduce
16/01/26 01:34:17 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 01:34:17 INFO mapred.Merger: Down to the last merge-pass, with 1
segments left of total size: 68556 bytes
16/01/26 01:34:17 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 01:34:17 INFO mapreduce.Job: Job job_local2107275415_0001 running
in uber mode : false
16/01/26 01:34:17 INFO mapreduce.Job:  map 100% reduce 0%
16/01/26 01:34:17 INFO mapred.Task: Task:attempt_local2107275415_0001_r_00
0000_0 is done. And is in the process of committing
16/01/26 01:34:17 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 01:34:17 INFO mapred.Task: Task attempt_local2107275415_0001_r_00
0000_0 is allowed to commit now
16/01/26 01:34:17 INFO output.FileOutputCommitter: Saved output of task 'a
ttempt_local2107275415_0001_r_000000_0' to hdfs://localhost:54310/user/roo
t/wk2/hw22/output_1/_temporary/0/task_local2107275415_0001_r_000000
16/01/26 01:34:17 INFO mapred.LocalJobRunner: reduce > reduce
16/01/26 01:34:17 INFO mapred.Task: Task 'attempt_local2107275415_0001_r_0
00000_0' done.
16/01/26 01:34:17 INFO mapred.LocalJobRunner: Finishing task: attempt_loca
l2107275415_0001_r_000000_0
16/01/26 01:34:17 INFO mapred.LocalJobRunner: reduce task executor complet
e.
16/01/26 01:34:18 INFO mapreduce.Job:  map 100% reduce 100%
16/01/26 01:34:18 INFO mapreduce.Job: Job job_local2107275415_0001 complet
ed successfully
16/01/26 01:34:18 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=348994
                FILE: Number of bytes written=973680
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=114156
                HDFS: Number of bytes written=57078
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=5740
                Map output records=5740
                Map output bytes=57078
                Map output materialized bytes=68564
                Input split bytes=111
                Combine input records=0
                Combine output records=0
                Reduce input groups=107
                Reduce shuffle bytes=68564
                Reduce input records=5740
                Reduce output records=5740
                Spilled Records=11480
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=37
                Total committed heap usage (bytes)=1013448704
        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=57078
        File Output Format Counters
                Bytes Written=57078
16/01/26 01:34:18 INFO streaming.StreamJob: Output directory: /user/root/w
k2/hw22/output_1
```

In [117]:

```
!rm part*
!hdfs dfs -copyToLocal /user/root/wk2/hw22/output_1/part*
```

```
16/01/26 01:34:25 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 01:34:26 WARN hdfs.DFSClient: DFSInputStream has been closed alre
ady
```

Printing top 10 occurring tokens

```
In [120]: !tail -10 part-00000 |sort -nrk1
```

```
1246    the
961     to
661     and
560     of
529     a
427     you
415     in
391     your
373     for
260     this
```

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

$$Pr(X = \text{``}assistance\text{''}|Y = SPAM)$$

is calculated as follows:

$$\frac{\text{the number of times ``assistance'' occurs in SPAM labeled documents}}{\text{the number of words in documents labeled SPAM}}$$

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 [245]:

```
%%writefile mapper1.py
#!/usr/bin/python
## mapper231.py
## Author: Prabhakar Gundugola
## Description: mapper-1 code for HW2.3

import sys
import re
import string

WORD_RE = re.compile(r"[\w']+") # Compile regex to easily parse complete words

spam_email_count = 0
ham_email_count = 0

total_ham_count = 0
total_spam_count = 0

for line in sys.stdin:

    # Tokenize each line. Line format - DOC_ID <tab> SPAM <tab> subject <tab> body
    tokens = line.lower().strip().split('\t')

    spam = tokens[1]
    if spam == '1':
        spam_email_count += 1
    else:
        ham_email_count += 1

    # Concatenate subject and body fields and store it in word_string
    word_string = tokens[2].strip() + ' ' + tokens[3].strip()

    # Remove punctuation
    word_string = word_string.translate(string.maketrans("", ""),
                                        string.punctuation)

    word_list = WORD_RE.findall(word_string)
    # Get unique words
    unique_words = set(word_list)

    for word in unique_words:
        word_freq = word_list.count(word)
        print('{0}\t{1}\t{2}'.format(word, spam, str(word_freq)))

print('{0}\t{1}\t{2}'.format('0000000DOC_CLASS',
                                   str(spam_email_count),
                                   str(ham_email_count)))
```

```
Overwriting mapper1.py
```

In [246]:

```
%%writefile reducer1.py
#!/usr/bin/python
## reducer231.py
## Author: Prabhakar Gundugola
## Description: reducer-1 code for HW2.3

import sys
import math

spam_email_count = 0
ham_email_count = 0

spam_word_count = 0
ham_word_count = 0

total_spam_count = 0
total_ham_count = 0

total_docs = 0

spam_words_freq = {}
ham_words_freq = {}

priors_calc = 0

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

    if tokens[0] == "0000000DOC_CLASS":
        spam_email_count = int(tokens[1])
        ham_email_count = int(tokens[2])
        continue

    try:
        count = int(tokens[2])
        spam = int(tokens[1])
    except ValueError:
        continue


    word, spam, frequency = tokens[0], spam, count
    if spam == 1:
        total_spam_count += frequency
        if word in spam_words_freq:
            spam_words_freq[word] += frequency
        else:
            spam_words_freq[word] = frequency

        if word not in ham_words_freq:
            ham_words_freq[word] = 0
    else:
        total_ham_count += frequency
        if word in ham_words_freq:
            ham_words_freq[word] += frequency
        else:
            ham_words_freq[word] = frequency

        if word not in spam_words_freq:
            spam_words_freq[word] = 0

prior_spam = math.log(1.0*spam_email_count/(spam_email_count + ham_email_co
unt))
prior_ham = math.log(1.0*ham_email_count/(spam_email_count + ham_email_coun
t))
print('{0}\t{1}\t{2}'.format("0000000PRIORS", str(prior_spam), str(prior_ha
m)))


# Calculate conditional probability
prob_spam_words = {}
prob_ham_words = {}

for word in ham_words_freq:
    if spam_words_freq[word] > 0:
        prob_spam_words[word] = math.log((1.0)*(spam_words_freq[word])/tota
l_spam_count)
    else:
        prob_spam_words[word] = 0
    if ham_words_freq[word] > 0:
        prob_ham_words[word] = math.log((1.0)*(ham_words_freq[word])/tota
l_ham_count)
    else:
        prob_ham_words[word] = 0

    print('{0}\t{1}\t{2}'.format(word, str(prob_spam_words[word]), str(pro
b_ham_words[word])))
```

```
Overwriting reducer1.py
```

In [260]:

```
%%writefile mapper2.py
#!/usr/bin/python
## mapper-2.py
## Author: Prabhakar Gundugola
## Description: mapper-2 code for HW2.3

import sys
from math import log, exp
import re
import string

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

words = {}
zero_probs_spam = 0
zero_probs_ham = 0

with open('hw23out.txt', 'rb') as fp:
    for line in fp.readlines():
        tokens = line.strip().split('\t')
        if tokens[0] == "0000000PRIORS":
            prior_spam, prior_ham = float(tokens[1]), float(tokens[2])
            continue

        words[tokens[0]] = {'p_spam':float(tokens[1]), 'p_ham':float(token
s[2])}

count = 0
for line in sys.stdin:
    tokens = line.strip().split('\t')
    count += 1

    # Concatenate subject and body fields and store it in word_string
    word_string = tokens[2] + ' ' + tokens[3].strip()

    # Remove punctuation
    word_string = word_string.translate(string.maketrans("", ""),
                                        string.punctuation)

    word_list = WORD_RE.findall(word_string)

    prior_spam_doc, prior_ham_doc = prior_spam, prior_ham

    for word in word_list:
        word = word.lower().strip()
        try:
            if words[word]['p_spam'] <> 0:
                prior_spam_doc += words[word]['p_spam']
            else:
                prior_spam_doc += float('-inf')
                zero_probs_spam += 1

            if words[word]['p_ham'] <> 0:
                prior_ham_doc += words[word]['p_ham']
            else:
                prior_ham_doc += float('-inf')
                zero_probs_ham += 1
        except:
            continue


    predicted = 0
    if prior_spam_doc == float('-inf'):
        predicted = 0
        prior_spam_doc = 0
    elif prior_ham_doc == float('-inf'):
        predicted = 1
        prior_ham_doc = 0
    elif prior_spam_doc > prior_ham_doc:
        predicted = 1

    values = tokens[1] + '\t' + str(predicted) + '\t' + str(prior_spam_doc)
    values += '\t' + str(prior_ham_doc) + '\t' + str(zero_probs_spam)
    values += '\t' + str(zero_probs_ham)
    print tokens[0] + '\t' + values
```

```
Overwriting mapper2.py
```

In [269]:

```
%%writefile reducer2.py
#!/usr/bin/python
## reducer-2.py
## Author: Prabhakar Gundugola
## Description: reducer-2 code for HW2.3

import sys

count = 0
correct = 0
wrong = 0
zero_probs_spam = 0
zero_probs_ham = 0

for line in sys.stdin:
    tokens = line.strip().split('\t')
    spam, predicted = int(tokens[1]), int(tokens[2])
    count += 1
    zero_probs_spam += int(tokens[5])
    zero_probs_ham += int(tokens[6])
    if spam == predicted:
        correct += 1
    else:
        wrong += 1

    print line

training_error = 100.0*wrong/count
accuracy = 100.0*correct/count

print 'Training error: {0}%'.format(training_error)
print 'Accuracy: {0}%'.format(accuracy)
```

```
Overwriting reducer2.py
```

In [236]:

```
!chmod a+x mapper1.py
!chmod a+x mapper2.py
!chmod a+x reducer1.py
!chmod a+x reducer2.py
```

Command line testing

In [248]:

```
!cat enronemail_1h.txt|./mapper1.py|sort|./reducer1.py > hw23out.txt
```

In [265]:

```
!cat enronemail_1h.txt|./mapper2.py|sort|./reducer2.py
```

```
Training error: 0.0%
Accuracy: 100.0%
```

In [270]:

```
!cat enronemail_1h.txt|./mapper2.py|sort|./reducer2.py > histogram.txt
!grep assistance histogram.txt
```

Running the Hadoop Streaming for the first MapReduce job

In [272]:

```
# Ensure the output folder doesn't exist
!hdfs dfs -rm -r /user/root/wk2/hw23
!hdfs dfs -mkdir -p /user/root/wk2/hw23/input

!hdfs dfs -put enronemail_1h.txt /user/root/wk2/hw23/input
```

```
16/01/26 08:27:48 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
rm: `/user/root/wk2/hw23': No such file or directory
16/01/26 08:27:51 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 08:27:54 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
```

In [273]:

```
# Delete output folder if exists
!hdfs dfs -rm -r /user/root/wk2/hw23/output_1

# Run Hadoop Streaming job
!hadoop jar hadoop-streaming-2.7.1.jar \
-mapper mapper1.py \
-reducer reducer1.py \
-input /user/root/wk2/hw23/input \
-output /user/root/wk2/hw23/output_1
```

```
16/01/26 08:28:44 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
rm: `/user/root/wk2/hw23/output_1': No such file or directory
16/01/26 08:28:46 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 08:28:47 INFO Configuration.deprecation: session.id is deprecate
d. Instead, use dfs.metrics.session-id
16/01/26 08:28:47 INFO jvm.JvmMetrics: Initializing JVM Metrics with proce
ssName=JobTracker, sessionId=
16/01/26 08:28:47 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with
processName=JobTracker, sessionId= - already initialized
16/01/26 08:28:48 INFO mapred.FileInputFormat: Total input paths to proces
s : 1
16/01/26 08:28:48 INFO mapreduce.JobSubmitter: number of splits:1
16/01/26 08:28:48 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_local998646625_0001
16/01/26 08:28:48 INFO mapreduce.Job: The url to track the job: http://loc
alhost:8080/
16/01/26 08:28:48 INFO mapreduce.Job: Running job: job_local998646625_0001
16/01/26 08:28:48 INFO mapred.LocalJobRunner: OutputCommitter set in confi
g null
16/01/26 08:28:48 INFO mapred.LocalJobRunner: OutputCommitter is org.apach
e.hadoop.mapred.FileOutputCommitter
16/01/26 08:28:48 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 08:28:48 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/26 08:28:48 INFO mapred.LocalJobRunner: Starting task: attempt_local
998646625_0001_m_000000_0
16/01/26 08:28:48 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 08:28:48 INFO mapred.Task:  Using ResourceCalculatorProcessTree :
[ ]
16/01/26 08:28:48 INFO mapred.MapTask: Processing split: hdfs://localhos
t:54310/user/root/wk2/hw23/input/enronemail_1h.txt:0+203954
16/01/26 08:28:48 INFO mapred.MapTask: numReduceTasks: 1
16/01/26 08:28:49 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/26 08:28:49 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/26 08:28:49 INFO mapred.MapTask: soft limit at 83886080
16/01/26 08:28:49 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/26 08:28:49 INFO mapred.MapTask: kvstart = 26214396; length = 655360
0
16/01/26 08:28:49 INFO mapred.MapTask: Map output collector class = org.ap
ache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/26 08:28:49 INFO streaming.PipeMapRed: PipeMapRed exec [/root/hw
2/./mapper1.py]
16/01/26 08:28:49 INFO Configuration.deprecation: mapred.task.id is deprec
ated. Instead, use mapreduce.task.attempt.id
16/01/26 08:28:49 INFO Configuration.deprecation: user.name is deprecated.
Instead, use mapreduce.job.user.name
16/01/26 08:28:49 INFO Configuration.deprecation: map.input.start is depre
cated. Instead, use mapreduce.map.input.start
16/01/26 08:28:49 INFO Configuration.deprecation: mapred.task.is.map is de
precated. Instead, use mapreduce.task.ismap
16/01/26 08:28:49 INFO Configuration.deprecation: mapred.tip.id is depreca
ted. Instead, use mapreduce.task.id
```

```
16/01/26 08:28:49 INFO Configuration.deprecation: mapred.skip.on is deprec
ated. Instead, use mapreduce.job.skiprecords
16/01/26 08:28:49 INFO Configuration.deprecation: mapred.task.partition is
deprecated. Instead, use mapreduce.task.partition
16/01/26 08:28:49 INFO Configuration.deprecation: map.input.length is depr
ecated. Instead, use mapreduce.map.input.length
16/01/26 08:28:49 INFO Configuration.deprecation: mapred.local.dir is depr
ecated. Instead, use mapreduce.cluster.local.dir
16/01/26 08:28:49 INFO Configuration.deprecation: mapred.work.output.dir i
s deprecated. Instead, use mapreduce.task.output.dir
16/01/26 08:28:49 INFO Configuration.deprecation: map.input.file is deprec
ated. Instead, use mapreduce.map.input.file
16/01/26 08:28:49 INFO Configuration.deprecation: mapred.job.id is depreca
ted. Instead, use mapreduce.job.id
16/01/26 08:28:49 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 08:28:49 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 08:28:49 INFO streaming.PipeMapRed: Records R/W=72/1
16/01/26 08:28:49 INFO streaming.PipeMapRed: R/W/S=100/1999/0 in:NA [re
c/s] out:NA [rec/s]
16/01/26 08:28:49 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 08:28:49 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 08:28:49 INFO mapred.LocalJobRunner:
16/01/26 08:28:49 INFO mapred.MapTask: Starting flush of map output
16/01/26 08:28:49 INFO mapred.MapTask: Spilling map output
16/01/26 08:28:49 INFO mapred.MapTask: bufstart = 0; bufend = 164166; bufv
oid = 104857600
16/01/26 08:28:49 INFO mapred.MapTask: kvstart = 26214396(104857584); kven
d = 26153056(104612224); length = 61341/6553600
16/01/26 08:28:49 INFO mapred.MapTask: Finished spill 0
16/01/26 08:28:49 INFO mapred.Task: Task:attempt_local998646625_0001_m_000
000_0 is done. And is in the process of committing
16/01/26 08:28:49 INFO mapred.LocalJobRunner: Records R/W=72/1
16/01/26 08:28:49 INFO mapred.Task: Task 'attempt_local998646625_0001_m_00
0000_0' done.
16/01/26 08:28:49 INFO mapred.LocalJobRunner: Finishing task: attempt_loca
l998646625_0001_m_000000_0
16/01/26 08:28:49 INFO mapred.LocalJobRunner: map task executor complete.
16/01/26 08:28:49 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/26 08:28:49 INFO mapred.LocalJobRunner: Starting task: attempt_local
998646625_0001_r_000000_0
16/01/26 08:28:49 INFO mapreduce.Job: Job job_local998646625_0001 running
in uber mode : false
16/01/26 08:28:49 INFO mapreduce.Job:  map 100% reduce 0%
16/01/26 08:28:49 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 08:28:49 INFO mapred.Task:  Using ResourceCalculatorProcessTree :
[ ]
16/01/26 08:28:49 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: or
g.apache.hadoop.mapreduce.task.reduce.Shuffle@7ddaca36
16/01/26 08:28:49 INFO reduce.MergeManagerImpl: MergerManager: memoryLimi
t=353422528, maxSingleShuffleLimit=88355632, mergeThreshold=233258880, ioS
ortFactor=10, memToMemMergeOutputsThreshold=10
16/01/26 08:28:49 INFO reduce.EventFetcher: attempt_local998646625_000
1_r_000000_0 Thread started: EventFetcher for fetching Map Completion Even
```

```
ts
16/01/26 08:28:49 INFO reduce.LocalFetcher: localfetcher#1 about to shuffl
e output of map attempt_local998646625_0001_m_000000_0 decomp: 194840 len:
194844 to MEMORY
16/01/26 08:28:49 INFO reduce.InMemoryMapOutput: Read 194840 bytes from ma
p-output for attempt_local998646625_0001_m_000000_0
16/01/26 08:28:49 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-o
utput of size: 194840, inMemoryMapOutputs.size() -> 1, commitMemory -> 0,
usedMemory ->194840
16/01/26 08:28:49 INFO reduce.EventFetcher: EventFetcher is interrupted..
Returning
16/01/26 08:28:49 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 08:28:49 INFO reduce.MergeManagerImpl: finalMerge called with 1 i
n-memory map-outputs and 0 on-disk map-outputs
16/01/26 08:28:49 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 08:28:49 INFO mapred.Merger: Down to the last merge-pass, with 1
segments left of total size: 194836 bytes
16/01/26 08:28:50 INFO reduce.MergeManagerImpl: Merged 1 segments, 194840
bytes to disk to satisfy reduce memory limit
16/01/26 08:28:50 INFO reduce.MergeManagerImpl: Merging 1 files, 194844 by
tes from disk
16/01/26 08:28:50 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 byte
s from memory into reduce
16/01/26 08:28:50 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 08:28:50 INFO mapred.Merger: Down to the last merge-pass, with 1
segments left of total size: 194836 bytes
16/01/26 08:28:50 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 08:28:50 INFO streaming.PipeMapRed: PipeMapRed exec [/root/hw
2/./reducer1.py]
16/01/26 08:28:50 INFO Configuration.deprecation: mapred.job.tracker is de
precated. Instead, use mapreduce.jobtracker.address
16/01/26 08:28:50 INFO Configuration.deprecation: mapred.map.tasks is depr
ecated. Instead, use mapreduce.job.maps
16/01/26 08:28:50 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 08:28:50 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 08:28:50 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] o
ut:NA [rec/s]
16/01/26 08:28:50 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s]
out:NA [rec/s]
16/01/26 08:28:50 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s]
out:NA [rec/s]
16/01/26 08:28:50 INFO streaming.PipeMapRed: Records R/W=15336/1
16/01/26 08:28:50 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 08:28:50 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 08:28:50 INFO mapred.Task: Task:attempt_local998646625_0001_r_000
000_0 is done. And is in the process of committing
16/01/26 08:28:50 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 08:28:50 INFO mapred.Task: Task attempt_local998646625_0001_r_000
000_0 is allowed to commit now
16/01/26 08:28:50 INFO output.FileOutputCommitter: Saved output of task 'a
ttempt_local998646625_0001_r_000000_0' to hdfs://localhost:54310/user/roo
t/wk2/hw23/output_1/_temporary/0/task_local998646625_0001_r_000000
16/01/26 08:28:50 INFO mapred.LocalJobRunner: Records R/W=15336/1 > reduce
16/01/26 08:28:50 INFO mapred.Task: Task 'attempt_local998646625_0001_r_00
```

```
0000_0' done.
16/01/26 08:28:50 INFO mapred.LocalJobRunner: Finishing task: attempt_loca
l998646625_0001_r_000000_0
16/01/26 08:28:50 INFO mapred.LocalJobRunner: reduce task executor complet
e.
16/01/26 08:28:50 INFO mapreduce.Job:  map 100% reduce 100%
16/01/26 08:28:50 INFO mapreduce.Job: Job job_local998646625_0001 complete
d successfully
16/01/26 08:28:50 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=601568
                FILE: Number of bytes written=1350020
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=407908
                HDFS: Number of bytes written=155886
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=100
                Map output records=15336
                Map output bytes=164166
                Map output materialized bytes=194844
                Input split bytes=117
                Combine input records=0
                Combine output records=0
                Reduce input groups=5741
                Reduce shuffle bytes=194844
                Reduce input records=15336
                Reduce output records=5741
                Spilled Records=30672
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=38
                Total committed heap usage (bytes)=1013448704
        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=203954
        File Output Format Counters
                Bytes Written=155886
16/01/26 08:28:50 INFO streaming.StreamJob: Output directory: /user/root/w
k2/hw23/output_1
```

In [275]:

```
!rm part*

# Copy the mapper output to local directory
!hdfs dfs -copyToLocal /user/root/wk2/hw23/output_1/part*
!mv part-00000 hw23out.txt
```

```
16/01/26 08:32:11 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 08:32:12 WARN hdfs.DFSClient: DFSInputStream has been closed alre
ady
```

In [276]:

```
# Delete output folder if exists
!hdfs dfs -rm -r /user/root/wk2/hw23/output_2

# Run Hadoop Streaming job
!hadoop jar hadoop-streaming-2.7.1.jar \
-mapper mapper2.py \
-reducer reducer2.py \
-input /user/root/wk2/hw23/input \
-output /user/root/wk2/hw23/output_2
```

```
16/01/26 08:33:00 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
rm: `/user/root/wk2/hw23/output_2': No such file or directory
16/01/26 08:33:03 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 08:33:04 INFO Configuration.deprecation: session.id is deprecate
d. Instead, use dfs.metrics.session-id
16/01/26 08:33:04 INFO jvm.JvmMetrics: Initializing JVM Metrics with proce
ssName=JobTracker, sessionId=
16/01/26 08:33:04 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with
processName=JobTracker, sessionId= - already initialized
16/01/26 08:33:04 INFO mapred.FileInputFormat: Total input paths to proces
s : 1
16/01/26 08:33:04 INFO mapreduce.JobSubmitter: number of splits:1
16/01/26 08:33:04 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_local1126097630_0001
16/01/26 08:33:05 INFO mapreduce.Job: The url to track the job: http://loc
alhost:8080/
16/01/26 08:33:05 INFO mapreduce.Job: Running job: job_local1126097630_000
1
16/01/26 08:33:05 INFO mapred.LocalJobRunner: OutputCommitter set in confi
g null
16/01/26 08:33:05 INFO mapred.LocalJobRunner: OutputCommitter is org.apach
e.hadoop.mapred.FileOutputCommitter
16/01/26 08:33:05 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 08:33:05 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/26 08:33:05 INFO mapred.LocalJobRunner: Starting task: attempt_local
1126097630_0001_m_000000_0
16/01/26 08:33:05 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 08:33:05 INFO mapred.Task:  Using ResourceCalculatorProcessTree :
[ ]
16/01/26 08:33:05 INFO mapred.MapTask: Processing split: hdfs://localhos
t:54310/user/root/wk2/hw23/input/enronemail_1h.txt:0+203954
16/01/26 08:33:05 INFO mapred.MapTask: numReduceTasks: 1
16/01/26 08:33:05 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/26 08:33:05 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/26 08:33:05 INFO mapred.MapTask: soft limit at 83886080
16/01/26 08:33:05 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/26 08:33:05 INFO mapred.MapTask: kvstart = 26214396; length = 655360
0
16/01/26 08:33:05 INFO mapred.MapTask: Map output collector class = org.ap
ache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/26 08:33:05 INFO streaming.PipeMapRed: PipeMapRed exec [/root/hw
2/./mapper2.py]
16/01/26 08:33:05 INFO Configuration.deprecation: mapred.task.id is deprec
ated. Instead, use mapreduce.task.attempt.id
16/01/26 08:33:05 INFO Configuration.deprecation: user.name is deprecated.
Instead, use mapreduce.job.user.name
16/01/26 08:33:05 INFO Configuration.deprecation: map.input.start is depre
cated. Instead, use mapreduce.map.input.start
16/01/26 08:33:05 INFO Configuration.deprecation: mapred.task.is.map is de
precated. Instead, use mapreduce.task.ismap
16/01/26 08:33:05 INFO Configuration.deprecation: mapred.tip.id is depreca
```

```
ted. Instead, use mapreduce.task.id
16/01/26 08:33:05 INFO Configuration.deprecation: mapred.skip.on is deprec
ated. Instead, use mapreduce.job.skiprecords
16/01/26 08:33:05 INFO Configuration.deprecation: mapred.task.partition is
deprecated. Instead, use mapreduce.task.partition
16/01/26 08:33:05 INFO Configuration.deprecation: map.input.length is depr
ecated. Instead, use mapreduce.map.input.length
16/01/26 08:33:05 INFO Configuration.deprecation: mapred.local.dir is depr
ecated. Instead, use mapreduce.cluster.local.dir
16/01/26 08:33:05 INFO Configuration.deprecation: mapred.work.output.dir i
s deprecated. Instead, use mapreduce.task.output.dir
16/01/26 08:33:05 INFO Configuration.deprecation: map.input.file is deprec
ated. Instead, use mapreduce.map.input.file
16/01/26 08:33:05 INFO Configuration.deprecation: mapred.job.id is depreca
ted. Instead, use mapreduce.job.id
16/01/26 08:33:05 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 08:33:05 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 08:33:05 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] o
ut:NA [rec/s]
16/01/26 08:33:05 INFO streaming.PipeMapRed: Records R/W=100/1
16/01/26 08:33:05 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 08:33:05 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 08:33:05 INFO mapred.LocalJobRunner:
16/01/26 08:33:05 INFO mapred.MapTask: Starting flush of map output
16/01/26 08:33:05 INFO mapred.MapTask: Spilling map output
16/01/26 08:33:05 INFO mapred.MapTask: bufstart = 0; bufend = 5275; bufvoi
d = 104857600
16/01/26 08:33:05 INFO mapred.MapTask: kvstart = 26214396(104857584); kven
d = 26214000(104856000); length = 397/6553600
16/01/26 08:33:05 INFO mapred.MapTask: Finished spill 0
16/01/26 08:33:05 INFO mapred.Task: Task:attempt_local1126097630_0001_m_00
0000_0 is done. And is in the process of committing
16/01/26 08:33:05 INFO mapred.LocalJobRunner: Records R/W=100/1
16/01/26 08:33:05 INFO mapred.Task: Task 'attempt_local1126097630_0001_m_0
00000_0' done.
16/01/26 08:33:05 INFO mapred.LocalJobRunner: Finishing task: attempt_loca
l1126097630_0001_m_000000_0
16/01/26 08:33:05 INFO mapred.LocalJobRunner: map task executor complete.
16/01/26 08:33:05 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/26 08:33:05 INFO mapred.LocalJobRunner: Starting task: attempt_local
1126097630_0001_r_000000_0
16/01/26 08:33:05 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 08:33:05 INFO mapred.Task:  Using ResourceCalculatorProcessTree :
[ ]
16/01/26 08:33:05 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: or
g.apache.hadoop.mapreduce.task.reduce.Shuffle@714005f6
16/01/26 08:33:05 INFO reduce.MergeManagerImpl: MergerManager: memoryLimi
t=353422528, maxSingleShuffleLimit=88355632, mergeThreshold=233258880, ioS
ortFactor=10, memToMemMergeOutputsThreshold=10
16/01/26 08:33:05 INFO reduce.EventFetcher: attempt_local1126097630_000
1_r_000000_0 Thread started: EventFetcher for fetching Map Completion Even
ts
16/01/26 08:33:05 INFO reduce.LocalFetcher: localfetcher#1 about to shuffl
```

```
e output of map attempt_local1126097630_0001_m_000000_0 decomp: 5477 len:
5481 to MEMORY
16/01/26 08:33:05 INFO reduce.InMemoryMapOutput: Read 5477 bytes from map-
output for attempt_local1126097630_0001_m_000000_0
16/01/26 08:33:05 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-o
utput of size: 5477, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, us
edMemory ->5477
16/01/26 08:33:05 INFO reduce.EventFetcher: EventFetcher is interrupted..
Returning
16/01/26 08:33:05 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 08:33:05 INFO reduce.MergeManagerImpl: finalMerge called with 1 i
n-memory map-outputs and 0 on-disk map-outputs
16/01/26 08:33:05 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 08:33:05 INFO mapred.Merger: Down to the last merge-pass, with 1
segments left of total size: 5452 bytes
16/01/26 08:33:05 INFO reduce.MergeManagerImpl: Merged 1 segments, 5477 by
tes to disk to satisfy reduce memory limit
16/01/26 08:33:05 INFO reduce.MergeManagerImpl: Merging 1 files, 5481 byte
s from disk
16/01/26 08:33:05 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 byte
s from memory into reduce
16/01/26 08:33:05 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 08:33:05 INFO mapred.Merger: Down to the last merge-pass, with 1
segments left of total size: 5452 bytes
16/01/26 08:33:05 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 08:33:05 INFO streaming.PipeMapRed: PipeMapRed exec [/root/hw
2/./reducer2.py]
16/01/26 08:33:05 INFO Configuration.deprecation: mapred.job.tracker is de
precated. Instead, use mapreduce.jobtracker.address
16/01/26 08:33:05 INFO Configuration.deprecation: mapred.map.tasks is depr
ecated. Instead, use mapreduce.job.maps
16/01/26 08:33:06 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 08:33:06 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 08:33:06 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] o
ut:NA [rec/s]
16/01/26 08:33:06 INFO streaming.PipeMapRed: Records R/W=100/1
16/01/26 08:33:06 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 08:33:06 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 08:33:06 INFO mapreduce.Job: Job job_local1126097630_0001 running
in uber mode : false
16/01/26 08:33:06 INFO mapreduce.Job:  map 100% reduce 0%
16/01/26 08:33:06 INFO mapred.Task: Task:attempt_local1126097630_0001_r_00
0000_0 is done. And is in the process of committing
16/01/26 08:33:06 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 08:33:06 INFO mapred.Task: Task attempt_local1126097630_0001_r_00
0000_0 is allowed to commit now
16/01/26 08:33:06 INFO output.FileOutputCommitter: Saved output of task 'a
ttempt_local1126097630_0001_r_000000_0' to hdfs://localhost:54310/user/roo
t/wk2/hw23/output_2/_temporary/0/task_local1126097630_0001_r_000000
16/01/26 08:33:06 INFO mapred.LocalJobRunner: Records R/W=100/1 > reduce
16/01/26 08:33:06 INFO mapred.Task: Task 'attempt_local1126097630_0001_r_0
00000_0' done.
16/01/26 08:33:06 INFO mapred.LocalJobRunner: Finishing task: attempt_loca
l1126097630_0001_r_000000_0
```

```
16/01/26 08:33:06 INFO mapred.LocalJobRunner: reduce task executor complet
e.
16/01/26 08:33:07 INFO mapreduce.Job:  map 100% reduce 100%
16/01/26 08:33:07 INFO mapreduce.Job: Job job_local1126097630_0001 complet
ed successfully
16/01/26 08:33:07 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=222842
                FILE: Number of bytes written=784939
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=407908
                HDFS: Number of bytes written=5515
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=100
                Map output records=100
                Map output bytes=5275
                Map output materialized bytes=5481
                Input split bytes=117
                Combine input records=0
                Combine output records=0
                Reduce input groups=100
                Reduce shuffle bytes=5481
                Reduce input records=100
                Reduce output records=202
                Spilled Records=200
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=36
                Total committed heap usage (bytes)=1013448704
        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=203954
        File Output Format Counters
                Bytes Written=5515
16/01/26 08:33:07 INFO streaming.StreamJob: Output directory: /user/root/w
k2/hw23/output_2
```

In [287]:

```
!hdfs dfs -cat /user/root/wk2/hw23/output_2/part-00000|tail -3
```

```
16/01/26 09:00:21 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable

Training error: 0.0%
Accuracy: 100.0%
```

In [279]:

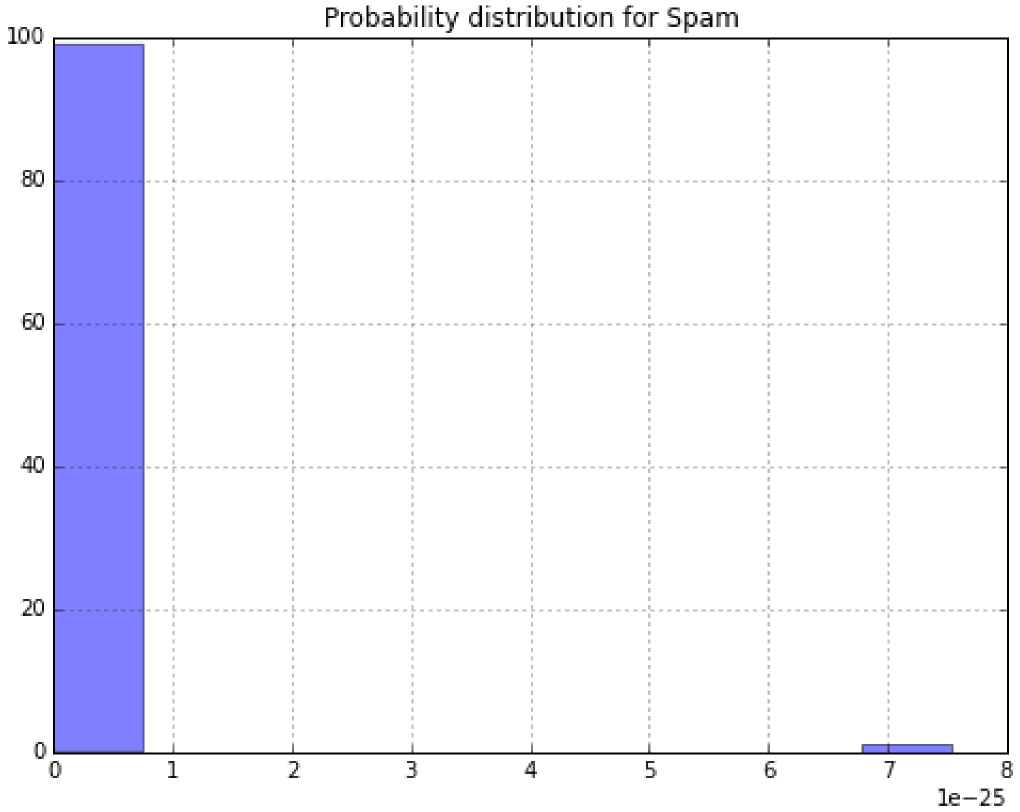
```
!rm histogram*

# Copy the mapper output to local directory
!hdfs dfs -copyToLocal /user/root/wk2/hw23/output_2/part*
!mv part-00000 histogram.txt

!head -n $(($(wc -l < histogram.txt) - 1)) < histogram.txt > histogram_plo
t.txt
```

```
16/01/26 08:38:55 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 08:38:56 WARN hdfs.DFSClient: DFSInputStream has been closed alre
ady
```

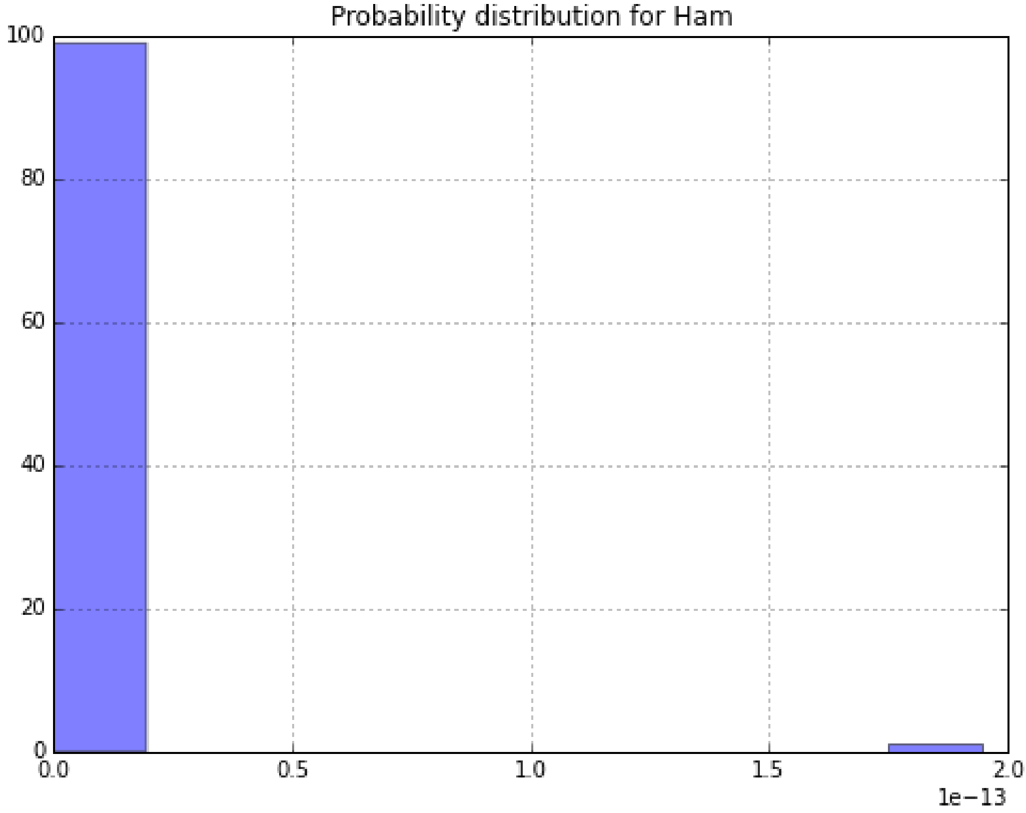
In [281]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline

df = pd.read_csv("histogram_plot.txt", sep='\t', header=None)
df.columns = ['docid', 'spam', 'predicted', 'pr_spam', 'pr_ham', 'spam0',
'ham0']
e_spam = np.exp(df.pr_spam[df.pr_spam!=0])
df['e_spam'] = np.where(df['pr_spam']==0, 0.0, np.exp(df['pr_spam']))
df['e_ham'] = np.where(df['pr_ham']==0, 0.0, np.exp(df['pr_ham']))

plt.figure()
plt.figure(figsize=(8,6), dpi=30)
plt.title('Probability distribution for Spam')
df.e_spam.hist(alpha=0.5)

plt.figure()
plt.figure(figsize=(8,6), dpi=30)
plt.title('Probability distribution for Ham')
df.e_ham.hist(alpha=0.5)
```

```
Out[281]: <matplotlib.axes._subplots.AxesSubplot at 0x7f2ee268c350>

          <matplotlib.figure.Figure at 0x7f2ee267bc10>
```


Probability distribution for Spam
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
1e−25

```
          <matplotlib.figure.Figure at 0x7f2ee268ce50>
```


Probability distribution for Ham
100
80
60
40
20
0
0.0
0.5
1.0
1.5
2.0
1e−13

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)

The problem statement is similar to HW2.3 and therefore the structure for this problem is almost the same except for reducer1.py. We will reuse the remaining mappers and reducers

In [282]:

```
%%writefile reducer3.py
#!/usr/bin/python
## reducer3.py
## Author: Prabhakar Gundugola
## Description: reducer3 code for HW2.4

import sys
import math

spam_email_count = 0
ham_email_count = 0

spam_word_count = 0
ham_word_count = 0

total_spam_count = 0
total_ham_count = 0

total_docs = 0

spam_words_freq = {}
ham_words_freq = {}

priors_calc = 0

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

    if tokens[0] == "0000000DOC_CLASS":
        spam_email_count = int(tokens[1])
        ham_email_count = int(tokens[2])
        continue

    try:
        count = int(tokens[2])
        spam = int(tokens[1])
    except ValueError:
        continue


    word, spam, frequency = tokens[0], spam, count
    if spam == 1:
        total_spam_count += frequency
        if word in spam_words_freq:
            spam_words_freq[word] += frequency
        else:
            spam_words_freq[word] = frequency

        if word not in ham_words_freq:
            ham_words_freq[word] = 0
    else:
        total_ham_count += frequency
        if word in ham_words_freq:
            ham_words_freq[word] += frequency
        else:
            ham_words_freq[word] = frequency

        if word not in spam_words_freq:
            spam_words_freq[word] = 0

prior_spam = math.log(1.0*spam_email_count/(spam_email_count + ham_email_co
unt))
prior_ham = math.log(1.0*ham_email_count/(spam_email_count + ham_email_coun
t))
print('{0}\t{1}\t{2}'.format("0000000PRIORS", str(prior_spam), str(prior_ha
m)))


# Calculate conditional probability.
# Applied Laplace smoothing to math.log function in the numerator
prob_spam_words = {}
prob_ham_words = {}

for word in ham_words_freq:
    if spam_words_freq[word] > 0:
        prob_spam_words[word] = math.log((1.0)*(spam_words_freq[word]+1)/to
tal_spam_count)
    else:
        prob_spam_words[word] = 0
    if ham_words_freq[word] > 0:
        prob_ham_words[word] = math.log((1.0)*(ham_words_freq[word]+1)/tota
l_ham_count)
    else:
        prob_ham_words[word] = 0

    print('{0}\t{1}\t{2}'.format(word, str(prob_spam_words[word]), str(pro
b_ham_words[word])))
```

```
Writing reducer3.py
```

In [283]:

```
# Ensure the input folder doesn't exist
!hdfs dfs -rm -r /user/root/wk2/hw24
!hdfs dfs -mkdir -p /user/root/wk2/hw24/input

!hdfs dfs -put enronemail_1h.txt /user/root/wk2/hw24/input
```

```
16/01/26 08:56:09 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
rm: `/user/root/wk2/hw24': No such file or directory
16/01/26 08:56:12 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 08:56:15 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
```

In [284]:

```
# Delete output folder if exists
!hdfs dfs -rm -r /user/root/wk2/hw24/output_1

# Run Hadoop Streaming job
!hadoop jar hadoop-streaming-2.7.1.jar \
-mapper mapper1.py \
-reducer reducer3.py \
-input /user/root/wk2/hw24/input \
-output /user/root/wk2/hw24/output_1
```

```
16/01/26 08:56:51 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
rm: `/user/root/wk2/hw24/output_1': No such file or directory
16/01/26 08:56:54 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 08:56:55 INFO Configuration.deprecation: session.id is deprecate
d. Instead, use dfs.metrics.session-id
16/01/26 08:56:55 INFO jvm.JvmMetrics: Initializing JVM Metrics with proce
ssName=JobTracker, sessionId=
16/01/26 08:56:55 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with
processName=JobTracker, sessionId= - already initialized
16/01/26 08:56:55 INFO mapred.FileInputFormat: Total input paths to proces
s : 1
16/01/26 08:56:55 INFO mapreduce.JobSubmitter: number of splits:1
16/01/26 08:56:55 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_local392316276_0001
16/01/26 08:56:56 INFO mapreduce.Job: The url to track the job: http://loc
alhost:8080/
16/01/26 08:56:56 INFO mapred.LocalJobRunner: OutputCommitter set in confi
g null
16/01/26 08:56:56 INFO mapreduce.Job: Running job: job_local392316276_0001
16/01/26 08:56:56 INFO mapred.LocalJobRunner: OutputCommitter is org.apach
e.hadoop.mapred.FileOutputCommitter
16/01/26 08:56:56 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 08:56:56 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/26 08:56:56 INFO mapred.LocalJobRunner: Starting task: attempt_local
392316276_0001_m_000000_0
16/01/26 08:56:56 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 08:56:56 INFO mapred.Task:  Using ResourceCalculatorProcessTree :
[ ]
16/01/26 08:56:56 INFO mapred.MapTask: Processing split: hdfs://localhos
t:54310/user/root/wk2/hw24/input/enronemail_1h.txt:0+203954
16/01/26 08:56:56 INFO mapred.MapTask: numReduceTasks: 1
16/01/26 08:56:56 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/26 08:56:56 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/26 08:56:56 INFO mapred.MapTask: soft limit at 83886080
16/01/26 08:56:56 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/26 08:56:56 INFO mapred.MapTask: kvstart = 26214396; length = 655360
0
16/01/26 08:56:56 INFO mapred.MapTask: Map output collector class = org.ap
ache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/26 08:56:56 INFO streaming.PipeMapRed: PipeMapRed exec [/root/hw
2/./mapper1.py]
16/01/26 08:56:56 INFO Configuration.deprecation: mapred.task.id is deprec
ated. Instead, use mapreduce.task.attempt.id
16/01/26 08:56:56 INFO Configuration.deprecation: user.name is deprecated.
Instead, use mapreduce.job.user.name
16/01/26 08:56:56 INFO Configuration.deprecation: map.input.start is depre
cated. Instead, use mapreduce.map.input.start
16/01/26 08:56:56 INFO Configuration.deprecation: mapred.task.is.map is de
precated. Instead, use mapreduce.task.ismap
16/01/26 08:56:56 INFO Configuration.deprecation: mapred.tip.id is depreca
ted. Instead, use mapreduce.task.id
```

```
16/01/26 08:56:56 INFO Configuration.deprecation: mapred.skip.on is deprec
ated. Instead, use mapreduce.job.skiprecords
16/01/26 08:56:56 INFO Configuration.deprecation: mapred.task.partition is
deprecated. Instead, use mapreduce.task.partition
16/01/26 08:56:56 INFO Configuration.deprecation: map.input.length is depr
ecated. Instead, use mapreduce.map.input.length
16/01/26 08:56:56 INFO Configuration.deprecation: mapred.local.dir is depr
ecated. Instead, use mapreduce.cluster.local.dir
16/01/26 08:56:56 INFO Configuration.deprecation: mapred.work.output.dir i
s deprecated. Instead, use mapreduce.task.output.dir
16/01/26 08:56:56 INFO Configuration.deprecation: map.input.file is deprec
ated. Instead, use mapreduce.map.input.file
16/01/26 08:56:56 INFO Configuration.deprecation: mapred.job.id is depreca
ted. Instead, use mapreduce.job.id
16/01/26 08:56:56 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 08:56:56 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 08:56:56 INFO streaming.PipeMapRed: Records R/W=72/1
16/01/26 08:56:56 INFO streaming.PipeMapRed: R/W/S=100/2164/0 in:NA [re
c/s] out:NA [rec/s]
16/01/26 08:56:56 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 08:56:56 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 08:56:56 INFO mapred.LocalJobRunner:
16/01/26 08:56:56 INFO mapred.MapTask: Starting flush of map output
16/01/26 08:56:56 INFO mapred.MapTask: Spilling map output
16/01/26 08:56:56 INFO mapred.MapTask: bufstart = 0; bufend = 164166; bufv
oid = 104857600
16/01/26 08:56:56 INFO mapred.MapTask: kvstart = 26214396(104857584); kven
d = 26153056(104612224); length = 61341/6553600
16/01/26 08:56:57 INFO mapreduce.Job: Job job_local392316276_0001 running
in uber mode : false
16/01/26 08:56:57 INFO mapreduce.Job:  map 0% reduce 0%
16/01/26 08:56:57 INFO mapred.MapTask: Finished spill 0
16/01/26 08:56:57 INFO mapred.Task: Task:attempt_local392316276_0001_m_000
000_0 is done. And is in the process of committing
16/01/26 08:56:57 INFO mapred.LocalJobRunner: Records R/W=72/1
16/01/26 08:56:57 INFO mapred.Task: Task 'attempt_local392316276_0001_m_00
0000_0' done.
16/01/26 08:56:57 INFO mapred.LocalJobRunner: Finishing task: attempt_loca
l392316276_0001_m_000000_0
16/01/26 08:56:57 INFO mapred.LocalJobRunner: map task executor complete.
16/01/26 08:56:57 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/26 08:56:57 INFO mapred.LocalJobRunner: Starting task: attempt_local
392316276_0001_r_000000_0
16/01/26 08:56:57 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 08:56:57 INFO mapred.Task:  Using ResourceCalculatorProcessTree :
[ ]
16/01/26 08:56:57 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: or
g.apache.hadoop.mapreduce.task.reduce.Shuffle@1ece5b86
16/01/26 08:56:57 INFO reduce.MergeManagerImpl: MergerManager: memoryLimi
t=353422528, maxSingleShuffleLimit=88355632, mergeThreshold=233258880, ioS
ortFactor=10, memToMemMergeOutputsThreshold=10
16/01/26 08:56:57 INFO reduce.EventFetcher: attempt_local392316276_000
1_r_000000_0 Thread started: EventFetcher for fetching Map Completion Even
```

```
ts
16/01/26 08:56:57 INFO reduce.LocalFetcher: localfetcher#1 about to shuffl
e output of map attempt_local392316276_0001_m_000000_0 decomp: 194840 len:
194844 to MEMORY
16/01/26 08:56:57 INFO reduce.InMemoryMapOutput: Read 194840 bytes from ma
p-output for attempt_local392316276_0001_m_000000_0
16/01/26 08:56:57 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-o
utput of size: 194840, inMemoryMapOutputs.size() -> 1, commitMemory -> 0,
usedMemory ->194840
16/01/26 08:56:57 INFO reduce.EventFetcher: EventFetcher is interrupted..
Returning
16/01/26 08:56:57 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 08:56:57 INFO reduce.MergeManagerImpl: finalMerge called with 1 i
n-memory map-outputs and 0 on-disk map-outputs
16/01/26 08:56:57 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 08:56:57 INFO mapred.Merger: Down to the last merge-pass, with 1
segments left of total size: 194836 bytes
16/01/26 08:56:57 INFO reduce.MergeManagerImpl: Merged 1 segments, 194840
bytes to disk to satisfy reduce memory limit
16/01/26 08:56:57 INFO reduce.MergeManagerImpl: Merging 1 files, 194844 by
tes from disk
16/01/26 08:56:57 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 byte
s from memory into reduce
16/01/26 08:56:57 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 08:56:57 INFO mapred.Merger: Down to the last merge-pass, with 1
segments left of total size: 194836 bytes
16/01/26 08:56:57 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 08:56:57 INFO streaming.PipeMapRed: PipeMapRed exec [/root/hw
2/./reducer3.py]
16/01/26 08:56:57 INFO Configuration.deprecation: mapred.job.tracker is de
precated. Instead, use mapreduce.jobtracker.address
16/01/26 08:56:57 INFO Configuration.deprecation: mapred.map.tasks is depr
ecated. Instead, use mapreduce.job.maps
16/01/26 08:56:57 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 08:56:57 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 08:56:57 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] o
ut:NA [rec/s]
16/01/26 08:56:57 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s]
out:NA [rec/s]
16/01/26 08:56:57 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s]
out:NA [rec/s]
16/01/26 08:56:57 INFO streaming.PipeMapRed: Records R/W=15336/1
16/01/26 08:56:57 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 08:56:57 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 08:56:58 INFO mapred.Task: Task:attempt_local392316276_0001_r_000
000_0 is done. And is in the process of committing
16/01/26 08:56:58 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 08:56:58 INFO mapred.Task: Task attempt_local392316276_0001_r_000
000_0 is allowed to commit now
16/01/26 08:56:58 INFO output.FileOutputCommitter: Saved output of task 'a
ttempt_local392316276_0001_r_000000_0' to hdfs://localhost:54310/user/roo
t/wk2/hw24/output_1/_temporary/0/task_local392316276_0001_r_000000
16/01/26 08:56:58 INFO mapred.LocalJobRunner: Records R/W=15336/1 > reduce
16/01/26 08:56:58 INFO mapred.Task: Task 'attempt_local392316276_0001_r_00
```

```
0000_0' done.
16/01/26 08:56:58 INFO mapred.LocalJobRunner: Finishing task: attempt_loca
l392316276_0001_r_000000_0
16/01/26 08:56:58 INFO mapred.LocalJobRunner: reduce task executor complet
e.
16/01/26 08:56:58 INFO mapreduce.Job:  map 100% reduce 100%
16/01/26 08:56:58 INFO mapreduce.Job: Job job_local392316276_0001 complete
d successfully
16/01/26 08:56:58 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=601568
                FILE: Number of bytes written=1350020
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=407908
                HDFS: Number of bytes written=155887
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=100
                Map output records=15336
                Map output bytes=164166
                Map output materialized bytes=194844
                Input split bytes=117
                Combine input records=0
                Combine output records=0
                Reduce input groups=5741
                Reduce shuffle bytes=194844
                Reduce input records=15336
                Reduce output records=5741
                Spilled Records=30672
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=49
                Total committed heap usage (bytes)=1013448704
        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=203954
        File Output Format Counters
                Bytes Written=155887
16/01/26 08:56:58 INFO streaming.StreamJob: Output directory: /user/root/w
k2/hw24/output_1
```

In [285]:

```
!rm part*

# Copy the mapper output to local directory
!hdfs dfs -copyToLocal /user/root/wk2/hw24/output_1/part*
!mv part-00000 hw23out.txt
```

```
rm: cannot remove 'part*': No such file or directory
16/01/26 08:57:36 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 08:57:37 WARN hdfs.DFSClient: DFSInputStream has been closed alre
ady
```

In [286]:

```
# Delete output folder if exists
!hdfs dfs -rm -r /user/root/wk2/hw24/output_2

# Run Hadoop Streaming job
!hadoop jar hadoop-streaming-2.7.1.jar \
-mapper mapper2.py \
-reducer reducer2.py \
-input /user/root/wk2/hw24/input \
-output /user/root/wk2/hw24/output_2
```

```
16/01/26 08:58:23 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
rm: `/user/root/wk2/hw24/output_2': No such file or directory
16/01/26 08:58:26 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 08:58:27 INFO Configuration.deprecation: session.id is deprecate
d. Instead, use dfs.metrics.session-id
16/01/26 08:58:27 INFO jvm.JvmMetrics: Initializing JVM Metrics with proce
ssName=JobTracker, sessionId=
16/01/26 08:58:27 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with
processName=JobTracker, sessionId= - already initialized
16/01/26 08:58:28 INFO mapred.FileInputFormat: Total input paths to proces
s : 1
16/01/26 08:58:28 INFO mapreduce.JobSubmitter: number of splits:1
16/01/26 08:58:28 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_local214088061_0001
16/01/26 08:58:28 INFO mapreduce.Job: The url to track the job: http://loc
alhost:8080/
16/01/26 08:58:28 INFO mapreduce.Job: Running job: job_local214088061_0001
16/01/26 08:58:28 INFO mapred.LocalJobRunner: OutputCommitter set in confi
g null
16/01/26 08:58:28 INFO mapred.LocalJobRunner: OutputCommitter is org.apach
e.hadoop.mapred.FileOutputCommitter
16/01/26 08:58:28 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 08:58:28 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/26 08:58:28 INFO mapred.LocalJobRunner: Starting task: attempt_local
214088061_0001_m_000000_0
16/01/26 08:58:28 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 08:58:28 INFO mapred.Task:  Using ResourceCalculatorProcessTree :
[ ]
16/01/26 08:58:28 INFO mapred.MapTask: Processing split: hdfs://localhos
t:54310/user/root/wk2/hw24/input/enronemail_1h.txt:0+203954
16/01/26 08:58:28 INFO mapred.MapTask: numReduceTasks: 1
16/01/26 08:58:28 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/26 08:58:28 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/26 08:58:28 INFO mapred.MapTask: soft limit at 83886080
16/01/26 08:58:28 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/26 08:58:28 INFO mapred.MapTask: kvstart = 26214396; length = 655360
0
16/01/26 08:58:28 INFO mapred.MapTask: Map output collector class = org.ap
ache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/26 08:58:28 INFO streaming.PipeMapRed: PipeMapRed exec [/root/hw
2/./mapper2.py]
16/01/26 08:58:28 INFO Configuration.deprecation: mapred.task.id is deprec
ated. Instead, use mapreduce.task.attempt.id
16/01/26 08:58:28 INFO Configuration.deprecation: user.name is deprecated.
Instead, use mapreduce.job.user.name
16/01/26 08:58:28 INFO Configuration.deprecation: map.input.start is depre
cated. Instead, use mapreduce.map.input.start
16/01/26 08:58:28 INFO Configuration.deprecation: mapred.task.is.map is de
precated. Instead, use mapreduce.task.ismap
16/01/26 08:58:28 INFO Configuration.deprecation: mapred.tip.id is depreca
ted. Instead, use mapreduce.task.id
```

```
16/01/26 08:58:29 INFO Configuration.deprecation: mapred.skip.on is deprec
ated. Instead, use mapreduce.job.skiprecords
16/01/26 08:58:29 INFO Configuration.deprecation: mapred.task.partition is
deprecated. Instead, use mapreduce.task.partition
16/01/26 08:58:29 INFO Configuration.deprecation: map.input.length is depr
ecated. Instead, use mapreduce.map.input.length
16/01/26 08:58:29 INFO Configuration.deprecation: mapred.local.dir is depr
ecated. Instead, use mapreduce.cluster.local.dir
16/01/26 08:58:29 INFO Configuration.deprecation: mapred.work.output.dir i
s deprecated. Instead, use mapreduce.task.output.dir
16/01/26 08:58:29 INFO Configuration.deprecation: map.input.file is deprec
ated. Instead, use mapreduce.map.input.file
16/01/26 08:58:29 INFO Configuration.deprecation: mapred.job.id is depreca
ted. Instead, use mapreduce.job.id
16/01/26 08:58:29 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 08:58:29 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 08:58:29 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] o
ut:NA [rec/s]
16/01/26 08:58:29 INFO streaming.PipeMapRed: Records R/W=100/1
16/01/26 08:58:29 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 08:58:29 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 08:58:29 INFO mapred.LocalJobRunner:
16/01/26 08:58:29 INFO mapred.MapTask: Starting flush of map output
16/01/26 08:58:29 INFO mapred.MapTask: Spilling map output
16/01/26 08:58:29 INFO mapred.MapTask: bufstart = 0; bufend = 5277; bufvoi
d = 104857600
16/01/26 08:58:29 INFO mapred.MapTask: kvstart = 26214396(104857584); kven
d = 26214000(104856000); length = 397/6553600
16/01/26 08:58:29 INFO mapred.MapTask: Finished spill 0
16/01/26 08:58:29 INFO mapred.Task: Task:attempt_local214088061_0001_m_000
000_0 is done. And is in the process of committing
16/01/26 08:58:29 INFO mapred.LocalJobRunner: Records R/W=100/1
16/01/26 08:58:29 INFO mapred.Task: Task 'attempt_local214088061_0001_m_00
0000_0' done.
16/01/26 08:58:29 INFO mapred.LocalJobRunner: Finishing task: attempt_loca
l214088061_0001_m_000000_0
16/01/26 08:58:29 INFO mapred.LocalJobRunner: map task executor complete.
16/01/26 08:58:29 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/26 08:58:29 INFO mapred.LocalJobRunner: Starting task: attempt_local
214088061_0001_r_000000_0
16/01/26 08:58:29 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 08:58:29 INFO mapred.Task:  Using ResourceCalculatorProcessTree :
[ ]
16/01/26 08:58:29 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: or
g.apache.hadoop.mapreduce.task.reduce.Shuffle@2956842e
16/01/26 08:58:29 INFO reduce.MergeManagerImpl: MergerManager: memoryLimi
t=353422528, maxSingleShuffleLimit=88355632, mergeThreshold=233258880, ioS
ortFactor=10, memToMemMergeOutputsThreshold=10
16/01/26 08:58:29 INFO reduce.EventFetcher: attempt_local214088061_000
1_r_000000_0 Thread started: EventFetcher for fetching Map Completion Even
ts
16/01/26 08:58:29 INFO reduce.LocalFetcher: localfetcher#1 about to shuffl
e output of map attempt_local214088061_0001_m_000000_0 decomp: 5479 len: 5
```

```
483 to MEMORY
16/01/26 08:58:29 INFO reduce.InMemoryMapOutput: Read 5479 bytes from map-
output for attempt_local214088061_0001_m_000000_0
16/01/26 08:58:29 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-o
utput of size: 5479, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, us
edMemory ->5479
16/01/26 08:58:29 INFO reduce.EventFetcher: EventFetcher is interrupted..
Returning
16/01/26 08:58:29 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 08:58:29 INFO reduce.MergeManagerImpl: finalMerge called with 1 i
n-memory map-outputs and 0 on-disk map-outputs
16/01/26 08:58:29 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 08:58:29 INFO mapred.Merger: Down to the last merge-pass, with 1
segments left of total size: 5454 bytes
16/01/26 08:58:29 INFO reduce.MergeManagerImpl: Merged 1 segments, 5479 by
tes to disk to satisfy reduce memory limit
16/01/26 08:58:29 INFO reduce.MergeManagerImpl: Merging 1 files, 5483 byte
s from disk
16/01/26 08:58:29 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 byte
s from memory into reduce
16/01/26 08:58:29 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 08:58:29 INFO mapred.Merger: Down to the last merge-pass, with 1
segments left of total size: 5454 bytes
16/01/26 08:58:29 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 08:58:29 INFO streaming.PipeMapRed: PipeMapRed exec [/root/hw
2/./reducer2.py]
16/01/26 08:58:29 INFO Configuration.deprecation: mapred.job.tracker is de
precated. Instead, use mapreduce.jobtracker.address
16/01/26 08:58:29 INFO Configuration.deprecation: mapred.map.tasks is depr
ecated. Instead, use mapreduce.job.maps
16/01/26 08:58:29 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 08:58:29 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 08:58:29 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] o
ut:NA [rec/s]
16/01/26 08:58:29 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 08:58:29 INFO streaming.PipeMapRed: Records R/W=100/1
16/01/26 08:58:29 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 08:58:29 INFO mapred.Task: Task:attempt_local214088061_0001_r_000
000_0 is done. And is in the process of committing
16/01/26 08:58:29 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 08:58:29 INFO mapred.Task: Task attempt_local214088061_0001_r_000
000_0 is allowed to commit now
16/01/26 08:58:29 INFO mapreduce.Job: Job job_local214088061_0001 running
in uber mode : false
16/01/26 08:58:29 INFO mapreduce.Job:  map 100% reduce 0%
16/01/26 08:58:29 INFO output.FileOutputCommitter: Saved output of task 'a
ttempt_local214088061_0001_r_000000_0' to hdfs://localhost:54310/user/roo
t/wk2/hw24/output_2/_temporary/0/task_local214088061_0001_r_000000
16/01/26 08:58:29 INFO mapred.LocalJobRunner: Records R/W=100/1 > reduce
16/01/26 08:58:29 INFO mapred.Task: Task 'attempt_local214088061_0001_r_00
0000_0' done.
16/01/26 08:58:29 INFO mapred.LocalJobRunner: Finishing task: attempt_loca
l214088061_0001_r_000000_0
16/01/26 08:58:29 INFO mapred.LocalJobRunner: reduce task executor complet
```

```
e.
16/01/26 08:58:30 INFO mapreduce.Job:  map 100% reduce 100%
16/01/26 08:58:30 INFO mapreduce.Job: Job job_local214088061_0001 complete
d successfully
16/01/26 08:58:30 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=222846
                FILE: Number of bytes written=781937
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=407908
                HDFS: Number of bytes written=5517
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=100
                Map output records=100
                Map output bytes=5277
                Map output materialized bytes=5483
                Input split bytes=117
                Combine input records=0
                Combine output records=0
                Reduce input groups=100
                Reduce shuffle bytes=5483
                Reduce input records=100
                Reduce output records=202
                Spilled Records=200
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=34
                Total committed heap usage (bytes)=1013448704
        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=203954
        File Output Format Counters
                Bytes Written=5517
16/01/26 08:58:30 INFO streaming.StreamJob: Output directory: /user/root/w
k2/hw24/output_2
```

In [288]:

```
!hdfs dfs -cat /user/root/wk2/hw24/output_2/part-00000|tail -3
```

```
16/01/26 09:00:46 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable

Training error: 0.0%
Accuracy: 100.0%
```

HW 2.5.

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:

We will reuse HW2.3 mappers and reducers except reducer1.py.

In [305]:

```
%%writefile reducer4.py
#!/usr/bin/python
## reducer4.py
## Author: Prabhakar Gundugola
## Description: reducer4 code for HW2.5

import sys
import math

spam_email_count = 0
ham_email_count = 0

spam_word_count = 0
ham_word_count = 0

total_spam_count = 0
total_ham_count = 0
total_word_count = 0

total_docs = 0

spam_words_freq = {}
ham_words_freq = {}

priors_calc = 0

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

    if tokens[0] == "0000000DOC_CLASS":
        spam_email_count = int(tokens[1])
        ham_email_count = int(tokens[2])
        continue

    try:
        count = int(tokens[2])
        spam = int(tokens[1])
    except ValueError:
        continue


    word, spam, frequency = tokens[0], spam, count

    if not(word in spam_words_freq or word in ham_words_freq):
        total_word_count += 1


    if spam == 1:
        total_spam_count += frequency
        if word in spam_words_freq:
            spam_words_freq[word] += frequency
        else:
            spam_words_freq[word] = frequency

        if word not in ham_words_freq:
            ham_words_freq[word] = 0
    else:
        total_ham_count += frequency
        if word in ham_words_freq:
            ham_words_freq[word] += frequency
        else:
            ham_words_freq[word] = frequency

        if word not in spam_words_freq:
            spam_words_freq[word] = 0

prior_spam = math.log(1.0*spam_email_count/(spam_email_count + ham_email_co
unt))
prior_ham = math.log(1.0*ham_email_count/(spam_email_count + ham_email_coun
t))
print('{0}\t{1}\t{2}'.format("0000000PRIORS", str(prior_spam), str(prior_ha
m)))


# Calculate conditional probability.
# Applied Laplace smoothing to math.log function in the numerator
prob_spam_words = {}
prob_ham_words = {}

for word in ham_words_freq:
    ham_count = ham_words_freq[word]
    spam_count = spam_words_freq[word]
    if (ham_count+spam_count < 3 ):
        total_word_count -= 1
        continue
    if spam_words_freq[word] > 0:
        prob_spam_words[word] = math.log((1.0)*(spam_words_freq[word]+1)/to
tal_spam_count + total_word_count)
    else:
        prob_spam_words[word] = 0
    if ham_words_freq[word] > 0:
        prob_ham_words[word] = math.log((1.0)*(ham_words_freq[word]+1)/tota
l_ham_count + total_word_count)
    else:
        prob_ham_words[word] = 0

    print('{0}\t{1}\t{2}'.format(word, str(prob_spam_words[word]), str(pro
b_ham_words[word])))
```

```
Overwriting reducer4.py
```

In [306]:

```
# Ensure the input folder doesn't exist
!hdfs dfs -rm -r /user/root/wk2/hw25
!hdfs dfs -mkdir -p /user/root/wk2/hw25/input

!hdfs dfs -put enronemail_1h.txt /user/root/wk2/hw25/input
```

```
16/01/26 09:42:30 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 09:42:31 INFO fs.TrashPolicyDefault: Namenode trash configuratio
n: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/root/wk2/hw25
16/01/26 09:42:33 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 09:42:36 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
```

In [307]:

```
# Delete output folder if exists
!hdfs dfs -rm -r /user/root/wk2/hw25/output_1

# Run Hadoop Streaming job
!hadoop jar hadoop-streaming-2.7.1.jar \
-mapper mapper1.py \
-reducer reducer4.py \
-input /user/root/wk2/hw25/input \
-output /user/root/wk2/hw25/output_1
```

```
16/01/26 09:42:39 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
rm: `/user/root/wk2/hw25/output_1': No such file or directory
16/01/26 09:42:42 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 09:42:43 INFO Configuration.deprecation: session.id is deprecate
d. Instead, use dfs.metrics.session-id
16/01/26 09:42:43 INFO jvm.JvmMetrics: Initializing JVM Metrics with proce
ssName=JobTracker, sessionId=
16/01/26 09:42:43 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with
processName=JobTracker, sessionId= - already initialized
16/01/26 09:42:43 INFO mapred.FileInputFormat: Total input paths to proces
s : 1
16/01/26 09:42:43 INFO mapreduce.JobSubmitter: number of splits:1
16/01/26 09:42:43 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_local1511549212_0001
16/01/26 09:42:43 INFO mapreduce.Job: The url to track the job: http://loc
alhost:8080/
16/01/26 09:42:43 INFO mapreduce.Job: Running job: job_local1511549212_000
1
16/01/26 09:42:43 INFO mapred.LocalJobRunner: OutputCommitter set in confi
g null
16/01/26 09:42:43 INFO mapred.LocalJobRunner: OutputCommitter is org.apach
e.hadoop.mapred.FileOutputCommitter
16/01/26 09:42:43 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 09:42:44 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/26 09:42:44 INFO mapred.LocalJobRunner: Starting task: attempt_local
1511549212_0001_m_000000_0
16/01/26 09:42:44 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 09:42:44 INFO mapred.Task:  Using ResourceCalculatorProcessTree :
[ ]
16/01/26 09:42:44 INFO mapred.MapTask: Processing split: hdfs://localhos
t:54310/user/root/wk2/hw25/input/enronemail_1h.txt:0+203954
16/01/26 09:42:44 INFO mapred.MapTask: numReduceTasks: 1
16/01/26 09:42:44 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/26 09:42:44 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/26 09:42:44 INFO mapred.MapTask: soft limit at 83886080
16/01/26 09:42:44 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/26 09:42:44 INFO mapred.MapTask: kvstart = 26214396; length = 655360
0
16/01/26 09:42:44 INFO mapred.MapTask: Map output collector class = org.ap
ache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/26 09:42:44 INFO streaming.PipeMapRed: PipeMapRed exec [/root/hw
2/./mapper1.py]
16/01/26 09:42:44 INFO Configuration.deprecation: mapred.task.id is deprec
ated. Instead, use mapreduce.task.attempt.id
16/01/26 09:42:44 INFO Configuration.deprecation: user.name is deprecated.
Instead, use mapreduce.job.user.name
16/01/26 09:42:44 INFO Configuration.deprecation: map.input.start is depre
cated. Instead, use mapreduce.map.input.start
16/01/26 09:42:44 INFO Configuration.deprecation: mapred.task.is.map is de
precated. Instead, use mapreduce.task.ismap
16/01/26 09:42:44 INFO Configuration.deprecation: mapred.tip.id is depreca
```

```
ted. Instead, use mapreduce.task.id
16/01/26 09:42:44 INFO Configuration.deprecation: mapred.skip.on is deprec
ated. Instead, use mapreduce.job.skiprecords
16/01/26 09:42:44 INFO Configuration.deprecation: mapred.task.partition is
deprecated. Instead, use mapreduce.task.partition
16/01/26 09:42:44 INFO Configuration.deprecation: map.input.length is depr
ecated. Instead, use mapreduce.map.input.length
16/01/26 09:42:44 INFO Configuration.deprecation: mapred.local.dir is depr
ecated. Instead, use mapreduce.cluster.local.dir
16/01/26 09:42:44 INFO Configuration.deprecation: mapred.work.output.dir i
s deprecated. Instead, use mapreduce.task.output.dir
16/01/26 09:42:44 INFO Configuration.deprecation: map.input.file is deprec
ated. Instead, use mapreduce.map.input.file
16/01/26 09:42:44 INFO Configuration.deprecation: mapred.job.id is depreca
ted. Instead, use mapreduce.job.id
16/01/26 09:42:44 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 09:42:44 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 09:42:44 INFO streaming.PipeMapRed: Records R/W=72/1
16/01/26 09:42:44 INFO streaming.PipeMapRed: R/W/S=100/2196/0 in:NA [re
c/s] out:NA [rec/s]
16/01/26 09:42:44 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 09:42:44 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 09:42:44 INFO mapred.LocalJobRunner:
16/01/26 09:42:44 INFO mapred.MapTask: Starting flush of map output
16/01/26 09:42:44 INFO mapred.MapTask: Spilling map output
16/01/26 09:42:44 INFO mapred.MapTask: bufstart = 0; bufend = 164166; bufv
oid = 104857600
16/01/26 09:42:44 INFO mapred.MapTask: kvstart = 26214396(104857584); kven
d = 26153056(104612224); length = 61341/6553600
16/01/26 09:42:44 INFO mapred.MapTask: Finished spill 0
16/01/26 09:42:44 INFO mapreduce.Job: Job job_local1511549212_0001 running
in uber mode : false
16/01/26 09:42:44 INFO mapreduce.Job:  map 0% reduce 0%
16/01/26 09:42:44 INFO mapred.Task: Task:attempt_local1511549212_0001_m_00
0000_0 is done. And is in the process of committing
16/01/26 09:42:44 INFO mapred.LocalJobRunner: Records R/W=72/1
16/01/26 09:42:44 INFO mapred.Task: Task 'attempt_local1511549212_0001_m_0
00000_0' done.
16/01/26 09:42:44 INFO mapred.LocalJobRunner: Finishing task: attempt_loca
l1511549212_0001_m_000000_0
16/01/26 09:42:44 INFO mapred.LocalJobRunner: map task executor complete.
16/01/26 09:42:44 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/26 09:42:44 INFO mapred.LocalJobRunner: Starting task: attempt_local
1511549212_0001_r_000000_0
16/01/26 09:42:44 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 09:42:44 INFO mapred.Task:  Using ResourceCalculatorProcessTree :
[ ]
16/01/26 09:42:44 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: or
g.apache.hadoop.mapreduce.task.reduce.Shuffle@2711dfc7
16/01/26 09:42:45 INFO reduce.MergeManagerImpl: MergerManager: memoryLimi
t=353422528, maxSingleShuffleLimit=88355632, mergeThreshold=233258880, ioS
ortFactor=10, memToMemMergeOutputsThreshold=10
16/01/26 09:42:45 INFO reduce.EventFetcher: attempt_local1511549212_000
```

```
1_r_000000_0 Thread started: EventFetcher for fetching Map Completion Even
ts
16/01/26 09:42:45 INFO reduce.LocalFetcher: localfetcher#1 about to shuffl
e output of map attempt_local1511549212_0001_m_000000_0 decomp: 194840 le
n: 194844 to MEMORY
16/01/26 09:42:45 INFO reduce.InMemoryMapOutput: Read 194840 bytes from ma
p-output for attempt_local1511549212_0001_m_000000_0
16/01/26 09:42:45 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-o
utput of size: 194840, inMemoryMapOutputs.size() -> 1, commitMemory -> 0,
usedMemory ->194840
16/01/26 09:42:45 INFO reduce.EventFetcher: EventFetcher is interrupted..
Returning
16/01/26 09:42:45 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 09:42:45 INFO reduce.MergeManagerImpl: finalMerge called with 1 i
n-memory map-outputs and 0 on-disk map-outputs
16/01/26 09:42:45 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 09:42:45 INFO mapred.Merger: Down to the last merge-pass, with 1
segments left of total size: 194836 bytes
16/01/26 09:42:45 INFO reduce.MergeManagerImpl: Merged 1 segments, 194840
bytes to disk to satisfy reduce memory limit
16/01/26 09:42:45 INFO reduce.MergeManagerImpl: Merging 1 files, 194844 by
tes from disk
16/01/26 09:42:45 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 byte
s from memory into reduce
16/01/26 09:42:45 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 09:42:45 INFO mapred.Merger: Down to the last merge-pass, with 1
segments left of total size: 194836 bytes
16/01/26 09:42:45 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 09:42:45 INFO streaming.PipeMapRed: PipeMapRed exec [/root/hw
2/./reducer4.py]
16/01/26 09:42:45 INFO Configuration.deprecation: mapred.job.tracker is de
precated. Instead, use mapreduce.jobtracker.address
16/01/26 09:42:45 INFO Configuration.deprecation: mapred.map.tasks is depr
ecated. Instead, use mapreduce.job.maps
16/01/26 09:42:45 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 09:42:45 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 09:42:45 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] o
ut:NA [rec/s]
16/01/26 09:42:45 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s]
out:NA [rec/s]
16/01/26 09:42:45 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s]
out:NA [rec/s]
16/01/26 09:42:45 INFO streaming.PipeMapRed: Records R/W=15336/1
16/01/26 09:42:45 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 09:42:45 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 09:42:45 INFO mapred.Task: Task:attempt_local1511549212_0001_r_00
0000_0 is done. And is in the process of committing
16/01/26 09:42:45 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 09:42:45 INFO mapred.Task: Task attempt_local1511549212_0001_r_00
0000_0 is allowed to commit now
16/01/26 09:42:45 INFO output.FileOutputCommitter: Saved output of task 'a
ttempt_local1511549212_0001_r_000000_0' to hdfs://localhost:54310/user/roo
t/wk2/hw25/output_1/_temporary/0/task_local1511549212_0001_r_000000
16/01/26 09:42:45 INFO mapred.LocalJobRunner: Records R/W=15336/1 > reduce
```

```
16/01/26 09:42:45 INFO mapred.Task: Task 'attempt_local1511549212_0001_r_0
00000_0' done.
16/01/26 09:42:45 INFO mapred.LocalJobRunner: Finishing task: attempt_loca
l1511549212_0001_r_000000_0
16/01/26 09:42:45 INFO mapred.LocalJobRunner: reduce task executor complet
e.
16/01/26 09:42:45 INFO mapreduce.Job:  map 100% reduce 100%
16/01/26 09:42:45 INFO mapreduce.Job: Job job_local1511549212_0001 complet
ed successfully
16/01/26 09:42:45 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=601568
                FILE: Number of bytes written=1353028
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=407908
                HDFS: Number of bytes written=53028
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=100
                Map output records=15336
                Map output bytes=164166
                Map output materialized bytes=194844
                Input split bytes=117
                Combine input records=0
                Combine output records=0
                Reduce input groups=5741
                Reduce shuffle bytes=194844
                Reduce input records=15336
                Reduce output records=1830
                Spilled Records=30672
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=33
                Total committed heap usage (bytes)=1013448704
        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=203954
        File Output Format Counters
                Bytes Written=53028
16/01/26 09:42:45 INFO streaming.StreamJob: Output directory: /user/root/w
k2/hw25/output_1
```

In [308]:

```
!rm hw23out.txt

# Copy the mapper output to local directory
!hdfs dfs -copyToLocal /user/root/wk2/hw25/output_1/part*
!mv part-00000 hw23out.txt
```

```
rm: cannot remove 'hw23out.txt': No such file or directory
16/01/26 09:43:10 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 09:43:11 WARN hdfs.DFSClient: DFSInputStream has been closed alre
ady
```

In [309]:

```
# Delete output folder if exists
!hdfs dfs -rm -r /user/root/wk2/hw25/output_2

# Run Hadoop Streaming job
!hadoop jar hadoop-streaming-2.7.1.jar \
-mapper mapper2.py \
-reducer reducer2.py \
-input /user/root/wk2/hw25/input \
-output /user/root/wk2/hw25/output_2
```

```
16/01/26 09:43:17 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
rm: `/user/root/wk2/hw25/output_2': No such file or directory
16/01/26 09:43:20 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
16/01/26 09:43:21 INFO Configuration.deprecation: session.id is deprecate
d. Instead, use dfs.metrics.session-id
16/01/26 09:43:21 INFO jvm.JvmMetrics: Initializing JVM Metrics with proce
ssName=JobTracker, sessionId=
16/01/26 09:43:21 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with
processName=JobTracker, sessionId= - already initialized
16/01/26 09:43:22 INFO mapred.FileInputFormat: Total input paths to proces
s : 1
16/01/26 09:43:22 INFO mapreduce.JobSubmitter: number of splits:1
16/01/26 09:43:22 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_local1707625718_0001
16/01/26 09:43:22 INFO mapreduce.Job: The url to track the job: http://loc
alhost:8080/
16/01/26 09:43:22 INFO mapreduce.Job: Running job: job_local1707625718_000
1
16/01/26 09:43:22 INFO mapred.LocalJobRunner: OutputCommitter set in confi
g null
16/01/26 09:43:22 INFO mapred.LocalJobRunner: OutputCommitter is org.apach
e.hadoop.mapred.FileOutputCommitter
16/01/26 09:43:22 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 09:43:22 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/26 09:43:22 INFO mapred.LocalJobRunner: Starting task: attempt_local
1707625718_0001_m_000000_0
16/01/26 09:43:22 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 09:43:22 INFO mapred.Task:  Using ResourceCalculatorProcessTree :
[ ]
16/01/26 09:43:22 INFO mapred.MapTask: Processing split: hdfs://localhos
t:54310/user/root/wk2/hw25/input/enronemail_1h.txt:0+203954
16/01/26 09:43:22 INFO mapred.MapTask: numReduceTasks: 1
16/01/26 09:43:22 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/26 09:43:22 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/26 09:43:22 INFO mapred.MapTask: soft limit at 83886080
16/01/26 09:43:22 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/26 09:43:22 INFO mapred.MapTask: kvstart = 26214396; length = 655360
0
16/01/26 09:43:23 INFO mapred.MapTask: Map output collector class = org.ap
ache.hadoop.mapred.MapTask$MapOutputBuffer
16/01/26 09:43:23 INFO streaming.PipeMapRed: PipeMapRed exec [/root/hw
2/./mapper2.py]
16/01/26 09:43:23 INFO Configuration.deprecation: mapred.task.id is deprec
ated. Instead, use mapreduce.task.attempt.id
16/01/26 09:43:23 INFO Configuration.deprecation: user.name is deprecated.
Instead, use mapreduce.job.user.name
16/01/26 09:43:23 INFO Configuration.deprecation: map.input.start is depre
cated. Instead, use mapreduce.map.input.start
16/01/26 09:43:23 INFO Configuration.deprecation: mapred.task.is.map is de
precated. Instead, use mapreduce.task.ismap
16/01/26 09:43:23 INFO Configuration.deprecation: mapred.tip.id is depreca
```

```
ted. Instead, use mapreduce.task.id
16/01/26 09:43:23 INFO Configuration.deprecation: mapred.skip.on is deprec
ated. Instead, use mapreduce.job.skiprecords
16/01/26 09:43:23 INFO Configuration.deprecation: mapred.task.partition is
deprecated. Instead, use mapreduce.task.partition
16/01/26 09:43:23 INFO Configuration.deprecation: map.input.length is depr
ecated. Instead, use mapreduce.map.input.length
16/01/26 09:43:23 INFO Configuration.deprecation: mapred.local.dir is depr
ecated. Instead, use mapreduce.cluster.local.dir
16/01/26 09:43:23 INFO Configuration.deprecation: mapred.work.output.dir i
s deprecated. Instead, use mapreduce.task.output.dir
16/01/26 09:43:23 INFO Configuration.deprecation: map.input.file is deprec
ated. Instead, use mapreduce.map.input.file
16/01/26 09:43:23 INFO Configuration.deprecation: mapred.job.id is depreca
ted. Instead, use mapreduce.job.id
16/01/26 09:43:23 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 09:43:23 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 09:43:23 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] o
ut:NA [rec/s]
16/01/26 09:43:23 INFO streaming.PipeMapRed: Records R/W=100/1
16/01/26 09:43:23 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 09:43:23 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 09:43:23 INFO mapred.LocalJobRunner:
16/01/26 09:43:23 INFO mapred.MapTask: Starting flush of map output
16/01/26 09:43:23 INFO mapred.MapTask: Spilling map output
16/01/26 09:43:23 INFO mapred.MapTask: bufstart = 0; bufend = 5166; bufvoi
d = 104857600
16/01/26 09:43:23 INFO mapred.MapTask: kvstart = 26214396(104857584); kven
d = 26214000(104856000); length = 397/6553600
16/01/26 09:43:23 INFO mapred.MapTask: Finished spill 0
16/01/26 09:43:23 INFO mapred.Task: Task:attempt_local1707625718_0001_m_00
0000_0 is done. And is in the process of committing
16/01/26 09:43:23 INFO mapred.LocalJobRunner: Records R/W=100/1
16/01/26 09:43:23 INFO mapred.Task: Task 'attempt_local1707625718_0001_m_0
00000_0' done.
16/01/26 09:43:23 INFO mapred.LocalJobRunner: Finishing task: attempt_loca
l1707625718_0001_m_000000_0
16/01/26 09:43:23 INFO mapred.LocalJobRunner: map task executor complete.
16/01/26 09:43:23 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/26 09:43:23 INFO mapred.LocalJobRunner: Starting task: attempt_local
1707625718_0001_r_000000_0
16/01/26 09:43:23 INFO output.FileOutputCommitter: File Output Committer A
lgorithm version is 1
16/01/26 09:43:23 INFO mapred.Task:  Using ResourceCalculatorProcessTree :
[ ]
16/01/26 09:43:23 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: or
g.apache.hadoop.mapreduce.task.reduce.Shuffle@564d06b4
16/01/26 09:43:23 INFO reduce.MergeManagerImpl: MergerManager: memoryLimi
t=353422528, maxSingleShuffleLimit=88355632, mergeThreshold=233258880, ioS
ortFactor=10, memToMemMergeOutputsThreshold=10
16/01/26 09:43:23 INFO reduce.EventFetcher: attempt_local1707625718_000
1_r_000000_0 Thread started: EventFetcher for fetching Map Completion Even
ts
16/01/26 09:43:23 INFO reduce.LocalFetcher: localfetcher#1 about to shuffl
```

```
e output of map attempt_local1707625718_0001_m_000000_0 decomp: 5368 len:
5372 to MEMORY
16/01/26 09:43:23 INFO reduce.InMemoryMapOutput: Read 5368 bytes from map-
output for attempt_local1707625718_0001_m_000000_0
16/01/26 09:43:23 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-o
utput of size: 5368, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, us
edMemory ->5368
16/01/26 09:43:23 INFO reduce.EventFetcher: EventFetcher is interrupted..
Returning
16/01/26 09:43:23 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 09:43:23 INFO reduce.MergeManagerImpl: finalMerge called with 1 i
n-memory map-outputs and 0 on-disk map-outputs
16/01/26 09:43:23 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 09:43:23 INFO mapred.Merger: Down to the last merge-pass, with 1
segments left of total size: 5343 bytes
16/01/26 09:43:23 INFO reduce.MergeManagerImpl: Merged 1 segments, 5368 by
tes to disk to satisfy reduce memory limit
16/01/26 09:43:23 INFO reduce.MergeManagerImpl: Merging 1 files, 5372 byte
s from disk
16/01/26 09:43:23 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 byte
s from memory into reduce
16/01/26 09:43:23 INFO mapred.Merger: Merging 1 sorted segments
16/01/26 09:43:23 INFO mapred.Merger: Down to the last merge-pass, with 1
segments left of total size: 5343 bytes
16/01/26 09:43:23 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 09:43:23 INFO streaming.PipeMapRed: PipeMapRed exec [/root/hw
2/./reducer2.py]
16/01/26 09:43:23 INFO Configuration.deprecation: mapred.job.tracker is de
precated. Instead, use mapreduce.jobtracker.address
16/01/26 09:43:23 INFO Configuration.deprecation: mapred.map.tasks is depr
ecated. Instead, use mapreduce.job.maps
16/01/26 09:43:23 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 09:43:23 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] ou
t:NA [rec/s]
16/01/26 09:43:23 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] o
ut:NA [rec/s]
16/01/26 09:43:23 INFO streaming.PipeMapRed: MRErrorThread done
16/01/26 09:43:23 INFO streaming.PipeMapRed: Records R/W=100/1
16/01/26 09:43:23 INFO streaming.PipeMapRed: mapRedFinished
16/01/26 09:43:23 INFO mapreduce.Job: Job job_local1707625718_0001 running
in uber mode : false
16/01/26 09:43:23 INFO mapreduce.Job:  map 100% reduce 0%
16/01/26 09:43:23 INFO mapred.Task: Task:attempt_local1707625718_0001_r_00
0000_0 is done. And is in the process of committing
16/01/26 09:43:23 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/26 09:43:23 INFO mapred.Task: Task attempt_local1707625718_0001_r_00
0000_0 is allowed to commit now
16/01/26 09:43:23 INFO output.FileOutputCommitter: Saved output of task 'a
ttempt_local1707625718_0001_r_000000_0' to hdfs://localhost:54310/user/roo
t/wk2/hw25/output_2/_temporary/0/task_local1707625718_0001_r_000000
16/01/26 09:43:23 INFO mapred.LocalJobRunner: Records R/W=100/1 > reduce
16/01/26 09:43:23 INFO mapred.Task: Task 'attempt_local1707625718_0001_r_0
00000_0' done.
16/01/26 09:43:23 INFO mapred.LocalJobRunner: Finishing task: attempt_loca
l1707625718_0001_r_000000_0
```

```
16/01/26 09:43:23 INFO mapred.LocalJobRunner: reduce task executor complet
e.
16/01/26 09:43:24 INFO mapreduce.Job:  map 100% reduce 100%
16/01/26 09:43:24 INFO mapreduce.Job: Job job_local1707625718_0001 complet
ed successfully
16/01/26 09:43:24 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=222624
                FILE: Number of bytes written=784612
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=407908
                HDFS: Number of bytes written=5405
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=100
                Map output records=100
                Map output bytes=5166
                Map output materialized bytes=5372
                Input split bytes=117
                Combine input records=0
                Combine output records=0
                Reduce input groups=100
                Reduce shuffle bytes=5372
                Reduce input records=100
                Reduce output records=202
                Spilled Records=200
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=28
                Total committed heap usage (bytes)=1013448704
        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=203954
        File Output Format Counters
                Bytes Written=5405
16/01/26 09:43:24 INFO streaming.StreamJob: Output directory: /user/root/w
k2/hw25/output_2
```

```
In [310]: !hdfs dfs -cat /user/root/wk2/hw25/output_2/part-00000|tail -3
```

```
16/01/26 09:43:38 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable

Training error: 1.0%
Accuracy: 99.0%
```

The training error is 1.0% when we ignored the words with frequency less than 3

HW 2.6. Benchmark your code with the Python SciKit-Learn implementation of the multinomial Naive Bayes algorithm

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

In this exercise, please complete the following:

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

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

In [300]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB, BernoulliNB
import string

trainX = []
trainY = []

# Take only the subject and body here to simulate the mapper code
with open('enronemail_1h.txt', 'r') as myfile:
    for line in myfile.readlines():
        try:
            # Tokenize each line. Line format - DOC_ID <tab> SPAM <tab> sub
ject <tab> body
            tokens = line.lower().strip().split('\t')
            label = tokens[1]

            # Concatenate subject and body fields and store it in word_stri
ng
            word_string = tokens[2].strip() + ' ' + tokens[3].strip()

            # Remove punctuation
            word_string = word_string.translate(string.maketrans("", ""),
                                                string.punctuation)

            trainX.append(word_string)
        except ValueError:
            email_id, label, body = line.split('\t')
            X_train.append(body)
        # extract only words from the combined subject and body text
        trainY.append(int(label))

# Use the TfidVectorizer to create the feature vectors
# We should override the tokenizer regular expression to make it the same a
s what we used
# in our poor man's mapper code
vectorizer = TfidfVectorizer(token_pattern = "[\w']+")
vf = vectorizer.fit(trainX,trainY)

clf = MultinomialNB()
clf.fit(vf.fit_transform(trainX), trainY)
print "Multinomial Bayes Training Error: ", 1.0 - clf.score(vf.fit_transfor
m(trainX), trainY)
```

```
Multinomial Bayes Training Error:  0.0
```

HW 2.6 - Summary of Results

Classification Methodology	Training error
scikit-learn MultinomialNB	0%
Multinomial NB MapReduce Implementation	0%