MIDS-W261-2015-HWK-Week02-Sakhamuri

January 30, 2016

0.1 Name : Vamsi Sakhamuri

0.2 E-mail : vamsi@ischool.berkeley.edu

0.3 Class Name : W261-3

0.4 Week Number : 2

0.5 Date of submission : 01/26/2016

HW2.0.
What is a race condition in the context of parallel computation? Give an example. What is MapReduce? How does it differ from Hadoop? Which programming paradigm is Hadoop based on? Explain and give a simple example in code and show the code running.

- A race condition , in the context of parallel computation, can mean that the results of our computation will not be deterministic. It can occur when different threads attempt to modify a shared resource.

  For example, if two threads want to multiply a variable by 2 and assign the result back to the variable.

  So , if thread A computes X = X * 2

  and thread B also computes X = X * 2

  Assuming the initial value of X is 1, then there are two results possible.

  1st Possibility : Both thread A and thread B read the previous value of X. In this case, the new value of X will be 2 after both the threads finish their computation.

  2nd Possibility : Thread A reads the previous value of X and writes back the result to X (so X=2). Thread B reads this updated value of X. And so the new value of X will be 4 after both threads finish their computation.

  Since the result is not deterministic (2 vs 4), there is a race condition. This can be eliminated by the use of mutex's where the threads need to acquire a lock on the shared resource before proceeding their respective computation.

- Map reduce is a programming model which provides an abstraction to the user by hiding away the system level details from the programmer. The programmer does not have to worry about setting up barriers or worry about race conditions or deadlocks and can solely focus on the job of writing the mappers,reducers,combiners and partitioners.

- Hadoop is a framework whereas mapreduce is a programming model. Hadoop allows for the distributed processing of large-scale datasets over a cluster of commodity servers via a programming model, where the programming model is map-reduce.

A simple map-reduce program using hadoop which does a word count is shown below.

In [198]:
```
%%writefile mapper.py
#!/usr/bin/python
import sys
```

```
import re
# input comes from STDIN (standard input)
for line in sys.stdin:
    word = re.split(r'[\s+]',line)
    for w in word:
        print w,1   #emit word,1
```

```
Overwriting mapper.py
```

```
In [199]: !chmod a+x mapper.py
```

```
In [200]: %%writefile reducer.py
#!/usr/bin/python
import re
import sys

prev_word = None
count = 0
# input comes from STDIN
for line in sys.stdin:
    w_t = re.split(r'[\s+]',line)
    if(prev_word !=None):
        if(prev_word !=w_t[0]):
            print prev_word,count
            count = 0
    count = count+1
    prev_word = w_t[0]
print prev_word,count
```

```
Overwriting reducer.py
```

```
In [201]: !chmod a+x reducer.py
```

```
In [202]: !echo "hello hi hey hello hi hi hello hi hi hi hey" > simple_text.txt
```

```
In [203]: !/Users/Vamsi/Downloads/hadoop-2.7.1/sbin/start-dfs.sh
```

```
16/01/30 03:45:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
Starting namenodes on [localhost]
localhost: namenode running as process 1615. Stop it first.
localhost: datanode running as process 1714. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 1353. Stop it first.
16/01/30 03:45:40 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

```
In [204]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -rm -r -f  /user/vamsi/hw2
```

```
16/01/30 03:45:42 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/01/30 03:45:43 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minut
Deleted /user/vamsi/hw2
```

```
In [205]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -mkdir -p /user/vamsi/hw2
```

```
16/01/30 03:45:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

```
In [206]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -put simple_text.txt /user/vamsi/hw2
```

```
16/01/30 03:45:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

```
In [207]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hadoop jar /Users/Vamsi/Downloads/hadoop-2.7.1/bin/ha
          -D mapred.reduce.tasks=1 \
          -mapper mapper.py \
          -reducer reducer.py \
          -input /user/vamsi/hw2/simple_text.txt \
          -output /user/vamsi/hw2/output_2_0
```

```
16/01/30 03:45:49 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/01/30 03:45:49 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.ses
16/01/30 03:45:49 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/01/30 03:45:49 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessi
16/01/30 03:45:50 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/30 03:45:50 INFO mapreduce.JobSubmitter: number of splits:1
16/01/30 03:45:50 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapred
16/01/30 03:45:50 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1399621981_0001
16/01/30 03:45:50 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/30 03:45:50 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/30 03:45:50 INFO mapreduce.Job: Running job: job_local1399621981_0001
16/01/30 03:45:50 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
16/01/30 03:45:50 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/30 03:45:50 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/30 03:45:50 INFO mapred.LocalJobRunner: Starting task: attempt_local1399621981_0001_m_000000_0
16/01/30 03:45:50 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/30 03:45:50 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/01/30 03:45:50 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/30 03:45:50 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/vamsi/hw2/simple_tex
16/01/30 03:45:50 INFO mapred.MapTask: numReduceTasks: 1
16/01/30 03:45:50 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/30 03:45:50 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/30 03:45:50 INFO mapred.MapTask: soft limit at 83886080
16/01/30 03:45:50 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/30 03:45:50 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/30 03:45:50 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Map
16/01/30 03:45:50 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw2/./mapper.p
16/01/30 03:45:51 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.ta
16/01/30 03:45:51 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce
16/01/30 03:45:51 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.m
16/01/30 03:45:51 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.j
16/01/30 03:45:51 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce
16/01/30 03:45:51 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use map
16/01/30 03:45:51 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce
16/01/30 03:45:51 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.jo
16/01/30 03:45:51 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.us
16/01/30 03:45:51 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapredu
16/01/30 03:45:51 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.t
16/01/30 03:45:51 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapr
16/01/30 03:45:51 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:45:51 INFO streaming.PipeMapRed: Records R/W=1/1
16/01/30 03:45:51 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 03:45:51 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 03:45:51 INFO mapred.LocalJobRunner:
16/01/30 03:45:51 INFO mapred.MapTask: Starting flush of map output
16/01/30 03:45:51 INFO mapred.MapTask: Spilling map output
```

```
16/01/30 03:45:51 INFO mapred.MapTask: bufstart = 0; bufend = 81; bufvoid = 104857600
16/01/30 03:45:51 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214352(104857408); leng
16/01/30 03:45:51 INFO mapred.MapTask: Finished spill 0
16/01/30 03:45:51 INFO mapred.Task: Task:attempt_local1399621981_0001_m_000000_0 is done. And is in the p
16/01/30 03:45:51 INFO mapred.LocalJobRunner: Records R/W=1/1
16/01/30 03:45:51 INFO mapred.Task: Task 'attempt_local1399621981_0001_m_000000_0' done.
16/01/30 03:45:51 INFO mapred.LocalJobRunner: Finishing task: attempt_local1399621981_0001_m_000000_0
16/01/30 03:45:51 INFO mapred.LocalJobRunner: map task executor complete.
16/01/30 03:45:51 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/30 03:45:51 INFO mapred.LocalJobRunner: Starting task: attempt_local1399621981_0001_r_000000_0
16/01/30 03:45:51 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/30 03:45:51 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/01/30 03:45:51 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/30 03:45:51 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/01/30 03:45:51 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleL
16/01/30 03:45:51 INFO reduce.EventFetcher: attempt_local1399621981_0001_r_000000_0 Thread started: Event
16/01/30 03:45:51 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1
16/01/30 03:45:51 INFO reduce.InMemoryMapOutput: Read 107 bytes from map-output for attempt_local139962
16/01/30 03:45:51 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 107, inMemoryMa
16/01/30 03:45:51 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/30 03:45:51 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 03:45:51 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on
16/01/30 03:45:51 INFO mapred.Merger: Merging 1 sorted segments
16/01/30 03:45:51 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
16/01/30 03:45:51 INFO reduce.MergeManagerImpl: Merged 1 segments, 107 bytes to disk to satisfy reduce
16/01/30 03:45:51 INFO reduce.MergeManagerImpl: Merging 1 files, 111 bytes from disk
16/01/30 03:45:51 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/01/30 03:45:51 INFO mapred.Merger: Merging 1 sorted segments
16/01/30 03:45:51 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
16/01/30 03:45:51 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 03:45:51 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw2/./reducer
16/01/30 03:45:51 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapredu
16/01/30 03:45:51 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduc
16/01/30 03:45:51 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:45:51 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:45:51 INFO streaming.PipeMapRed: Records R/W=12/1
16/01/30 03:45:51 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 03:45:51 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 03:45:51 INFO mapred.Task: Task:attempt_local1399621981_0001_r_000000_0 is done. And is in the p
16/01/30 03:45:51 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 03:45:51 INFO mapred.Task: Task attempt_local1399621981_0001_r_000000_0 is allowed to commit now
16/01/30 03:45:51 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1399621981_0001_r_
16/01/30 03:45:51 INFO mapred.LocalJobRunner: Records R/W=12/1 > reduce
16/01/30 03:45:51 INFO mapred.Task: Task 'attempt_local1399621981_0001_r_000000_0' done.
16/01/30 03:45:51 INFO mapred.LocalJobRunner: Finishing task: attempt_local1399621981_0001_r_000000_0
16/01/30 03:45:51 INFO mapred.LocalJobRunner: reduce task executor complete.
16/01/30 03:45:51 INFO mapreduce.Job: Job job_local1399621981_0001 running in uber mode : false
16/01/30 03:45:51 INFO mapreduce.Job:  map 100% reduce 100%
16/01/30 03:45:51 INFO mapreduce.Job: Job job_local1399621981_0001 completed successfully
16/01/30 03:45:51 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=212040
                FILE: Number of bytes written=773449
                FILE: Number of read operations=0
```

```
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=88
                HDFS: Number of bytes written=26
                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=1
                Map output records=12
                Map output bytes=81
                Map output materialized bytes=111
                Input split bytes=104
                Combine input records=0
                Combine output records=0
                Reduce input groups=4
                Reduce shuffle bytes=111
                Reduce input records=12
                Reduce output records=4
                Spilled Records=24
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=568328192
        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=44
        File Output Format Counters
                Bytes Written=26
16/01/30 03:45:51 INFO streaming.StreamJob: Output directory: /user/vamsi/hw2/output_2_0
```

```
In [208]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/vamsi/hw2/output_2_0/part-00000
```

```
16/01/30 03:45:53 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
 1
hello 3
hey 2
hi 6
```

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.

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 [209]: #Generating a input file with 10000 random integers between 0 and 1,000,000
          #each line is of the form <integer,"NA">
```

```
from random import randint

N = 10000

with open('numbers.txt', 'w+') as f:
    for i in range(N):
        x = "<" + str(randint(0,1000000)) + "," + "\"NA\">\n"    #pick random numbers between
        f.write(x)
```

In [210]:
```
%%writefile mapper.py
#!/usr/bin/python
import sys
import re
# input comes from STDIN (standard input)
for line in sys.stdin:
    # remove the leading '<'
    line = line.lstrip('<')

    # remove the ending ,"NA">
    line = re.sub(',"NA">$',"",line)
    line = line.strip()
    print int(line)
```

```
Overwriting mapper.py
```

In [211]:
```
!chmod a+x mapper.py
```

In [212]:
```
%%writefile reducer.py
#!/usr/bin/python
from operator import itemgetter
import sys

# input comes from STDIN
for line in sys.stdin:
    line = line.strip()
    print "<" + line + "," + "\"NA\">"
```

```
Overwriting reducer.py
```

In [213]:
```
!chmod a+x reducer.py
```

In [214]:
```
!/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -put numbers.txt /user/vamsi/hw2
```

```
16/01/30 03:45:58 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

In [215]:
```
!/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hadoop jar /Users/Vamsi/Downloads/hadoop-2.7.1/bin/ha
-D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
-D mapred.reduce.tasks=1 \
-D mapred.text.key.comparator.options=-nr \
-mapper mapper.py \
-reducer reducer.py \
-input /user/vamsi/hw2/numbers.txt \
-output /user/vamsi/hw2/output_2_1
```

```
16/01/30 03:46:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/01/30 03:46:02 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.ses
16/01/30 03:46:02 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/01/30 03:46:02 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessio
16/01/30 03:46:02 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/30 03:46:02 INFO mapreduce.JobSubmitter: number of splits:1
16/01/30 03:46:02 INFO Configuration.deprecation: mapred.text.key.comparator.options is deprecated. Inst
16/01/30 03:46:02 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapred
16/01/30 03:46:02 INFO Configuration.deprecation: mapred.output.key.comparator.class is deprecated. Inst
16/01/30 03:46:02 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local600626759_0001
16/01/30 03:46:03 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/30 03:46:03 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/30 03:46:03 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
16/01/30 03:46:03 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/30 03:46:03 INFO mapreduce.Job: Running job: job_local600626759_0001
16/01/30 03:46:03 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/30 03:46:03 INFO mapred.LocalJobRunner: Starting task: attempt_local600626759_0001_m_000000_0
16/01/30 03:46:03 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/30 03:46:03 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
16/01/30 03:46:03 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/30 03:46:03 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/vamsi/hw2/numbers.tx
16/01/30 03:46:03 INFO mapred.MapTask: numReduceTasks: 1
16/01/30 03:46:03 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/30 03:46:03 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/30 03:46:03 INFO mapred.MapTask: soft limit at 83886080
16/01/30 03:46:03 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/30 03:46:03 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/30 03:46:03 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Map
16/01/30 03:46:03 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw2/./mapper.p
16/01/30 03:46:03 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.ta
16/01/30 03:46:03 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce
16/01/30 03:46:03 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.m
16/01/30 03:46:03 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.j
16/01/30 03:46:03 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce
16/01/30 03:46:03 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use map
16/01/30 03:46:03 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.
16/01/30 03:46:03 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.jo
16/01/30 03:46:03 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.us
16/01/30 03:46:03 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapredu
16/01/30 03:46:03 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.t
16/01/30 03:46:03 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapr
16/01/30 03:46:03 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:03 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:03 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:03 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:03 INFO streaming.PipeMapRed: Records R/W=9434/1
16/01/30 03:46:03 INFO streaming.PipeMapRed: R/W/S=10000/746/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:03 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 03:46:03 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 03:46:03 INFO mapred.LocalJobRunner:
16/01/30 03:46:03 INFO mapred.MapTask: Starting flush of map output
16/01/30 03:46:03 INFO mapred.MapTask: Spilling map output
16/01/30 03:46:03 INFO mapred.MapTask: bufstart = 0; bufend = 78923; bufvoid = 104857600
16/01/30 03:46:03 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26174400(104697600); lengt
```

```
16/01/30 03:46:03 INFO mapred.MapTask: Finished spill 0
16/01/30 03:46:03 INFO mapred.Task: Task:attempt_local600626759_0001_m_000000_0 is done. And is in the pr
16/01/30 03:46:03 INFO mapred.LocalJobRunner: Records R/W=9434/1
16/01/30 03:46:03 INFO mapred.Task: Task 'attempt_local600626759_0001_m_000000_0' done.
16/01/30 03:46:03 INFO mapred.LocalJobRunner: Finishing task: attempt_local600626759_0001_m_000000_0
16/01/30 03:46:03 INFO mapred.LocalJobRunner: map task executor complete.
16/01/30 03:46:03 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/30 03:46:03 INFO mapred.LocalJobRunner: Starting task: attempt_local600626759_0001_r_000000_0
16/01/30 03:46:03 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/30 03:46:03 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/01/30 03:46:03 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/30 03:46:03 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/01/30 03:46:03 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleL
16/01/30 03:46:03 INFO reduce.EventFetcher: attempt_local600626759_0001_r_000000_0 Thread started: EventF
16/01/30 03:46:04 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local6
16/01/30 03:46:04 INFO reduce.InMemoryMapOutput: Read 98925 bytes from map-output for attempt_local60062
16/01/30 03:46:04 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 98925, inMemory
16/01/30 03:46:04 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/30 03:46:04 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 03:46:04 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-
16/01/30 03:46:04 INFO mapred.Merger: Merging 1 sorted segments
16/01/30 03:46:04 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 9
16/01/30 03:46:04 INFO reduce.MergeManagerImpl: Merged 1 segments, 98925 bytes to disk to satisfy reduce
16/01/30 03:46:04 INFO reduce.MergeManagerImpl: Merging 1 files, 98929 bytes from disk
16/01/30 03:46:04 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/01/30 03:46:04 INFO mapred.Merger: Merging 1 sorted segments
16/01/30 03:46:04 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 9
16/01/30 03:46:04 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 03:46:04 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw2/./reducer
16/01/30 03:46:04 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapredu
16/01/30 03:46:04 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce
16/01/30 03:46:04 INFO mapreduce.Job: Job job_local600626759_0001 running in uber mode : false
16/01/30 03:46:04 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:04 INFO mapreduce.Job:  map 100% reduce 0%
16/01/30 03:46:04 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:04 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:04 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:04 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:04 INFO streaming.PipeMapRed: Records R/W=10000/1
16/01/30 03:46:04 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 03:46:04 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 03:46:04 INFO mapred.Task: Task:attempt_local600626759_0001_r_000000_0 is done. And is in the pr
16/01/30 03:46:04 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 03:46:04 INFO mapred.Task: Task attempt_local600626759_0001_r_000000_0 is allowed to commit now
16/01/30 03:46:04 INFO output.FileOutputCommitter: Saved output of task 'attempt_local600626759_0001_r_00
16/01/30 03:46:04 INFO mapred.LocalJobRunner: Records R/W=10000/1 > reduce
16/01/30 03:46:04 INFO mapred.Task: Task 'attempt_local600626759_0001_r_000000_0' done.
16/01/30 03:46:04 INFO mapred.LocalJobRunner: Finishing task: attempt_local600626759_0001_r_000000_0
16/01/30 03:46:04 INFO mapred.LocalJobRunner: reduce task executor complete.
16/01/30 03:46:05 INFO mapreduce.Job:  map 100% reduce 100%
16/01/30 03:46:05 INFO mapreduce.Job: Job job_local600626759_0001 completed successfully
16/01/30 03:46:05 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=409674
```

```
                FILE: Number of bytes written=1068931
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=277846
                HDFS: Number of bytes written=148923
                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=78923
                Map output materialized bytes=98929
                Input split bytes=100
                Combine input records=0
                Combine output records=0
                Reduce input groups=9947
                Reduce shuffle bytes=98929
                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)=0
                Total committed heap usage (bytes)=567279616
        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=138923
        File Output Format Counters
                Bytes Written=148923
16/01/30 03:46:05 INFO streaming.StreamJob: Output directory: /user/vamsi/hw2/output_2_1
```

```
In [216]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/vamsi/hw2/output_2_1/part-00000
```

```
16/01/30 03:46:06 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
<999987,"NA">
<999717,"NA">
<999664,"NA">
<999618,"NA">
<999608,"NA">
<999586,"NA">
<999575,"NA">
<999536,"NA">
<999408,"NA">
<999397,"NA">
cat: Unable to write to output stream.
```

```
In [217]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/vamsi/hw2/output_2_1/part-00000
```

```
16/01/30 03:46:08 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
<1026,"NA">
<830,"NA">
<704,"NA">
<657,"NA">
<512,"NA">
<350,"NA">
<272,"NA">
<264,"NA">
<192,"NA">
<91,"NA">
```

If there are multiple reducers, yes we would need additional steps. The output from these multiple reducers would need to be passed into another map-reduce task.

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.

```
In [218]: %%writefile mapper.py
          #!/usr/bin/python
          import sys
          import re

          for line in sys.stdin:
              words = []  #empty list for words
              email   = re.split('\t+',line)
              if(len(email)==4):
                  subject = re.split(r'[\s.,]+',email[2].strip())
                  body    = re.split(r'[\s.,]+',email[3].strip())
                  for s in subject:
                      words.append(s)    #appending list of words occuring in the subject
                  for b in body:
                      words.append(b)       #appending list of words occuring in the body
                  for word in words:
                          if(re.search('\w+',word)):
                              print "%s,1" %word        #emit word,1 to the reducer
```

```
Overwriting mapper.py
```

```
In [219]: !chmod a+x mapper.py
```

```
In [220]: %%writefile reducer.py
          #!/usr/bin/python
          import re
          import sys

          token_prev = None

          token_count = 0
          # input comes from STDIN
          for line in sys.stdin:
              linea = re.split(r',',line)

              if(token_prev!=None):
```

```
                if(token_prev != linea[0]):
                    print "%s,%d" %(token_prev,token_count)
                    token_count = 0

            token_count += int(linea[1])
            token_prev = linea[0]
```

```
Overwriting reducer.py
```

```
In [221]: !chmod a+x reducer.py
```

```
In [222]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -put enronemail_1h.txt /user/vamsi/hw2
```

```
16/01/30 03:46:11 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

```
In [223]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hadoop jar /Users/Vamsi/Downloads/hadoop-2.7.1/bin/ha
          -D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
          -D mapred.reduce.tasks=1 \
          -mapper mapper.py \
          -reducer reducer.py \
          -input /user/vamsi/hw2/enronemail_1h.txt \
          -output /user/vamsi/hw2/output_2_2
```

```
16/01/30 03:46:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/01/30 03:46:14 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.ses
16/01/30 03:46:14 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/01/30 03:46:14 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessio
16/01/30 03:46:15 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/30 03:46:15 INFO mapreduce.JobSubmitter: number of splits:1
16/01/30 03:46:15 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapred
16/01/30 03:46:15 INFO Configuration.deprecation: mapred.output.key.comparator.class is deprecated. Inst
16/01/30 03:46:15 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1188899611_0001
16/01/30 03:46:15 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/30 03:46:15 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/30 03:46:15 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
16/01/30 03:46:15 INFO mapreduce.Job: Running job: job_local1188899611_0001
16/01/30 03:46:15 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/30 03:46:15 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/30 03:46:15 INFO mapred.LocalJobRunner: Starting task: attempt_local1188899611_0001_m_000000_0
16/01/30 03:46:16 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/30 03:46:16 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/01/30 03:46:16 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/30 03:46:16 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/vamsi/hw2/enronemail
16/01/30 03:46:16 INFO mapred.MapTask: numReduceTasks: 1
16/01/30 03:46:16 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/30 03:46:16 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/30 03:46:16 INFO mapred.MapTask: soft limit at 83886080
16/01/30 03:46:16 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/30 03:46:16 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/30 03:46:16 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Map
16/01/30 03:46:16 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw2/./mapper.p
16/01/30 03:46:16 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.ta
16/01/30 03:46:16 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce
16/01/30 03:46:16 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.m
16/01/30 03:46:16 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.j
```

```
16/01/30 03:46:16 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce
16/01/30 03:46:16 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use map
16/01/30 03:46:16 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce
16/01/30 03:46:16 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.j
16/01/30 03:46:16 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.us
16/01/30 03:46:16 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapredu
16/01/30 03:46:16 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.
16/01/30 03:46:16 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapr
16/01/30 03:46:16 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:16 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:16 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:16 INFO streaming.PipeMapRed: Records R/W=101/1
16/01/30 03:46:16 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 03:46:16 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 03:46:16 INFO mapred.LocalJobRunner:
16/01/30 03:46:16 INFO mapred.MapTask: Starting flush of map output
16/01/30 03:46:16 INFO mapred.MapTask: Spilling map output
16/01/30 03:46:16 INFO mapred.MapTask: bufstart = 0; bufend = 275549; bufvoid = 104857600
16/01/30 03:46:16 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26092040(104368160); leng
16/01/30 03:46:16 INFO mapred.MapTask: Finished spill 0
16/01/30 03:46:16 INFO mapred.Task: Task:attempt_local1188899611_0001_m_000000_0 is done. And is in the p
16/01/30 03:46:16 INFO mapred.LocalJobRunner: Records R/W=101/1
16/01/30 03:46:16 INFO mapred.Task: Task 'attempt_local1188899611_0001_m_000000_0' done.
16/01/30 03:46:16 INFO mapred.LocalJobRunner: Finishing task: attempt_local1188899611_0001_m_000000_0
16/01/30 03:46:16 INFO mapred.LocalJobRunner: map task executor complete.
16/01/30 03:46:16 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/30 03:46:16 INFO mapred.LocalJobRunner: Starting task: attempt_local1188899611_0001_r_000000_0
16/01/30 03:46:16 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/30 03:46:16 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/01/30 03:46:16 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/30 03:46:16 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/01/30 03:46:16 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLi
16/01/30 03:46:16 INFO reduce.EventFetcher: attempt_local1188899611_0001_r_000000_0 Thread started: Event
16/01/30 03:46:16 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1
16/01/30 03:46:16 INFO reduce.InMemoryMapOutput: Read 336731 bytes from map-output for attempt_local1188
16/01/30 03:46:16 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 336731, inMemor
16/01/30 03:46:16 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/30 03:46:16 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 03:46:16 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-
16/01/30 03:46:16 INFO mapred.Merger: Merging 1 sorted segments
16/01/30 03:46:16 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 3
16/01/30 03:46:16 INFO mapreduce.Job: Job job_local1188899611_0001 running in uber mode : false
16/01/30 03:46:16 INFO mapreduce.Job:  map 100% reduce 0%
16/01/30 03:46:16 INFO reduce.MergeManagerImpl: Merged 1 segments, 336731 bytes to disk to satisfy reduc
16/01/30 03:46:16 INFO reduce.MergeManagerImpl: Merging 1 files, 336735 bytes from disk
16/01/30 03:46:16 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/01/30 03:46:16 INFO mapred.Merger: Merging 1 sorted segments
16/01/30 03:46:16 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 3
16/01/30 03:46:16 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 03:46:16 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw2/./reducer
16/01/30 03:46:16 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapredu
16/01/30 03:46:16 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce
16/01/30 03:46:16 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:16 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
```

```
16/01/30 03:46:16 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:17 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:17 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:17 INFO streaming.PipeMapRed: Records R/W=18181/1
16/01/30 03:46:17 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 03:46:17 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 03:46:17 INFO mapred.Task: Task:attempt_local1188899611_0001_r_000000_0 is done. And is in the p
16/01/30 03:46:17 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 03:46:17 INFO mapred.Task: Task attempt_local1188899611_0001_r_000000_0 is allowed to commit now
16/01/30 03:46:17 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1188899611_0001_r_0
16/01/30 03:46:17 INFO mapred.LocalJobRunner: Records R/W=18181/1 > reduce
16/01/30 03:46:17 INFO mapred.Task: Task 'attempt_local1188899611_0001_r_000000_0' done.
16/01/30 03:46:17 INFO mapred.LocalJobRunner: Finishing task: attempt_local1188899611_0001_r_000000_0
16/01/30 03:46:17 INFO mapred.LocalJobRunner: reduce task executor complete.
16/01/30 03:46:17 INFO mapreduce.Job:  map 100% reduce 100%
16/01/30 03:46:17 INFO mapreduce.Job: Job job_local1188899611_0001 completed successfully
16/01/30 03:46:17 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=885298
                FILE: Number of bytes written=1784543
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=407962
                HDFS: Number of bytes written=69393
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=101
                Map output records=30590
                Map output bytes=275549
                Map output materialized bytes=336735
                Input split bytes=106
                Combine input records=0
                Combine output records=0
                Reduce input groups=6236
                Reduce shuffle bytes=336735
                Reduce input records=30590
                Reduce output records=6235
                Spilled Records=61180
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=567279616
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
```

```
                Bytes Read=203981
        File Output Format Counters
                Bytes Written=69393
16/01/30 03:46:17 INFO streaming.StreamJob: Output directory: /user/vamsi/hw2/output_2_2
```

```
In [224]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/vamsi/hw2/output_2_2/part-00000
```

```
16/01/30 03:46:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
assistance,9
```

```
In [225]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/vamsi/hw2/output_2_2/part-00000 |
```

```
16/01/30 03:46:21 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

HW2.2.1 Using Hadoop MapReduce and your wordcount job (from HW2.2) determine the top-10 occurring tokens (most frequent tokens)

```
In [226]: %%writefile mapper.py
          #!/usr/bin/python
          import sys
          import re

          for line in sys.stdin:
              line = re.split(',',line)
              print "%s,%s" %(line[0].strip(),line[1].strip())
```

```
Overwriting mapper.py
```

```
In [227]: !chmod a+x mapper.py
```

```
In [228]: %%writefile reducer.py
          #!/usr/bin/python
          import re
          import sys

          # input comes from STDIN
          for line in sys.stdin:
              line = re.split(',',line)
              print "%s,%s" %(line[0].strip(),line[1].strip())
```

```
Overwriting reducer.py
```

```
In [229]: !chmod a+x reducer.py
```

```
In [230]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -put word_counts /user/vamsi/hw2
```

```
16/01/30 03:46:24 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

```
In [231]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hadoop jar /Users/Vamsi/Downloads/hadoop-2.7.1/bin/ha
          -D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
          -D stream.map.output.field.separator=, \
          -D stream.num.map.output.key.fields=2 \
          -D map.output.key.field.separator=, \
          -D mapred.text.key.comparator.options=-k2,2nr \
          -D mapred.reduce.tasks=1 \
          -mapper mapper.py \
          -reducer reducer.py \
          -input /user/vamsi/hw2/word_counts \
          -output /user/vamsi/hw2/output_2_2_1 \
```

```
16/01/30 03:46:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/01/30 03:46:27 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.ses
16/01/30 03:46:27 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/01/30 03:46:27 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessi
16/01/30 03:46:28 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/30 03:46:28 INFO mapreduce.JobSubmitter: number of splits:1
16/01/30 03:46:28 INFO Configuration.deprecation: map.output.key.field.separator is deprecated. Instead
16/01/30 03:46:28 INFO Configuration.deprecation: mapred.text.key.comparator.options is deprecated. Inst
16/01/30 03:46:28 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapred
16/01/30 03:46:28 INFO Configuration.deprecation: mapred.output.key.comparator.class is deprecated. Inst
16/01/30 03:46:28 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local136905951_0001
16/01/30 03:46:28 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/30 03:46:28 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/30 03:46:28 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
16/01/30 03:46:28 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/30 03:46:28 INFO mapreduce.Job: Running job: job_local136905951_0001
16/01/30 03:46:28 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/30 03:46:28 INFO mapred.LocalJobRunner: Starting task: attempt_local136905951_0001_m_000000_0
16/01/30 03:46:29 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/30 03:46:29 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/01/30 03:46:29 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/30 03:46:29 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/vamsi/hw2/word_count
16/01/30 03:46:29 INFO mapred.MapTask: numReduceTasks: 1
16/01/30 03:46:29 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/30 03:46:29 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/30 03:46:29 INFO mapred.MapTask: soft limit at 83886080
16/01/30 03:46:29 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/30 03:46:29 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/30 03:46:29 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Map
16/01/30 03:46:29 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw2/./mapper.p
16/01/30 03:46:29 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use map
16/01/30 03:46:29 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce
16/01/30 03:46:29 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapredu
16/01/30 03:46:29 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.t
16/01/30 03:46:29 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.ta
16/01/30 03:46:29 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce
16/01/30 03:46:29 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.m
16/01/30 03:46:29 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.j
16/01/30 03:46:29 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce
16/01/30 03:46:29 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.jo
16/01/30 03:46:29 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.us
16/01/30 03:46:29 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapr
16/01/30 03:46:29 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:29 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:29 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:29 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:29 INFO streaming.PipeMapRed: Records R/W=6235/1
16/01/30 03:46:29 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 03:46:29 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 03:46:29 INFO mapred.LocalJobRunner:
16/01/30 03:46:29 INFO mapred.MapTask: Starting flush of map output
16/01/30 03:46:29 INFO mapred.MapTask: Spilling map output
16/01/30 03:46:29 INFO mapred.MapTask: bufstart = 0; bufend = 69393; bufvoid = 104857600
16/01/30 03:46:29 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26189460(104757840); leng
```

```
16/01/30 03:46:29 INFO mapred.MapTask: Finished spill 0
16/01/30 03:46:29 INFO mapred.Task: Task:attempt_local136905951_0001_m_000000_0 is done. And is in the pr
16/01/30 03:46:29 INFO mapred.LocalJobRunner: Records R/W=6235/1
16/01/30 03:46:29 INFO mapred.Task: Task 'attempt_local136905951_0001_m_000000_0' done.
16/01/30 03:46:29 INFO mapred.LocalJobRunner: Finishing task: attempt_local136905951_0001_m_000000_0
16/01/30 03:46:29 INFO mapred.LocalJobRunner: map task executor complete.
16/01/30 03:46:29 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/30 03:46:29 INFO mapred.LocalJobRunner: Starting task: attempt_local136905951_0001_r_000000_0
16/01/30 03:46:29 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/30 03:46:29 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/01/30 03:46:29 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/30 03:46:29 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/01/30 03:46:29 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleL
16/01/30 03:46:29 INFO reduce.EventFetcher: attempt_local136905951_0001_r_000000_0 Thread started: EventF
16/01/30 03:46:29 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1
16/01/30 03:46:29 INFO reduce.InMemoryMapOutput: Read 81865 bytes from map-output for attempt_local13690
16/01/30 03:46:29 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 81865, inMemory
16/01/30 03:46:29 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/30 03:46:29 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 03:46:29 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on
16/01/30 03:46:29 INFO mapred.Merger: Merging 1 sorted segments
16/01/30 03:46:29 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 8
16/01/30 03:46:29 INFO reduce.MergeManagerImpl: Merged 1 segments, 81865 bytes to disk to satisfy reduce
16/01/30 03:46:29 INFO reduce.MergeManagerImpl: Merging 1 files, 81869 bytes from disk
16/01/30 03:46:29 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/01/30 03:46:29 INFO mapred.Merger: Merging 1 sorted segments
16/01/30 03:46:29 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 8
16/01/30 03:46:29 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 03:46:29 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw2/./reducer
16/01/30 03:46:29 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapredu
16/01/30 03:46:29 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce
16/01/30 03:46:29 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:29 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:29 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:29 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:29 INFO streaming.PipeMapRed: Records R/W=6235/1
16/01/30 03:46:29 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 03:46:29 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 03:46:29 INFO mapreduce.Job: Job job_local136905951_0001 running in uber mode : false
16/01/30 03:46:29 INFO mapreduce.Job:  map 100% reduce 0%
16/01/30 03:46:30 INFO mapred.Task: Task:attempt_local136905951_0001_r_000000_0 is done. And is in the pr
16/01/30 03:46:30 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 03:46:30 INFO mapred.Task: Task attempt_local136905951_0001_r_000000_0 is allowed to commit now
16/01/30 03:46:30 INFO output.FileOutputCommitter: Saved output of task 'attempt_local136905951_0001_r_00
16/01/30 03:46:30 INFO mapred.LocalJobRunner: Records R/W=6235/1 > reduce
16/01/30 03:46:30 INFO mapred.Task: Task 'attempt_local136905951_0001_r_000000_0' done.
16/01/30 03:46:30 INFO mapred.LocalJobRunner: Finishing task: attempt_local136905951_0001_r_000000_0
16/01/30 03:46:30 INFO mapred.LocalJobRunner: reduce task executor complete.
16/01/30 03:46:30 INFO mapreduce.Job:  map 100% reduce 100%
16/01/30 03:46:30 INFO mapreduce.Job: Job job_local136905951_0001 completed successfully
16/01/30 03:46:30 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=375554
                FILE: Number of bytes written=1020053
```

```
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=138786
                HDFS: Number of bytes written=69393
                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=6235
                Map output records=6235
                Map output bytes=69393
                Map output materialized bytes=81869
                Input split bytes=100
                Combine input records=0
                Combine output records=0
                Reduce input groups=6235
                Reduce shuffle bytes=81869
                Reduce input records=6235
                Reduce output records=6235
                Spilled Records=12470
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=567279616
        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=69393
        File Output Format Counters
                Bytes Written=69393
16/01/30 03:46:30 INFO streaming.StreamJob: Output directory: /user/vamsi/hw2/output_2_2_1

In [232]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/vamsi/hw2/output_2_2_1/part-00000

16/01/30 03:46:32 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
the,1201
to,894
and,620
of,535
a,516
in,402
you,401
your,379
for,354
on,251
cat: Unable to write to output stream.
```

HW2.3. Multinomial NAIVE BAYES with NO Smoothing Using the Enron data from HW1 and Hadoop MapReduce, write a mapper/reducer job(s) that will both learn Naive Bayes classifier and classify the Enron

email messages using the learnt Naive Bayes classifier. Use all white-space delimitted tokens as independent input variables (assume spaces, fullstops, commas as delimiters). Note: for multinomial Naive Bayes, the Pr(X=“assistance”|Y=SPAM) is calculated as follows:

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

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

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

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

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

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

```
In [233]: %%writefile mapper.py
          #!/usr/bin/python
          import sys
          import re

          #filename = sys.argv[1]
          # input comes from STDIN (standard input)

          for line in sys.stdin:
              words = []  #empty list for words
              email   = re.split('\t+',line)
              if(len(email)==4):
                  subject = re.split(r'[\s.,]+',email[2].strip())
                  body    = re.split(r'[\s.,]+',email[3].strip())
                  for s in subject:
                      words.append(s)    #appending list of words occuring in the subject
                  for b in body:
                      words.append(b)       #appending list of words occuring in the body
                  for word in words:
                          if(re.search('\w+',word)):
                              print "%s,%s,%s" %(email[0].strip(),email[1].strip(),word.strip())
                                          #emit email_ID,output_label,word to the reducer

Overwriting mapper.py

In [234]: !chmod a+x mapper.py

In [235]: %%writefile reducer.py
          #!/usr/bin/python
          import re
          import sys
          from math import log
```

```
email = {}
words = {}
spam_ec = 0
ham_ec = 0
total_ec = 0
spam_wc = 0
ham_wc = 0
total_spam_wc = 0
total_ham_wc = 0

# input comes from STDIN
for line in sys.stdin:
    line = re.split(r',',line)
    email_id = line[0].strip()
    spam = line[1].strip()
    word = line[2].strip()
    if word not in words.keys():
        words[word] = {'spam_count':0,'ham_count':0}

    if email_id not in email.keys():
        email[email_id] = {'spam':0,'words':[],'count':0}

    if(int(spam)==1):
        words[word]['spam_count'] += 1
        total_spam_wc +=1
    elif(int(spam)==0):
        words[word]['ham_count'] += 1
        total_ham_wc +=1
    email[email_id]['count'] += 1
    email[email_id]['spam'] = spam
    email[email_id]['words'].append(word)

#Computing priors

#P_prior_spam = Number of emails containing spam/total number of emails
#P_prior_ham = Number of emails containing ham/total number of emails
for e in email.keys():
    spam_ec += int(email[e]['spam'])
    total_ec += 1

P_prior_spam = float(spam_ec)/float(total_ec)
P_prior_ham = 1 - P_prior_spam

#Computing conditionals
#P(word|spam) and P(word|ham)

cond_probs = {}

for w in words.keys():
    wc_spam = words[w]['spam_count']
    wc_ham = words[w]['ham_count']
    p_w_spam = float(wc_spam)/(total_spam_wc)    #conditional probability of word given spam
    p_w_ham = float(wc_ham)/(total_ham_wc)       #conditional probability of word given ham
```

```
            cond_probs[w] = {'spam':p_w_spam,'ham':p_w_ham}

        #Now, onto predictions

        prediction = []

        for e in email.keys():
            p_spam_cond = 0
            p_ham_cond = 0
            for word in email[e]['words']:
                if(cond_probs[word]['spam'] !=float(0) and cond_probs[word]['ham'] !=float(0)):
                    p_spam_cond += log(cond_probs[word]['spam'])
                    p_ham_cond += log(cond_probs[word]['ham'])

            p_spam_given_word = log(P_prior_spam) + p_spam_cond
            p_ham_given_word = log(P_prior_ham) + p_ham_cond

            if(p_spam_given_word > p_ham_given_word):
                predict_spam = 1
            else:
                predict_spam = 0

            prediction.append([predict_spam,email[e]['spam'],e,p_spam_given_word,p_ham_given_word])

        correct =0
        total =0

        for p in prediction:
            if(p[0] == int(p[1])):
                correct += 1
            total +=1

        accuracy = 100*(float(correct)/total)

        print "The accuracy of the naive bayes classifier is %s" %accuracy

        for p in prediction:
            print p[0],p[1],p[2],p[3],p[4]
```

```
Overwriting reducer.py
```

```
In [236]: !chmod a+x reducer.py
```

```
In [237]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hadoop jar /Users/Vamsi/Downloads/hadoop-2.7.1/bin/ha
          -D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
          -D mapred.reduce.tasks=1 \
          -mapper mapper.py \
          -reducer reducer.py \
          -input /user/vamsi/hw2/enronemail_1h.txt \
          -output /user/vamsi/hw2/output_2_3
```

```
16/01/30 03:46:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/01/30 03:46:36 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.ses
16/01/30 03:46:36 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/01/30 03:46:36 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessio
```

```
16/01/30 03:46:36 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/30 03:46:37 INFO mapreduce.JobSubmitter: number of splits:1
16/01/30 03:46:37 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapred
16/01/30 03:46:37 INFO Configuration.deprecation: mapred.output.key.comparator.class is deprecated. Inst
16/01/30 03:46:37 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local43858998_0001
16/01/30 03:46:37 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/30 03:46:37 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/30 03:46:37 INFO mapreduce.Job: Running job: job_local43858998_0001
16/01/30 03:46:37 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
16/01/30 03:46:37 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/30 03:46:37 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/30 03:46:37 INFO mapred.LocalJobRunner: Starting task: attempt_local43858998_0001_m_000000_0
16/01/30 03:46:37 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/30 03:46:37 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/01/30 03:46:37 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/30 03:46:37 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/vamsi/hw2/enronemail
16/01/30 03:46:37 INFO mapred.MapTask: numReduceTasks: 1
16/01/30 03:46:37 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/30 03:46:37 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/30 03:46:37 INFO mapred.MapTask: soft limit at 83886080
16/01/30 03:46:37 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/30 03:46:37 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/30 03:46:37 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Map
16/01/30 03:46:37 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw2/./mapper.p
16/01/30 03:46:37 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.ta
16/01/30 03:46:37 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce
16/01/30 03:46:37 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.m
16/01/30 03:46:37 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.j
16/01/30 03:46:37 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce
16/01/30 03:46:37 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use map
16/01/30 03:46:37 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.
16/01/30 03:46:37 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.jo
16/01/30 03:46:37 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.us
16/01/30 03:46:37 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapredu
16/01/30 03:46:37 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.t
16/01/30 03:46:37 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapr
16/01/30 03:46:37 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:38 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:38 INFO streaming.PipeMapRed: Records R/W=73/1
16/01/30 03:46:38 INFO streaming.PipeMapRed: R/W/S=100/7029/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:38 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 03:46:38 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 03:46:38 INFO mapred.LocalJobRunner:
16/01/30 03:46:38 INFO mapred.MapTask: Starting flush of map output
16/01/30 03:46:38 INFO mapred.MapTask: Spilling map output
16/01/30 03:46:38 INFO mapred.MapTask: bufstart = 0; bufend = 999987; bufvoid = 104857600
16/01/30 03:46:38 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26092040(104368160); leng
16/01/30 03:46:38 INFO mapred.MapTask: Finished spill 0
16/01/30 03:46:38 INFO mapred.Task: Task:attempt_local43858998_0001_m_000000_0 is done. And is in the pro
16/01/30 03:46:38 INFO mapreduce.Job: Job job_local43858998_0001 running in uber mode : false
16/01/30 03:46:38 INFO mapreduce.Job:  map 0% reduce 0%
16/01/30 03:46:38 INFO mapred.LocalJobRunner: Records R/W=73/1
16/01/30 03:46:38 INFO mapred.Task: Task 'attempt_local43858998_0001_m_000000_0' done.
16/01/30 03:46:38 INFO mapred.LocalJobRunner: Finishing task: attempt_local43858998_0001_m_000000_0
```

```
16/01/30 03:46:38 INFO mapred.LocalJobRunner: map task executor complete.
16/01/30 03:46:38 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/30 03:46:38 INFO mapred.LocalJobRunner: Starting task: attempt_local43858998_0001_r_000000_0
16/01/30 03:46:38 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/30 03:46:38 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/01/30 03:46:38 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/30 03:46:38 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/01/30 03:46:38 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleL
16/01/30 03:46:38 INFO reduce.EventFetcher: attempt_local43858998_0001_r_000000_0 Thread started: EventFe
16/01/30 03:46:38 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local4
16/01/30 03:46:38 INFO reduce.InMemoryMapOutput: Read 1061171 bytes from map-output for attempt_local438
16/01/30 03:46:38 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1061171, inMem
16/01/30 03:46:38 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/30 03:46:38 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 03:46:38 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on
16/01/30 03:46:38 INFO mapred.Merger: Merging 1 sorted segments
16/01/30 03:46:38 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
16/01/30 03:46:38 INFO reduce.MergeManagerImpl: Merged 1 segments, 1061171 bytes to disk to satisfy redu
16/01/30 03:46:38 INFO reduce.MergeManagerImpl: Merging 1 files, 1061175 bytes from disk
16/01/30 03:46:38 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/01/30 03:46:38 INFO mapred.Merger: Merging 1 sorted segments
16/01/30 03:46:38 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
16/01/30 03:46:38 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 03:46:38 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw2/./reducer
16/01/30 03:46:38 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapredu
16/01/30 03:46:38 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduc
16/01/30 03:46:38 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:38 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:38 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:38 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:39 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:39 INFO mapreduce.Job:  map 100% reduce 0%
16/01/30 03:46:42 INFO streaming.PipeMapRed: Records R/W=30590/1
16/01/30 03:46:42 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 03:46:42 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 03:46:42 INFO mapred.Task: Task:attempt_local43858998_0001_r_000000_0 is done. And is in the pro
16/01/30 03:46:42 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 03:46:42 INFO mapred.Task: Task attempt_local43858998_0001_r_000000_0 is allowed to commit now
16/01/30 03:46:42 INFO output.FileOutputCommitter: Saved output of task 'attempt_local43858998_0001_r_000
16/01/30 03:46:42 INFO mapred.LocalJobRunner: Records R/W=30590/1 > reduce
16/01/30 03:46:42 INFO mapred.Task: Task 'attempt_local43858998_0001_r_000000_0' done.
16/01/30 03:46:42 INFO mapred.LocalJobRunner: Finishing task: attempt_local43858998_0001_r_000000_0
16/01/30 03:46:42 INFO mapred.LocalJobRunner: reduce task executor complete.
16/01/30 03:46:43 INFO mapreduce.Job:  map 100% reduce 100%
16/01/30 03:46:43 INFO mapreduce.Job: Job job_local43858998_0001 completed successfully
16/01/30 03:46:43 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=2334178
                FILE: Number of bytes written=3951839
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=407962
                HDFS: Number of bytes written=5692
```

```
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=101
                Map output records=30590
                Map output bytes=999987
                Map output materialized bytes=1061175
                Input split bytes=106
                Combine input records=0
                Combine output records=0
                Reduce input groups=15355
                Reduce shuffle bytes=1061175
                Reduce input records=30590
                Reduce output records=99
                Spilled Records=61180
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=568328192
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=203981
        File Output Format Counters
                Bytes Written=5692
16/01/30 03:46:43 INFO streaming.StreamJob: Output directory: /user/vamsi/hw2/output_2_3

In [238]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/vamsi/hw2/output_2_3/part-00000

16/01/30 03:46:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
The accuracy of the naive bayes classifier is 90.8163265306

In [239]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/vamsi/hw2/output_2_3/part-00000

16/01/30 03:46:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...

In [240]: %%writefile hist.py
          #!/usr/bin/python
          import re
          import sys
          import numpy as np
          import matplotlib.pyplot as plt

          filename = sys.argv[1]
          i = 0

          spam_post = []
          ham_post = []
```

```
with open(filename,'r') as f:
    for line in f:
        if(i==0):  #skip over the first line
            i=i+1
        else:
            fields = line.split()
            spam_post.append(float(fields[3]))
            ham_post.append(float(fields[4]))

spam_post_np = np.asarray(spam_post)
ham_post_np = np.asarray(ham_post)

#print type(spam_post_np)
#print spam_post_np.shape
# histogram of the spam posterior

#plt.hist(spam_post_np)
#plt.title('Class-SPAM Posterior Probability')
#plt.show()

# histogram of the ham posterior
#n, bins, patches = plt.hist(ham_post_np, 50, normed=1, facecolor='green', alpha=0.75)

#plt.xlabel('Class - HAM')
#plt.ylabel('Posterior Probability')
#plt.show()
```

```
Overwriting hist.py
```

```
In [241]: !python hist.py prediction.txt
```

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

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

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

```
In [242]: %%writefile mapper.py
#!/usr/bin/python
import sys
import re

#filename = sys.argv[1]
# input comes from STDIN (standard input)

for line in sys.stdin:
    words = []  #empty list for words
    email   = re.split('\t+',line)
    if(len(email)==4):
        subject = re.split(r'[\s.,]+',email[2].strip())
        body    = re.split(r'[\s.,]+',email[3].strip())
        for s in subject:
            words.append(s)    #appending list of words occuring in the subject
```

```
                for b in body:
                    words.append(b)       #appending list of words occuring in the body
                for word in words:
                        if(re.search('\w+',word)):
                            print "%s,%s,%s" %(email[0].strip(),email[1].strip(),word.strip())  #emit
```

```
Overwriting mapper.py
```

In [243]:
```
!chmod a+x mapper.py
```

In [244]:
```
%%writefile reducer.py
#!/usr/bin/python
import re
import sys
from math import log

email = {}
words = {}
spam_ec = 0
ham_ec = 0
total_ec = 0
spam_wc = 0
ham_wc = 0
total_spam_wc = 0
total_ham_wc = 0

# input comes from STDIN
for line in sys.stdin:
    line = re.split(r',',line)
    email_id = line[0].strip()
    spam = line[1].strip()
    word = line[2].strip()
    if word not in words.keys():
        words[word] = {'spam_count':0,'ham_count':0}

    if email_id not in email.keys():
        email[email_id] = {'spam':0,'words':[],'count':0}

    if(int(spam)==1):
        words[word]['spam_count'] += 1
        total_spam_wc +=1
    elif(int(spam)==0):
        words[word]['ham_count'] += 1
        total_ham_wc +=1
    email[email_id]['count'] += 1
    email[email_id]['spam'] = spam
    email[email_id]['words'].append(word)

#Computing priors

#P_prior_spam = Number of emails containing spam/total number of emails
#P_prior_ham = Number of emails containing ham/total number of emails
for e in email.keys():
    spam_ec += int(email[e]['spam'])
    total_ec += 1
```

```
P_prior_spam = float(spam_ec)/float(total_ec)
P_prior_ham = 1 - P_prior_spam

#Computing conditionals
#P(word|spam) and P(word|ham)

cond_probs = {}

for w in words.keys():
    wc_spam = words[w]['spam_count']
    wc_ham = words[w]['ham_count']
    p_w_spam = (float(wc_spam)+1)/(total_spam_wc+1)    #conditional probability of word given
    p_w_ham = (float(wc_ham)+1)/(total_ham_wc+1)       #conditional probability of word given
    cond_probs[w] = {'spam':p_w_spam,'ham':p_w_ham}

#Now, onto predictions

prediction = []

for e in email.keys():
    p_spam_cond = 0
    p_ham_cond = 0
    for word in email[e]['words']:
        #if(cond_probs[word]['spam'] !=float(0) and cond_probs[word]['ham'] !=float(0)):
        p_spam_cond += log(cond_probs[word]['spam'])
        p_ham_cond += log(cond_probs[word]['ham'])

    p_spam_given_word = log(P_prior_spam) + p_spam_cond
    p_ham_given_word = log(P_prior_ham) + p_ham_cond

    if(p_spam_given_word > p_ham_given_word):
        predict_spam = 1
    else:
        predict_spam = 0

    prediction.append([predict_spam,email[e]['spam'],e,p_spam_given_word,p_ham_given_word])

correct =0
total =0

for p in prediction:
    if(p[0] == int(p[1])):
        correct += 1
    total +=1

accuracy = 100*(float(correct)/total)

print "The accuracy of the naive bayes classifier is %s" %accuracy
```

```
Overwriting reducer.py
```

In [245]:
```
!chmod a+x reducer.py
```

In [246]:
```
!/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hadoop jar /Users/Vamsi/Downloads/hadoop-2.7.1/bin/ha
-D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
```

```
-D mapred.reduce.tasks=1 \
-mapper mapper.py \
-reducer reducer.py \
-input /user/vamsi/hw2/enronemail_1h.txt \
-output /user/vamsi/hw2/output_2_4
```

```
16/01/30 03:46:50 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/01/30 03:46:51 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.ses
16/01/30 03:46:51 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/01/30 03:46:51 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessio
16/01/30 03:46:52 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/30 03:46:52 INFO mapreduce.JobSubmitter: number of splits:1
16/01/30 03:46:52 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapred
16/01/30 03:46:52 INFO Configuration.deprecation: mapred.output.key.comparator.class is deprecated. Inst
16/01/30 03:46:52 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local570888281_0001
16/01/30 03:46:52 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/30 03:46:52 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
16/01/30 03:46:52 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/30 03:46:52 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/30 03:46:52 INFO mapreduce.Job: Running job: job_local570888281_0001
16/01/30 03:46:52 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/30 03:46:52 INFO mapred.LocalJobRunner: Starting task: attempt_local570888281_0001_m_000000_0
16/01/30 03:46:52 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/30 03:46:52 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/01/30 03:46:52 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/30 03:46:52 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/vamsi/hw2/enronemail
16/01/30 03:46:52 INFO mapred.MapTask: numReduceTasks: 1
16/01/30 03:46:52 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/30 03:46:52 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
16/01/30 03:46:52 INFO mapred.MapTask: soft limit at 83886080
16/01/30 03:46:52 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/30 03:46:52 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/30 03:46:52 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Map
16/01/30 03:46:52 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw2/./mapper.p
16/01/30 03:46:52 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.ta
16/01/30 03:46:52 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce
16/01/30 03:46:52 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.m
16/01/30 03:46:52 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.j
16/01/30 03:46:52 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce
16/01/30 03:46:52 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use map
16/01/30 03:46:52 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce
16/01/30 03:46:52 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.jo
16/01/30 03:46:52 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.us
16/01/30 03:46:52 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapredu
16/01/30 03:46:52 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.t
16/01/30 03:46:52 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapr
16/01/30 03:46:53 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:53 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:53 INFO streaming.PipeMapRed: Records R/W=73/1
16/01/30 03:46:53 INFO streaming.PipeMapRed: R/W/S=100/8082/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:53 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 03:46:53 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 03:46:53 INFO mapred.LocalJobRunner:
16/01/30 03:46:53 INFO mapred.MapTask: Starting flush of map output
```

```
16/01/30 03:46:53 INFO mapred.MapTask: Spilling map output
16/01/30 03:46:53 INFO mapred.MapTask: bufstart = 0; bufend = 999987; bufvoid = 104857600
16/01/30 03:46:53 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26092040(104368160); leng
16/01/30 03:46:53 INFO mapred.MapTask: Finished spill 0
16/01/30 03:46:53 INFO mapred.Task: Task:attempt_local570888281_0001_m_000000_0 is done. And is in the pr
16/01/30 03:46:53 INFO mapred.LocalJobRunner: Records R/W=73/1
16/01/30 03:46:53 INFO mapred.Task: Task 'attempt_local570888281_0001_m_000000_0' done.
16/01/30 03:46:53 INFO mapred.LocalJobRunner: Finishing task: attempt_local570888281_0001_m_000000_0
16/01/30 03:46:53 INFO mapred.LocalJobRunner: map task executor complete.
16/01/30 03:46:53 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/30 03:46:53 INFO mapred.LocalJobRunner: Starting task: attempt_local570888281_0001_r_000000_0
16/01/30 03:46:53 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/30 03:46:53 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/01/30 03:46:53 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/30 03:46:53 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/01/30 03:46:53 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleL
16/01/30 03:46:53 INFO reduce.EventFetcher: attempt_local570888281_0001_r_000000_0 Thread started: EventF
16/01/30 03:46:53 INFO mapreduce.Job: Job job_local570888281_0001 running in uber mode : false
16/01/30 03:46:53 INFO mapreduce.Job:  map 100% reduce 0%
16/01/30 03:46:53 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local5
16/01/30 03:46:53 INFO reduce.InMemoryMapOutput: Read 1061171 bytes from map-output for attempt_local570
16/01/30 03:46:53 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1061171, inMem
16/01/30 03:46:53 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/30 03:46:53 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 03:46:53 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on
16/01/30 03:46:53 INFO mapred.Merger: Merging 1 sorted segments
16/01/30 03:46:53 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
16/01/30 03:46:53 INFO reduce.MergeManagerImpl: Merged 1 segments, 1061171 bytes to disk to satisfy redu
16/01/30 03:46:53 INFO reduce.MergeManagerImpl: Merging 1 files, 1061175 bytes from disk
16/01/30 03:46:53 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/01/30 03:46:53 INFO mapred.Merger: Merging 1 sorted segments
16/01/30 03:46:53 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
16/01/30 03:46:53 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 03:46:53 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw2/./reducer
16/01/30 03:46:53 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapredu
16/01/30 03:46:53 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce
16/01/30 03:46:53 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:53 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:53 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:53 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:54 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:46:57 INFO streaming.PipeMapRed: Records R/W=30590/1
16/01/30 03:46:57 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 03:46:57 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 03:46:57 INFO mapred.Task: Task:attempt_local570888281_0001_r_000000_0 is done. And is in the pr
16/01/30 03:46:57 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 03:46:57 INFO mapred.Task: Task attempt_local570888281_0001_r_000000_0 is allowed to commit now
16/01/30 03:46:57 INFO output.FileOutputCommitter: Saved output of task 'attempt_local570888281_0001_r_00
16/01/30 03:46:57 INFO mapred.LocalJobRunner: Records R/W=30590/1 > reduce
16/01/30 03:46:57 INFO mapred.Task: Task 'attempt_local570888281_0001_r_000000_0' done.
16/01/30 03:46:57 INFO mapred.LocalJobRunner: Finishing task: attempt_local570888281_0001_r_000000_0
16/01/30 03:46:57 INFO mapred.LocalJobRunner: reduce task executor complete.
16/01/30 03:46:58 INFO mapreduce.Job:  map 100% reduce 100%
16/01/30 03:46:58 INFO mapreduce.Job: Job job_local570888281_0001 completed successfully
```

```
16/01/30 03:46:58 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=2334178
                FILE: Number of bytes written=3954851
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=407962
                HDFS: Number of bytes written=61
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=101
                Map output records=30590
                Map output bytes=999987
                Map output materialized bytes=1061175
                Input split bytes=106
                Combine input records=0
                Combine output records=0
                Reduce input groups=15355
                Reduce shuffle bytes=1061175
                Reduce input records=30590
                Reduce output records=1
                Spilled Records=61180
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=9
                Total committed heap usage (bytes)=492830720
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=203981
        File Output Format Counters
                Bytes Written=61
16/01/30 03:46:58 INFO streaming.StreamJob: Output directory: /user/vamsi/hw2/output_2_4

In [247]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/vamsi/hw2/output_2_4/part-00000

16/01/30 03:47:00 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
The accuracy of the naive bayes classifier is 98.9795918367
```

With laplace-add one smoothing, the accuracy is almost ˜99% whereas without smoothing it was around ˜91%. In general, smoothing helps by preventing a overfit of the training dataset. However , in this case since we train and test on the same data-set , the reason for differing accuracy is a bit different.

Without smoothing (2.3), we ignored tokens which did not occur in the email document to avoid log(0), which is undefined. With smoothing, we did not have to ignore those tokens because we would never hit log(0).

HW2.5. Repeat HW2.4. This time when modeling and classification ignore tokens with a frequency of less than three (3) in the training set. How does it affect the misclassifcation error of learnt naive multinomial Bayesian Classifier on the training dataset:

In [248]:
```
%%writefile mapper.py
#!/usr/bin/python
import sys
import re

#filename = sys.argv[1]
# input comes from STDIN (standard input)

for line in sys.stdin:
    words = []  #empty list for words
    email   = re.split('\t+',line)
    if(len(email)==4):
        subject = re.split(r'[\s.,]+',email[2].strip())
        body    = re.split(r'[\s.,]+',email[3].strip())
        for s in subject:
            words.append(s)    #appending list of words occuring in the subject
        for b in body:
            words.append(b)       #appending list of words occuring in the body
        for word in words:
                if(re.search('\w+',word)):
                    print "%s,%s,%s" %(email[0].strip(),email[1].strip(),word.strip())  #emit
```

```
Overwriting mapper.py
```

In [249]:
```
!chmod a+x mapper.py
```

In [250]:
```
%%writefile reducer.py
#!/usr/bin/python
import re
import sys
from math import log

email = {}
words = {}
spam_ec = 0
ham_ec = 0
total_ec = 0
spam_wc = 0
ham_wc = 0
total_spam_wc = 0
total_ham_wc = 0

# input comes from STDIN
for line in sys.stdin:
    line = re.split(r',',line)
    email_id = line[0].strip()
    spam = line[1].strip()
    word = line[2].strip()
    if word not in words.keys():
        words[word] = {'spam_count':0,'ham_count':0}
```

```
    if email_id not in email.keys():
        email[email_id] = {'spam':0,'words':[],'count':0}

    if(int(spam)==1):
        words[word]['spam_count'] += 1
        total_spam_wc +=1
    elif(int(spam)==0):
        words[word]['ham_count'] += 1
        total_ham_wc +=1

    email[email_id]['count'] += 1
    email[email_id]['spam'] = spam
    email[email_id]['words'].append(word)

#Computing priors

#P_prior_spam = Number of emails containing spam/total number of emails
#P_prior_ham = Number of emails containing ham/total number of emails
for e in email.keys():
    spam_ec += int(email[e]['spam'])
    total_ec += 1

P_prior_spam = float(spam_ec)/float(total_ec)
P_prior_ham = 1 - P_prior_spam

#Computing conditionals
#P(word|spam) and P(word|ham)

cond_probs = {}

for w in words.keys():
    wc_spam = words[w]['spam_count']
    wc_ham = words[w]['ham_count']
    p_w_spam = (float(wc_spam)+1)/(total_spam_wc+1)    #conditional probability of word given
    p_w_ham = (float(wc_ham)+1)/(total_ham_wc+1)       #conditional probability of word given
    if((wc_spam+wc_ham)<3):
        cond_probs[w] = {'spam':p_w_spam,'ignore_word':1,'ham':p_w_ham}
    else:
        cond_probs[w] = {'spam':p_w_spam,'ignore_word':0,'ham':p_w_ham}

#Now, onto predictions

prediction = []

for e in email.keys():
    p_spam_cond = 0
    p_ham_cond = 0
    for word in email[e]['words']:
        if(cond_probs[word]['ignore_word'] ==0):
            p_spam_cond += log(cond_probs[word]['spam'])
            p_ham_cond += log(cond_probs[word]['ham'])

    p_spam_given_word = log(P_prior_spam) + p_spam_cond
    p_ham_given_word = log(P_prior_ham) + p_ham_cond
```

```
            if(p_spam_given_word > p_ham_given_word):
                predict_spam = 1
            else:
                predict_spam = 0

            prediction.append([predict_spam,email[e]['spam'],e,p_spam_given_word,p_ham_given_word])

        correct = 0
        total =0

        for p in prediction:
            if(p[0] == int(p[1])):
                correct += 1
            total +=1

        accuracy = 100*(float(correct)/total)

        print "The accuracy of the naive bayes classifier is %s" %accuracy
```

```
Overwriting reducer.py
```

```
In [251]: !chmod a+x reducer.py
```

```
In [252]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hadoop jar /Users/Vamsi/Downloads/hadoop-2.7.1/bin/ha
          -D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
          -D mapred.reduce.tasks=1 \
          -mapper mapper.py \
          -reducer reducer.py \
          -input /user/vamsi/hw2/enronemail_1h.txt \
          -output /user/vamsi/hw2/output_2_5
```

```
16/01/30 03:47:02 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
16/01/30 03:47:03 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.ses
16/01/30 03:47:03 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
16/01/30 03:47:03 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessi
16/01/30 03:47:04 INFO mapred.FileInputFormat: Total input paths to process : 1
16/01/30 03:47:04 INFO mapreduce.JobSubmitter: number of splits:1
16/01/30 03:47:04 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapred
16/01/30 03:47:04 INFO Configuration.deprecation: mapred.output.key.comparator.class is deprecated. Inst
16/01/30 03:47:04 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1371474831_0001
16/01/30 03:47:04 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
16/01/30 03:47:04 INFO mapred.LocalJobRunner: OutputCommitter set in config null
16/01/30 03:47:04 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
16/01/30 03:47:04 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/30 03:47:04 INFO mapreduce.Job: Running job: job_local1371474831_0001
16/01/30 03:47:04 INFO mapred.LocalJobRunner: Waiting for map tasks
16/01/30 03:47:04 INFO mapred.LocalJobRunner: Starting task: attempt_local1371474831_0001_m_000000_0
16/01/30 03:47:04 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/30 03:47:04 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
16/01/30 03:47:04 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/30 03:47:04 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/vamsi/hw2/enronemail
16/01/30 03:47:04 INFO mapred.MapTask: numReduceTasks: 1
16/01/30 03:47:05 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
16/01/30 03:47:05 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
```

```
16/01/30 03:47:05 INFO mapred.MapTask: soft limit at 83886080
16/01/30 03:47:05 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
16/01/30 03:47:05 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
16/01/30 03:47:05 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Map
16/01/30 03:47:05 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw2/./mapper.p
16/01/30 03:47:05 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.ta
16/01/30 03:47:05 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce
16/01/30 03:47:05 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.m
16/01/30 03:47:05 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.j
16/01/30 03:47:05 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce
16/01/30 03:47:05 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use map
16/01/30 03:47:05 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce
16/01/30 03:47:05 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.jo
16/01/30 03:47:05 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.us
16/01/30 03:47:05 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapredu
16/01/30 03:47:05 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.t
16/01/30 03:47:05 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapr
16/01/30 03:47:05 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:47:05 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:47:05 INFO streaming.PipeMapRed: Records R/W=73/1
16/01/30 03:47:05 INFO streaming.PipeMapRed: R/W/S=100/6490/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:47:05 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 03:47:05 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 03:47:05 INFO mapred.LocalJobRunner:
16/01/30 03:47:05 INFO mapred.MapTask: Starting flush of map output
16/01/30 03:47:05 INFO mapred.MapTask: Spilling map output
16/01/30 03:47:05 INFO mapred.MapTask: bufstart = 0; bufend = 999987; bufvoid = 104857600
16/01/30 03:47:05 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26092040(104368160); leng
16/01/30 03:47:05 INFO mapred.MapTask: Finished spill 0
16/01/30 03:47:05 INFO mapred.Task: Task:attempt_local1371474831_0001_m_000000_0 is done. And is in the p
16/01/30 03:47:05 INFO mapred.LocalJobRunner: Records R/W=73/1
16/01/30 03:47:05 INFO mapred.Task: Task 'attempt_local1371474831_0001_m_000000_0' done.
16/01/30 03:47:05 INFO mapred.LocalJobRunner: Finishing task: attempt_local1371474831_0001_m_000000_0
16/01/30 03:47:05 INFO mapred.LocalJobRunner: map task executor complete.
16/01/30 03:47:05 INFO mapred.LocalJobRunner: Waiting for reduce tasks
16/01/30 03:47:05 INFO mapred.LocalJobRunner: Starting task: attempt_local1371474831_0001_r_000000_0
16/01/30 03:47:05 INFO mapreduce.Job: Job job_local1371474831_0001 running in uber mode : false
16/01/30 03:47:05 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
16/01/30 03:47:05 INFO mapreduce.Job:  map 100% reduce 0%
16/01/30 03:47:05 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
16/01/30 03:47:05 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
16/01/30 03:47:05 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
16/01/30 03:47:05 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLi
16/01/30 03:47:05 INFO reduce.EventFetcher: attempt_local1371474831_0001_r_000000_0 Thread started: Event
16/01/30 03:47:05 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1
16/01/30 03:47:05 INFO reduce.InMemoryMapOutput: Read 1061171 bytes from map-output for attempt_local137
16/01/30 03:47:05 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1061171, inMemo
16/01/30 03:47:05 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
16/01/30 03:47:05 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 03:47:05 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-
16/01/30 03:47:05 INFO mapred.Merger: Merging 1 sorted segments
16/01/30 03:47:05 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 1
16/01/30 03:47:06 INFO reduce.MergeManagerImpl: Merged 1 segments, 1061171 bytes to disk to satisfy redu
16/01/30 03:47:06 INFO reduce.MergeManagerImpl: Merging 1 files, 1061175 bytes from disk
```

```
16/01/30 03:47:06 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
16/01/30 03:47:06 INFO mapred.Merger: Merging 1 sorted segments
16/01/30 03:47:06 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
16/01/30 03:47:06 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 03:47:06 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/Vamsi/Documents/W261/hw2/./reducer
16/01/30 03:47:06 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapredu
16/01/30 03:47:06 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce
16/01/30 03:47:06 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:47:06 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:47:06 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:47:06 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:47:06 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
16/01/30 03:47:09 INFO streaming.PipeMapRed: Records R/W=30590/1
16/01/30 03:47:09 INFO streaming.PipeMapRed: MRErrorThread done
16/01/30 03:47:09 INFO streaming.PipeMapRed: mapRedFinished
16/01/30 03:47:10 INFO mapred.Task: Task:attempt_local1371474831_0001_r_000000_0 is done. And is in the p
16/01/30 03:47:10 INFO mapred.LocalJobRunner: 1 / 1 copied.
16/01/30 03:47:10 INFO mapred.Task: Task attempt_local1371474831_0001_r_000000_0 is allowed to commit now
16/01/30 03:47:10 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1371474831_0001_r_0
16/01/30 03:47:10 INFO mapred.LocalJobRunner: Records R/W=30590/1 > reduce
16/01/30 03:47:10 INFO mapred.Task: Task 'attempt_local1371474831_0001_r_000000_0' done.
16/01/30 03:47:10 INFO mapred.LocalJobRunner: Finishing task: attempt_local1371474831_0001_r_000000_0
16/01/30 03:47:10 INFO mapred.LocalJobRunner: reduce task executor complete.
16/01/30 03:47:10 INFO mapreduce.Job:  map 100% reduce 100%
16/01/30 03:47:10 INFO mapreduce.Job: Job job_local1371474831_0001 completed successfully
16/01/30 03:47:10 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=2334178
                FILE: Number of bytes written=3957863
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=407962
                HDFS: Number of bytes written=61
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=101
                Map output records=30590
                Map output bytes=999987
                Map output materialized bytes=1061175
                Input split bytes=106
                Combine input records=0
                Combine output records=0
                Reduce input groups=15355
                Reduce shuffle bytes=1061175
                Reduce input records=30590
                Reduce output records=1
                Spilled Records=61180
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
```

```
                Total committed heap usage (bytes)=567279616
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=203981
        File Output Format Counters
                Bytes Written=61
16/01/30 03:47:10 INFO streaming.StreamJob: Output directory: /user/vamsi/hw2/output_2_5
```

```
In [253]: !/Users/Vamsi/Downloads/hadoop-2.7.1/bin/hdfs dfs -cat /user/vamsi/hw2/output_2_5/part-00000

16/01/30 03:47:12 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
The accuracy of the naive bayes classifier is 98.9795918367
```

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

It always a good idea to benchmark your solutions against publicly available libraries such as SciKit-Learn, The Machine Learning toolkit available in Python. In this exercise, we benchmark ourselves against the SciKit-Learn implementation of multinomial Naive Bayes. For more information on this implementation see: http://scikit-learn.org/stable/modules/naive_bayes.html 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 [254]: %%writefile scikit_bm.py
          #!/usr/bin/python
          import sys
          import re
          from sklearn.naive_bayes import MultinomialNB
          from sklearn.feature_extraction.text import *
          import numpy as np

          filename = sys.argv[1]
          corpus = []   #empty list for words
          y_label = [] # output labels

          #Getting data in a format acceptable to scikit learn Multinomial Naive Bayes library
          with open(filename,"r") as f:
              for line in f:
                  words_p = []
                  email   = re.split('\t+',line)
                  if(len(email)==4):
                      entire_email = email[2] + email[3]
                      corpus.append(entire_email)
                      y_label.append(email[1])
```

```
mNB = MultinomialNB()
cv = CountVectorizer()

corpus_np = np.asarray(corpus)
y_label_np = np.asarray(y_label)
sp_matrix = cv.fit_transform(corpus_np)

model_mNB = mNB.fit(sp_matrix,y_label_np)

pred = mNB.predict(sp_matrix)

total = 0
correct = 0
for i in range(len(pred)):
    if(pred[i] == y_label_np[i]):
        correct = correct+1
    total=total+1

accuracy = 100*(float(correct)/float(total))
print "Accuracy using SciKit learn Multinomial Naive Bayes is %s",accuracy
```

```
Overwriting scikit_bm.py
```

```
In [255]: !python scikit_bm.py enronemail_1h.txt
```

```
Accuracy using SciKit learn Multinomial Naive Bayes is %s 100.0
```