hw2

September 15, 2015

1 DATASCI W261: Machine Learning at Scale

2 Gopala Tumuluri's Submission for HW2 and Parts of HW1

NOTE: I replicated a lot of 'reducer' code across problems to make it easy for the grader. I would never write redundant code in this fashion. Please be considerate on this point.

2.0.1 Start yarn and hdfs

```
In [234]: !/usr/local/Cellar/hadoop/2.7.0/sbin/start-yarn.sh
          !/usr/local/Cellar/hadoop/2.7.0/sbin/start-dfs.sh
```

```
starting yarn daemons
starting resourcemanager, logging to /usr/local/Cellar/hadoop/2.7.0/libexec/logs/yarn-gtumuluri-resourc
localhost: starting nodemanager, logging to /usr/local/Cellar/hadoop/2.7.0/libexec/logs/yarn-gtumuluri-
15/09/15 17:46:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/Cellar/hadoop/2.7.0/libexec/logs/hadoop-gtumuluri-na
localhost: starting datanode, logging to /usr/local/Cellar/hadoop/2.7.0/libexec/logs/hadoop-gtumuluri-da
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/Cellar/hadoop/2.7.0/libexec/logs/hadoop-gtumu
15/09/15 17:46:22 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

2.0.2 Create a folder in HDFS

```
In [235]: !hdfs dfs -mkdir -p /user/gtumuluri
```

```
15/09/15 17:46:30 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

2.1 Problem - 1.0.0

Define big data. Provide an example of a big data problem in your domain of expertise. In my mind, as a technology person, big data means content (loosely to mean bits/bytes) of value that don't fit on a single or a small set of machines, and certainly can't be computed on such resources. Data that requires multiple computers to store and compute qualifies as big data in my view.

I work in the healthcare industry where we bill and collect payments from patients online on behalf of hospitals and physicians offices. In my line work, I am routinely task with processing multiple gigabytes worth of web log data. This has not risen to the level of big data yet since we are a growing startup. But, with the exponential user base growth, I expect this problem to become a nice big data problem.

2.2 Problem - 1.0.1

Bias and Variance First let me address the irreducible error. This is the error that is inherent in any randomly produced/occurring/generated data. The variation that is purely due to chance and can't be eliminated/reduced by any model, especially for unseen observations. For any meaningful amount of data, a perfect fit can not be achieved even when using extremely high order polynomials.

Bias occurs when the model is under fit due to poor or limited features. This will manifest itself in the form of high training and test erorr. If one were to plot the curves in a reflected manner, they would see that the train and test errors are far apart from zero, and apart from each other. To address bias, one has to go for a more complex model.

Variance problem occurs when the model is overfit on the training data and perhaps performs perfectly on this data. But, has very large error when predicting test / unseen data. Using high order polynomials can result in training data being so 'well' fit that the error is minimal (near zero), but when the same model is used to predict test/new data, the error rate is very high. When you plot the erorr rates for train and test data, a high variance model will show that the train data has near zero error, and the test data will have a very high error rate, and even a growing error rate as the overfitting continues.

Model selection - First, plotting error rates and observing trends is crucial. One must select a model that strikes a good balance between train/test error. If the train error can't be reduced by much, one must consider this to be a bias problem and focus on additional feature selection. If the train error is low, but the test error starts to diverge away from zero, this should be considered a high variance problem, and the model should be simplified to not overfit the data.

2.3 Problem - 1.1: Read through pNaiveBayes.sh

```
In [288]: print 'done'
```

```
done
```

2.4 Problem 2.0

Race Condition A race condition in programming context is when one process that depends on a step, essentially races past it before that milestone has been met by yet another co-dependent process. This would result in unpredictable outputs. A simple example would be the map-reduce one where some mappers finish and if the reducer were to start its work before all mappers finished, the results would be unpredictable and wrong.

MapReduce MapReduce is a parallel, functional programming framework that allows for share-nothing distributed processing (to solve embarrassingly parallel problems). It differs from Hadoop greatly. Hadoop is a platform for stroing data in a highly distributed fashion with replication, redundancy and fault tolerance, and also allowing programs to operate on that data through centralized management and control. MapReduce on the other hand is an abstraction on top of Hadoop to actally perform the necessary computation on the data.

MapReduce Programming Paradigm MapReduce uses functional programming paradigm where one can write functions for map and reduce, and have those functions be applied to data in a specific step/order. In anotherway, mapreduce is also a 'gateway' type programming model where there are wait stages to ensure all parallel tasks complete a step before proceeding to the next.

2.5 Problem 2.1

2.5.1 Problem 2.1 - Random Number Generator

```
In [237]: import random

          # Generate 10,000 random integers between 0 and some large number
```

```
nums = [random.randint(0, 1000000) for i in range(0, 10000)]

# write them one line at a time to a file as key value
# in the format of 'number, NA'
file = open('2.1_randints.txt', 'w')
for num in nums:
    file.writelines(str(num) + ', NA\n')
```

2.5.2 Problem 2.1 - Hadoop Mapper for Sorting Numbers

```
In [238]: %%writefile 2.1_mapper.py
          #!/usr/bin/python
          import sys

          # Simply read the input from standard inpit and output
          # the number and the count (1) - the latter does not matter.
          for line in sys.stdin:
              line = line.strip()
              words = line.split(',')
              print '%s\t%s' % (words[0], 1)
```

```
Writing 2.1_mapper.py
```

2.5.3 Problem 2.1 - Hadoop Reducer for Sorting Numbers

```
In [239]: %%writefile 2.1_reducer.py
          #!/usr/bin/python
          import sys

          # Simply read the standard input and output the value
          # It comes in sorted order from the hadoop shuffle step
          for line in sys.stdin:
              line = line.strip()
              words = line.split()
              print words[0]
```

```
Writing 2.1_reducer.py
```

2.5.4 Problem 2.1 - File Upload of 10,000 Random Integers to HDFS

```
In [241]: !hdfs dfs -put 2.1_randints.txt /user/gtumuluri
```

```
15/09/15 17:47:11 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

2.5.5 Problem 2.1 - Run Hadoop MapReduce with Key Comparator Option to Perform Numeric Sort

Hadoop framework sorts key/value pairs output from the mappers using alphabetical sort order. This won't work for sorting integers. So, we change the key comparator to use a numeric sort in the shuffle phase so that the numbers appear in numerically sorted order at the reducer.

```
In [243]: !hadoop jar /usr/local/Cellar/hadoop/2.7.0/libexec/share/hadoop/tools/lib/hadoop-streaming-2.
```

```
15/09/15 17:48:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
15/09/15 17:48:01 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.se
15/09/15 17:48:01 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
```

```
15/09/15 17:48:01 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessi
15/09/15 17:48:02 INFO mapred.FileInputFormat: Total input paths to process : 1
15/09/15 17:48:02 INFO mapreduce.JobSubmitter: number of splits:1
15/09/15 17:48:02 INFO Configuration.deprecation: mapred.text.key.comparator.options is deprecated. Ins
15/09/15 17:48:02 INFO Configuration.deprecation: mapred.output.key.comparator.class is deprecated. Ins
15/09/15 17:48:02 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local540527564_0001
15/09/15 17:48:02 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
15/09/15 17:48:02 INFO mapred.LocalJobRunner: OutputCommitter set in config null
15/09/15 17:48:02 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
15/09/15 17:48:02 INFO mapreduce.Job: Running job: job_local540527564_0001
15/09/15 17:48:02 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
15/09/15 17:48:02 INFO mapred.LocalJobRunner: Waiting for map tasks
15/09/15 17:48:02 INFO mapred.LocalJobRunner: Starting task: attempt_local540527564_0001_m_000000_0
15/09/15 17:48:03 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
15/09/15 17:48:03 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
15/09/15 17:48:03 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
15/09/15 17:48:03 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/gtumuluri/randints.
15/09/15 17:48:03 INFO mapred.MapTask: numReduceTasks: 1
15/09/15 17:48:03 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
15/09/15 17:48:03 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
15/09/15 17:48:03 INFO mapred.MapTask: soft limit at 83886080
15/09/15 17:48:03 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
15/09/15 17:48:03 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
15/09/15 17:48:03 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Map
15/09/15 17:48:03 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/gtumuluri/Documents/BerkeleyMIDS/S
15/09/15 17:48:03 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.ta
15/09/15 17:48:03 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce
15/09/15 17:48:03 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.m
15/09/15 17:48:03 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.j
15/09/15 17:48:03 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce
15/09/15 17:48:03 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use map
15/09/15 17:48:03 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.
15/09/15 17:48:03 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.jo
15/09/15 17:48:03 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.us
15/09/15 17:48:03 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapredu
15/09/15 17:48:03 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.t
15/09/15 17:48:03 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapr
15/09/15 17:48:03 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 17:48:03 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 17:48:03 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 17:48:03 INFO streaming.PipeMapRed: MRErrorThread done
15/09/15 17:48:03 INFO streaming.PipeMapRed: Records R/W=100/1
15/09/15 17:48:03 INFO streaming.PipeMapRed: mapRedFinished
15/09/15 17:48:03 INFO mapred.LocalJobRunner:
15/09/15 17:48:03 INFO mapred.MapTask: Starting flush of map output
15/09/15 17:48:03 INFO mapred.MapTask: Spilling map output
15/09/15 17:48:03 INFO mapred.MapTask: bufstart = 0; bufend = 890; bufvoid = 104857600
15/09/15 17:48:03 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214000(104856000); leng
15/09/15 17:48:03 INFO mapred.MapTask: Finished spill 0
15/09/15 17:48:03 INFO mapred.Task: Task:attempt_local540527564_0001_m_000000_0 is done. And is in the pr
15/09/15 17:48:03 INFO mapred.LocalJobRunner: Records R/W=100/1
15/09/15 17:48:03 INFO mapred.Task: Task 'attempt_local540527564_0001_m_000000_0' done.
15/09/15 17:48:03 INFO mapred.LocalJobRunner: Finishing task: attempt_local540527564_0001_m_000000_0
15/09/15 17:48:03 INFO mapred.LocalJobRunner: map task executor complete.
```

```
15/09/15 17:48:03 INFO mapred.LocalJobRunner: Waiting for reduce tasks
15/09/15 17:48:03 INFO mapred.LocalJobRunner: Starting task: attempt_local540527564_0001_r_000000_0
15/09/15 17:48:03 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
15/09/15 17:48:03 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
15/09/15 17:48:03 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
15/09/15 17:48:03 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
15/09/15 17:48:03 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleL
15/09/15 17:48:03 INFO reduce.EventFetcher: attempt_local540527564_0001_r_000000_0 Thread started: EventF
15/09/15 17:48:03 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local5
15/09/15 17:48:03 INFO reduce.InMemoryMapOutput: Read 1092 bytes from map-output for attempt_local540527
15/09/15 17:48:03 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1092, inMemoryM
15/09/15 17:48:03 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
15/09/15 17:48:03 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 17:48:03 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-
15/09/15 17:48:03 INFO mapred.Merger: Merging 1 sorted segments
15/09/15 17:48:03 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
15/09/15 17:48:03 INFO reduce.MergeManagerImpl: Merged 1 segments, 1092 bytes to disk to satisfy reduce
15/09/15 17:48:03 INFO reduce.MergeManagerImpl: Merging 1 files, 1096 bytes from disk
15/09/15 17:48:03 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
15/09/15 17:48:03 INFO mapred.Merger: Merging 1 sorted segments
15/09/15 17:48:03 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
15/09/15 17:48:03 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 17:48:03 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/gtumuluri/Documents/BerkeleyMIDS/S
15/09/15 17:48:03 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapredu
15/09/15 17:48:03 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduc
15/09/15 17:48:03 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 17:48:03 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 17:48:03 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 17:48:03 INFO streaming.PipeMapRed: MRErrorThread done
15/09/15 17:48:03 INFO streaming.PipeMapRed: Records R/W=100/1
15/09/15 17:48:03 INFO streaming.PipeMapRed: mapRedFinished
15/09/15 17:48:03 INFO mapreduce.Job: Job job_local540527564_0001 running in uber mode : false
15/09/15 17:48:03 INFO mapreduce.Job:  map 100% reduce 0%
15/09/15 17:48:03 INFO mapred.Task: Task:attempt_local540527564_0001_r_000000_0 is done. And is in the pr
15/09/15 17:48:03 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 17:48:03 INFO mapred.Task: Task attempt_local540527564_0001_r_000000_0 is allowed to commit now
15/09/15 17:48:03 INFO output.FileOutputCommitter: Saved output of task 'attempt_local540527564_0001_r_0
15/09/15 17:48:03 INFO mapred.LocalJobRunner: Records R/W=100/1 > reduce
15/09/15 17:48:03 INFO mapred.Task: Task 'attempt_local540527564_0001_r_000000_0' done.
15/09/15 17:48:03 INFO mapred.LocalJobRunner: Finishing task: attempt_local540527564_0001_r_000000_0
15/09/15 17:48:03 INFO mapred.LocalJobRunner: reduce task executor complete.
15/09/15 17:48:04 INFO mapreduce.Job:  map 100% reduce 100%
15/09/15 17:48:04 INFO mapreduce.Job: Job job_local540527564_0001 completed successfully
15/09/15 17:48:04 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=214286
                FILE: Number of bytes written=801750
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2180
                HDFS: Number of bytes written=790
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
```

```
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=100
                Map output records=100
                Map output bytes=890
                Map output materialized bytes=1096
                Input split bytes=101
                Combine input records=0
                Combine output records=0
                Reduce input groups=100
                Reduce shuffle bytes=1096
                Reduce input records=100
                Reduce output records=100
                Spilled Records=200
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=6
                Total committed heap usage (bytes)=491782144
        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=1090
        File Output Format Counters
                Bytes Written=790
15/09/15 17:48:04 INFO streaming.StreamJob: Output directory: randintOutput
```

2.5.6 Problem 2.1 - Output of Sorted Numbers (Show First Few Lines)

```
In [244]: !hdfs dfs -cat randintOutput/part-00000 | head
```

```
15/09/15 17:48:10 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
6177
11787
12822
15793
18537
33585
42023
46559
96013
100725
```

2.6 Problem 2.2

2.6.1 Problem 2.2 - Mapper to Count Single Word Occurrence

```
In [256]: %%writefile 2.2_mapper.py
          #!/usr/bin/python
```

```
import sys

# Take input from the standard input
for line in sys.stdin:
    line = line.strip()
    items = line.split('\t')
    # Look for the 'body' portion of the email message
    if len(items) == 4:
        words = items[3].split()
        for word in words:
            # If the word in the body matches user specified,
            # output its occurrence
            if word.find(sys.argv[1]) == 0:
                print '%s\t%s' % (word, 1)
```

```
Overwriting 2.2_mapper.py
```

2.6.2 Problem 2.2 - Reducer to Count Single Word Occurrence

In [261]:
```
%%writefile 2.2_reducer.py
#!/usr/bin/python
import sys

# Initialize count, read lines from standard input
# and add up all words seen, print word and count
count = 0
for line in sys.stdin:
    if count == 0:
        words = line.strip()
        words = line.split()
    count += 1
print '%s\t%s' % (words[0], count)
```

```
Overwriting 2.2_reducer.py
```

2.6.3 Problem 2.2 - File Upload of Email Data Set to HDFS

In [262]:
```
!hdfs dfs -put enronemail_1h.txt /user/gtumuluri
```

```
15/09/15 17:59:29 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
put: ‘/user/gtumuluri/enronemail_1h.txt’: File exists
```

2.6.4 Problem 2.2 - Run Hadoop MapReduce with User Input Word

Passing ‘assistance’ as a user supplied word to the mapper.

In [263]:
```
!hadoop jar /usr/local/Cellar/hadoop/2.7.0/libexec/share/hadoop/tools/lib/hadoop-streaming-2.
```

```
15/09/15 17:59:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
15/09/15 17:59:35 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.ses
15/09/15 17:59:35 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
15/09/15 17:59:35 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessi
15/09/15 17:59:35 INFO mapred.FileInputFormat: Total input paths to process : 1
15/09/15 17:59:35 INFO mapreduce.JobSubmitter: number of splits:1
15/09/15 17:59:36 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local860394097_0001
15/09/15 17:59:36 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
```

```
15/09/15 17:59:36 INFO mapreduce.Job: Running job: job_local860394097_0001
15/09/15 17:59:36 INFO mapred.LocalJobRunner: OutputCommitter set in config null
15/09/15 17:59:36 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
15/09/15 17:59:36 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
15/09/15 17:59:36 INFO mapred.LocalJobRunner: Waiting for map tasks
15/09/15 17:59:36 INFO mapred.LocalJobRunner: Starting task: attempt_local860394097_0001_m_000000_0
15/09/15 17:59:36 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
15/09/15 17:59:36 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
15/09/15 17:59:36 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
15/09/15 17:59:36 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/gtumuluri/enronemail
15/09/15 17:59:36 INFO mapred.MapTask: numReduceTasks: 1
15/09/15 17:59:36 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
15/09/15 17:59:36 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
15/09/15 17:59:36 INFO mapred.MapTask: soft limit at 83886080
15/09/15 17:59:36 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
15/09/15 17:59:36 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
15/09/15 17:59:36 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Map
15/09/15 17:59:36 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/gtumuluri/Documents/BerkeleyMIDS/S
15/09/15 17:59:36 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.ta
15/09/15 17:59:36 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce
15/09/15 17:59:36 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.m
15/09/15 17:59:36 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.j
15/09/15 17:59:36 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce
15/09/15 17:59:36 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use map
15/09/15 17:59:36 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.
15/09/15 17:59:36 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.jo
15/09/15 17:59:36 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.us
15/09/15 17:59:36 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapredu
15/09/15 17:59:36 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.t
15/09/15 17:59:36 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapr
15/09/15 17:59:36 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 17:59:36 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 17:59:36 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 17:59:36 INFO streaming.PipeMapRed: MRErrorThread done
15/09/15 17:59:36 INFO streaming.PipeMapRed: Records R/W=101/1
15/09/15 17:59:36 INFO streaming.PipeMapRed: mapRedFinished
15/09/15 17:59:36 INFO mapred.LocalJobRunner:
15/09/15 17:59:36 INFO mapred.MapTask: Starting flush of map output
15/09/15 17:59:36 INFO mapred.MapTask: Spilling map output
15/09/15 17:59:36 INFO mapred.MapTask: bufstart = 0; bufend = 107; bufvoid = 104857600
15/09/15 17:59:36 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214368(104857472); leng
15/09/15 17:59:36 INFO mapred.MapTask: Finished spill 0
15/09/15 17:59:36 INFO mapred.Task: Task:attempt_local860394097_0001_m_000000_0 is done. And is in the pr
15/09/15 17:59:36 INFO mapred.LocalJobRunner: Records R/W=101/1
15/09/15 17:59:36 INFO mapred.Task: Task 'attempt_local860394097_0001_m_000000_0' done.
15/09/15 17:59:36 INFO mapred.LocalJobRunner: Finishing task: attempt_local860394097_0001_m_000000_0
15/09/15 17:59:36 INFO mapred.LocalJobRunner: map task executor complete.
15/09/15 17:59:36 INFO mapred.LocalJobRunner: Waiting for reduce tasks
15/09/15 17:59:36 INFO mapred.LocalJobRunner: Starting task: attempt_local860394097_0001_r_000000_0
15/09/15 17:59:36 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
15/09/15 17:59:36 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
15/09/15 17:59:36 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
15/09/15 17:59:36 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
15/09/15 17:59:36 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleL
```

```
15/09/15 17:59:36 INFO reduce.EventFetcher: attempt_local860394097_0001_r_000000_0 Thread started: EventF
15/09/15 17:59:36 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local8
15/09/15 17:59:36 INFO reduce.InMemoryMapOutput: Read 125 bytes from map-output for attempt_local8603940
15/09/15 17:59:36 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 125, inMemoryMa
15/09/15 17:59:36 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
15/09/15 17:59:36 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 17:59:36 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-
15/09/15 17:59:36 INFO mapred.Merger: Merging 1 sorted segments
15/09/15 17:59:36 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 1
15/09/15 17:59:36 INFO reduce.MergeManagerImpl: Merged 1 segments, 125 bytes to disk to satisfy reduce m
15/09/15 17:59:36 INFO reduce.MergeManagerImpl: Merging 1 files, 129 bytes from disk
15/09/15 17:59:36 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
15/09/15 17:59:36 INFO mapred.Merger: Merging 1 sorted segments
15/09/15 17:59:36 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 1
15/09/15 17:59:36 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 17:59:36 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/gtumuluri/Documents/BerkeleyMIDS/S
15/09/15 17:59:36 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapredu
15/09/15 17:59:36 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce
15/09/15 17:59:37 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 17:59:37 INFO streaming.PipeMapRed: Records R/W=8/1
15/09/15 17:59:37 INFO streaming.PipeMapRed: MRErrorThread done
15/09/15 17:59:37 INFO streaming.PipeMapRed: mapRedFinished
15/09/15 17:59:37 INFO mapred.Task: Task:attempt_local860394097_0001_r_000000_0 is done. And is in the pr
15/09/15 17:59:37 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 17:59:37 INFO mapred.Task: Task attempt_local860394097_0001_r_000000_0 is allowed to commit now
15/09/15 17:59:37 INFO output.FileOutputCommitter: Saved output of task 'attempt_local860394097_0001_r_00
15/09/15 17:59:37 INFO mapred.LocalJobRunner: Records R/W=8/1 > reduce
15/09/15 17:59:37 INFO mapred.Task: Task 'attempt_local860394097_0001_r_000000_0' done.
15/09/15 17:59:37 INFO mapred.LocalJobRunner: Finishing task: attempt_local860394097_0001_r_000000_0
15/09/15 17:59:37 INFO mapred.LocalJobRunner: reduce task executor complete.
15/09/15 17:59:37 INFO mapreduce.Job: Job job_local860394097_0001 running in uber mode : false
15/09/15 17:59:37 INFO mapreduce.Job:  map 100% reduce 100%
15/09/15 17:59:38 INFO mapreduce.Job: Job job_local860394097_0001 completed successfully
15/09/15 17:59:38 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=212364
                FILE: Number of bytes written=796797
                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=13
                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=8
                Map output bytes=107
                Map output materialized bytes=129
                Input split bytes=106
                Combine input records=0
                Combine output records=0
                Reduce input groups=3
```

```
                Reduce shuffle bytes=129
                Reduce input records=8
                Reduce output records=1
                Spilled Records=16
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=7
                Total committed heap usage (bytes)=491782144
        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=13
15/09/15 17:59:38 INFO streaming.StreamJob: Output directory: oneWordOutput
```

2.6.5 Problem 2.2 - Show Output of Single Word Count

Passing 'assistance' as a user supplied word to the mapper.

```
In [264]: !hdfs dfs -cat oneWordOutput/part-00000
```

```
15/09/15 17:59:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
assistance        8
```

2.7 Problem 2.3

2.7.1 Problem 2.3 - Mapper for Single Word Classification

```
In [265]: %%writefile 2.3_mapper.py
          #!/usr/bin/python
          import sys
          import string

          transtable = string.maketrans("","")

          # Read input from the standard input
          for line in sys.stdin:
              line = line.strip()
              items = line.split('\t')

              # If there is no content (as in subject/body in the data), skip
              if len(items) < 3:
                  continue
              if items[1] != '0' and items[1] != '1':
                  continue

              # Output a special word/keyword to allow reducer
              # to count the number of times a given class occurs.
```

```
    # Class is the second field in the data, so output
    # that by appending it to the 'class_' keyword string
    # and a count of 1 for each occurrence.
    print '%s\t%s' % ('class_' + items[1], 1)

    # If the line read has just subject, use that, otherwise
    # catenate with body also and use the entire content.
    if len(items) == 3:
        content = items[2]
    if len(items) == 4:
        content = items[2] + ' ' + items[3]

    # For each word in content, see if the word is same as user
    # chosen word, and then output the word and class to which
    # the document the word occurred in belongs to. This way, the
    # reducer can compute class frequencies for a given word.
    content = content.split()
    for word in content:
        # Remove punctuation
        word = word.translate(transtable, string.punctuation)
        if word.find(sys.argv[1]) == 0:
            print '%s\t%s' % (word, items[1])
```

```
Writing 2.3_mapper.py
```

2.7.2 Problem 2.3 - Reducer for Single Word Classification

```
In [271]: %%writefile 2.3_reducer.py
          #!/usr/bin/python
          import sys
          import math
          import string

          transtable = string.maketrans("","")

          # input comes from STDIN (standard input)

          # Placeholders for the vocabulary, frequencies
          # Dictionary is of form {vocab_word: {0: x, 1:y}} where
          # 0 and 1 are classes, and x and y are number of occurrences
          # of vocab word in respective classes.
          vocab = {}
          class0_freq = 0
          class1_freq = 0

          # Read each line from standard in and keep adding
          # class 0 and class 1 occurrences of the word into
          # the dictionary.
          for line in sys.stdin:
              words = line.strip('')
              words = line.split()
              if len(words) != 2:
                  continue
              vocab.setdefault(words[0], {0: 0, 1:0})
              if int(words[1]):
```

```
        vocab[words[0]][1] += 1
    else:
        vocab[words[0]][0] += 1

# Class frequencies come in special keywords from the mapper.
# Extract them and remove them from the dictionary.
class_0_freq = vocab['class_0'][1]
class_1_freq = vocab['class_1'][1]
vocab.pop('class_0')
vocab.pop('class_1')

# Compute class probabilities
class_0_prob = class_0_freq * 1.0 / (class_0_freq + class_1_freq)
class_1_prob = class_1_freq * 1.0 / (class_0_freq + class_1_freq)

# Comput size of the vocabulary for each class from the compiled
# dictionary above.
class_0_vocab = 0
class_1_vocab = 0
for key in vocab:
    class_0_vocab += vocab[key][0]
    class_1_vocab += vocab[key][1]

# The probability math implemented below to predict class given a document.
# P(Spam | Document) > P(Not Spam | Document)
# => ln(P(Spam | Document) / P(Not Spam | Document)) > 0
#
# So, we caclulate this value and the apply the above rule.
# ln(P(Spam | Document) / P(Not Spam | Document)) =
#   ln(P(Spam) / P(Not Spam)) + SUM(wi) {ln(P(word | Spam)/P(word | Not Spam))}

# P(Spam)/P(Not Spam) is always constant. Caclulate and store away.
ln_spam_not_spam = math.log(class_1_prob / class_0_prob)

# Read each document and compute the prediction using the algorithm above.
with open('enronemail_1h.txt') as infile:
    for document in infile:
        document = document.strip()
        document = document.split('\t')

        # If the document does not have subject/body fields, move on.
        if len(document) < 3 or len(document) > 4:
            continue

        # If it has the subject and body, catenate the two, otherwise use
        # the one available as the whole document.
        if len(document) == 4:
            content = document[2] + ' ' + document[3]
        else:
            content = document[2]

        # For each word in the document, compute the probability that the
        # word belongs to Spam/Not Spam classes.
        content = content.split()
```

```
    ln_word_spam_word_not_spam = 0
    for word in content:
        word = word.translate(transtable, string.punctuation)

        # If the word is in vocabulary, grab its frequency (plus one smoothing),
        # otherwise, just do plus one smoothing.
        if word in vocab:
            word_class_1_freq = vocab[word][1] + 1
            word_class_0_freq = vocab[word][0] + 1
        else:
            word_class_1_freq = 0 + 1
            word_class_0_freq = 0 + 1
        # Summation of the log ratios of word probabilities for each class.
        ln_word_spam_word_not_spam += math.log((word_class_1_freq * 1.0 /
                                                (class_1_vocab + len(vocab))) /
                                               (word_class_0_freq * 1.0 /
                                                (class_0_vocab + len(vocab))))

    # The final caculation of the log odds ratio of class. If this ratio is
    # greater than zero, we have class 1, otherwise, class 0.
    ln_doc_spam_not_spam = ln_spam_not_spam + ln_word_spam_word_not_spam
    if ln_doc_spam_not_spam > 0:
        print '%s\t%s\t%s' % (document[0], document[1], 1)
    else:
        print '%s\t%s\t%s' % (document[0], document[1], 0)
```

```
Overwriting 2.3_reducer.py
```

2.7.3 Problem 2.3 - Run Hadoop MapReduce Single Word Classifier with User Input Word

```
In [272]: !hadoop jar /usr/local/Cellar/hadoop/2.7.0/libexec/share/hadoop/tools/lib/hadoop-streaming-2.7
```

```
15/09/15 18:03:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
15/09/15 18:03:34 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.ses
15/09/15 18:03:34 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
15/09/15 18:03:34 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessio
15/09/15 18:03:35 INFO mapred.FileInputFormat: Total input paths to process : 1
15/09/15 18:03:35 INFO mapreduce.JobSubmitter: number of splits:1
15/09/15 18:03:35 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local205867234_0001
15/09/15 18:03:35 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
15/09/15 18:03:35 INFO mapred.LocalJobRunner: OutputCommitter set in config null
15/09/15 18:03:35 INFO mapreduce.Job: Running job: job_local205867234_0001
15/09/15 18:03:35 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
15/09/15 18:03:35 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
15/09/15 18:03:36 INFO mapred.LocalJobRunner: Waiting for map tasks
15/09/15 18:03:36 INFO mapred.LocalJobRunner: Starting task: attempt_local205867234_0001_m_000000_0
15/09/15 18:03:36 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
15/09/15 18:03:36 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only o
15/09/15 18:03:36 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
15/09/15 18:03:36 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/gtumuluri/enronemail
15/09/15 18:03:36 INFO mapred.MapTask: numReduceTasks: 1
15/09/15 18:03:36 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
15/09/15 18:03:36 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
15/09/15 18:03:36 INFO mapred.MapTask: soft limit at 83886080
15/09/15 18:03:36 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
```

```
15/09/15 18:03:36 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
15/09/15 18:03:36 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Ma
15/09/15 18:03:36 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/gtumuluri/Documents/BerkeleyMIDS/S
15/09/15 18:03:36 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.ta
15/09/15 18:03:36 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduc
15/09/15 18:03:36 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.m
15/09/15 18:03:36 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.j
15/09/15 18:03:36 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduc
15/09/15 18:03:36 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use map
15/09/15 18:03:36 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce
15/09/15 18:03:36 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.jo
15/09/15 18:03:36 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.us
15/09/15 18:03:36 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapredu
15/09/15 18:03:36 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.t
15/09/15 18:03:36 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapr
15/09/15 18:03:36 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:03:36 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:03:36 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:03:36 INFO streaming.PipeMapRed: MRErrorThread done
15/09/15 18:03:36 INFO streaming.PipeMapRed: Records R/W=101/1
15/09/15 18:03:36 INFO streaming.PipeMapRed: mapRedFinished
15/09/15 18:03:36 INFO mapred.LocalJobRunner:
15/09/15 18:03:36 INFO mapred.MapTask: Starting flush of map output
15/09/15 18:03:36 INFO mapred.MapTask: Spilling map output
15/09/15 18:03:36 INFO mapred.MapTask: bufstart = 0; bufend = 1117; bufvoid = 104857600
15/09/15 18:03:36 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26213964(104855856); leng
15/09/15 18:03:36 INFO mapred.MapTask: Finished spill 0
15/09/15 18:03:36 INFO mapred.Task: Task:attempt_local205867234_0001_m_000000_0 is done. And is in the pr
15/09/15 18:03:36 INFO mapred.LocalJobRunner: Records R/W=101/1
15/09/15 18:03:36 INFO mapred.Task: Task 'attempt_local205867234_0001_m_000000_0' done.
15/09/15 18:03:36 INFO mapred.LocalJobRunner: Finishing task: attempt_local205867234_0001_m_000000_0
15/09/15 18:03:36 INFO mapred.LocalJobRunner: map task executor complete.
15/09/15 18:03:36 INFO mapred.LocalJobRunner: Waiting for reduce tasks
15/09/15 18:03:36 INFO mapred.LocalJobRunner: Starting task: attempt_local205867234_0001_r_000000_0
15/09/15 18:03:36 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
15/09/15 18:03:36 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
15/09/15 18:03:36 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
15/09/15 18:03:36 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
15/09/15 18:03:36 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleL
15/09/15 18:03:36 INFO reduce.EventFetcher: attempt_local205867234_0001_r_000000_0 Thread started: EventF
15/09/15 18:03:36 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local2
15/09/15 18:03:36 INFO reduce.InMemoryMapOutput: Read 1337 bytes from map-output for attempt_local205867
15/09/15 18:03:36 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1337, inMemoryM
15/09/15 18:03:36 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
15/09/15 18:03:36 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 18:03:36 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-
15/09/15 18:03:36 INFO mapred.Merger: Merging 1 sorted segments
15/09/15 18:03:36 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 1
15/09/15 18:03:36 INFO reduce.MergeManagerImpl: Merged 1 segments, 1337 bytes to disk to satisfy reduce
15/09/15 18:03:36 INFO reduce.MergeManagerImpl: Merging 1 files, 1341 bytes from disk
15/09/15 18:03:36 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
15/09/15 18:03:36 INFO mapred.Merger: Merging 1 sorted segments
15/09/15 18:03:36 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 1
15/09/15 18:03:36 INFO mapred.LocalJobRunner: 1 / 1 copied.
```

```
15/09/15 18:03:36 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/gtumuluri/Documents/BerkeleyMIDS/S
15/09/15 18:03:36 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapredu
15/09/15 18:03:36 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduc
15/09/15 18:03:36 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:03:36 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:03:36 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:03:36 INFO streaming.PipeMapRed: Records R/W=109/1
15/09/15 18:03:36 INFO streaming.PipeMapRed: MRErrorThread done
15/09/15 18:03:36 INFO streaming.PipeMapRed: mapRedFinished
15/09/15 18:03:36 INFO mapreduce.Job: Job job_local205867234_0001 running in uber mode : false
15/09/15 18:03:36 INFO mapreduce.Job:  map 100% reduce 0%
15/09/15 18:03:37 INFO mapred.Task: Task:attempt_local205867234_0001_r_000000_0 is done. And is in the pr
15/09/15 18:03:37 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 18:03:37 INFO mapred.Task: Task attempt_local205867234_0001_r_000000_0 is allowed to commit now
15/09/15 18:03:37 INFO output.FileOutputCommitter: Saved output of task 'attempt_local205867234_0001_r_0
15/09/15 18:03:37 INFO mapred.LocalJobRunner: Records R/W=109/1 > reduce
15/09/15 18:03:37 INFO mapred.Task: Task 'attempt_local205867234_0001_r_000000_0' done.
15/09/15 18:03:37 INFO mapred.LocalJobRunner: Finishing task: attempt_local205867234_0001_r_000000_0
15/09/15 18:03:37 INFO mapred.LocalJobRunner: reduce task executor complete.
15/09/15 18:03:37 INFO mapreduce.Job:  map 100% reduce 100%
15/09/15 18:03:37 INFO mapreduce.Job: Job job_local205867234_0001 completed successfully
15/09/15 18:03:37 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=214788
                FILE: Number of bytes written=800457
                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=2672
                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=109
                Map output bytes=1117
                Map output materialized bytes=1341
                Input split bytes=106
                Combine input records=0
                Combine output records=0
                Reduce input groups=3
                Reduce shuffle bytes=1341
                Reduce input records=109
                Reduce output records=100
                Spilled Records=218
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=6
                Total committed heap usage (bytes)=488636416
        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=2672
15/09/15 18:03:37 INFO streaming.StreamJob: Output directory: classWordFreqOutput
```

2.7.4 Problem 2.3 - Show Output of Single Word Classifier

Passing 'assistance' as the only vocabulary word to classify by.

```
In [273]: !hdfs dfs -cat classWordFreqOutput/part-00000
```

```
15/09/15 18:03:43 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
0001.1999-12-10.farmer        0        0
0001.1999-12-10.kaminski        0        0
0001.2000-01-17.beck        0        0
0001.2000-06-06.lokay        0        0
0001.2001-02-07.kitchen        0        0
0001.2001-04-02.williams        0        0
0002.1999-12-13.farmer        0        0
0002.2001-02-07.kitchen        0        0
0002.2001-05-25.SA_and_HP        1        0
0002.2003-12-18.GP        1        0
0002.2004-08-01.BG        1        0
0003.1999-12-10.kaminski        0        0
0003.1999-12-14.farmer        0        0
0003.2000-01-17.beck        0        0
0003.2001-02-08.kitchen        0        0
0003.2003-12-18.GP        1        0
0003.2004-08-01.BG        1        0
0004.1999-12-10.kaminski        0        0
0004.1999-12-14.farmer        0        0
0004.2001-04-02.williams        0        0
0004.2001-06-12.SA_and_HP        1        0
0004.2004-08-01.BG        1        0
0005.1999-12-12.kaminski        0        0
0005.1999-12-14.farmer        0        0
0005.2000-06-06.lokay        0        0
0005.2001-02-08.kitchen        0        0
0005.2001-06-23.SA_and_HP        1        0
0005.2003-12-18.GP        1        0
0006.1999-12-13.kaminski        0        0
0006.2001-02-08.kitchen        0        0
0006.2001-04-03.williams        0        0
0006.2001-06-25.SA_and_HP        1        0
0006.2003-12-18.GP        1        0
0006.2004-08-01.BG        1        0
0007.1999-12-13.kaminski        0        0
0007.1999-12-14.farmer        0        0
0007.2000-01-17.beck        0        0
0007.2001-02-09.kitchen        0        0
```

```
0007.2003-12-18.GP        1        0
0007.2004-08-01.BG        1        0
0008.2001-02-09.kitchen       0        0
0008.2001-06-12.SA_and_HP       1        0
0008.2001-06-25.SA_and_HP       1        0
0008.2003-12-18.GP        1        0
0008.2004-08-01.BG        1        0
0009.1999-12-13.kaminski       0        0
0009.1999-12-14.farmer       0        0
0009.2000-06-07.lokay       0        0
0009.2001-02-09.kitchen       0        0
0009.2001-06-26.SA_and_HP       1        0
0009.2003-12-18.GP        1        0
0010.1999-12-14.farmer       0        0
0010.1999-12-14.kaminski       0        0
0010.2001-02-09.kitchen       0        0
0010.2001-06-28.SA_and_HP       1        0
0010.2003-12-18.GP        1        0
0010.2004-08-01.BG        1        0
0011.1999-12-14.farmer       0        0
0011.2001-06-28.SA_and_HP       1        0
0011.2001-06-29.SA_and_HP       1        0
0011.2003-12-18.GP        1        0
0011.2004-08-01.BG        1        0
0012.1999-12-14.farmer       0        0
0012.1999-12-14.kaminski       0        0
0012.2000-01-17.beck       0        0
0012.2000-06-08.lokay       0        0
0012.2001-02-09.kitchen       0        0
0012.2003-12-19.GP        1        0
0013.1999-12-14.farmer       0        0
0013.1999-12-14.kaminski       0        0
0013.2001-04-03.williams       0        0
0013.2001-06-30.SA_and_HP       1        0
0013.2004-08-01.BG        1        0
0014.1999-12-14.kaminski       0        0
0014.1999-12-15.farmer       0        0
0014.2001-02-12.kitchen       0        0
0014.2001-07-04.SA_and_HP       1        0
0014.2003-12-19.GP        1        0
0014.2004-08-01.BG        1        0
0015.1999-12-14.kaminski       0        0
0015.1999-12-15.farmer       0        0
0015.2000-06-09.lokay       0        0
0015.2001-02-12.kitchen       0        0
0015.2001-07-05.SA_and_HP       1        0
0015.2003-12-19.GP        1        0
0016.1999-12-15.farmer       0        0
0016.2001-02-12.kitchen       0        0
0016.2001-07-05.SA_and_HP       1        0
0016.2001-07-06.SA_and_HP       1        0
0016.2003-12-19.GP        1        0
0016.2004-08-01.BG        1        0
0017.1999-12-14.kaminski       0        0
```

```
0017.2000-01-17.beck        0        0
0017.2001-04-03.williams        0        0
0017.2003-12-18.GP        1        0
0017.2004-08-01.BG        1        0
0017.2004-08-02.BG        1        0
0018.1999-12-14.kaminski        0        0
0018.2001-07-13.SA_and_HP        1        0
0018.2003-12-18.GP        1        0
```

2.8 Problem 2.4

2.8.1 Problem 2.4 - Mapper for Multiple Word Classification

In [274]:
```
%%writefile 2.4_mapper.py
#!/usr/bin/python
import sys
import string

transtable = string.maketrans("","")

# Read input from the standard input
for line in sys.stdin:
    line = line.strip()
    items = line.split('\t')

    # If there is no content (as in subject/body in the data), skip
    if len(items) < 3:
        continue
    if items[1] != '0' and items[1] != '1':
        continue

    # Output a special word/keyword to allow reducer
    # to count the number of times a given class occurs.
    # Class is the second field in the data, so output
    # that by appending it to the 'class_' keyword string
    # and a count of 1 for each occurrence.
    print '%s\t%s' % ('class_' + items[1], 1)
    if len(items) == 3:
        content = items[2]
    if len(items) == 4:
        content = items[2] + ' ' + items[3]
    content = content.split()

    # For each word in content, see if the word is same as user
    # chosen word, and then output the word and class to which
    # the document the word occurred in belongs to. This way, the
    # reducer can compute class frequencies for a given word.
    for word in content:
        # Remove punctuation
        word = word.translate(transtable, string.punctuation)
        if word.find(sys.argv[1]) == 0 or word.find(sys.argv[2]) == 0 or word.find(sys.argv[3]
            print '%s\t%s' % (word, items[1])
```

```
Writing 2.4_mapper.py
```

2.8.2 Problem 2.4 - Reducer for Multiple Word Classification

In [275]:
```
%%writefile 2.4_reducer.py
#!/usr/bin/python
import sys
import math
import string

transtable = string.maketrans("","")

# input comes from STDIN (standard input)

# Placeholders for the vocabulary, frequencies
# Dictionary is of form {vocab_word: {0: x, 1:y}} where
# 0 and 1 are classes, and x and y are number of occurrences
# of vocab word in respective classes.
vocab = {}
class0_freq = 0
class1_freq = 0

# Read each line from standard in and keep adding
# class 0 and class 1 occurrences of the word into
# the dictionary.
for line in sys.stdin:
    words = line.strip('')
    words = line.split()
    if len(words) != 2:
        continue
    vocab.setdefault(words[0], {0: 0, 1:0})
    if int(words[1]):
        vocab[words[0]][1] += 1
    else:
        vocab[words[0]][0] += 1

# Class frequencies come in special keywords from the mapper.
# Extract them and remove them from the dictionary.
class_0_freq = vocab['class_0'][1]
class_1_freq = vocab['class_1'][1]
vocab.pop('class_0')
vocab.pop('class_1')

# Compute class probabilities
class_0_prob = class_0_freq * 1.0 / (class_0_freq + class_1_freq)
class_1_prob = class_1_freq * 1.0 / (class_0_freq + class_1_freq)

# Comput size of the vocabulary for each class from the compiled
# dictionary above.
class_0_vocab = 0
class_1_vocab = 0
for key in vocab:
    class_0_vocab += vocab[key][0]
    class_1_vocab += vocab[key][1]

# The probability math implemented below to predict class given a document.
```

```
# P(Spam | Document) > P(Not Spam | Document)
# => ln(P(Spam | Document) / P(Not Spam | Document)) > 0
#
# So, we caclulate this value and the apply the above rule.
# ln(P(Spam | Document) / P(Not Spam | Document)) =
#   ln(P(Spam) / P(Not Spam)) + SUM(wi) {ln(P(word | Spam)/P(word | Not Spam))}

# P(Spam)/P(Not Spam) is always constant. Caclulate and store away.
ln_spam_not_spam = math.log(class_1_prob / class_0_prob)

# Read each document and compute the prediction using the algorithm above.
with open('enronemail_1h.txt') as infile:
    for document in infile:
        document = document.strip()
        document = document.split('\t')

        # If the document does not have subject/body fields, move on.
        if len(document) < 3 or len(document) > 4:
            continue

        # If it has the subject and body, catenate the two, otherwise use
        # the one available as the whole document.
        if len(document) == 4:
            content = document[2] + ' ' + document[3]
        else:
            content = document[2]

        # For each word in the document, compute the probability that the
        # word belongs to Spam/Not Spam classes.
        content = content.split()
        ln_word_spam_word_not_spam = 0
        for word in content:
            word = word.translate(transtable, string.punctuation)

            # If the word is in vocabulary, grab its frequency (plus one smoothing),
            # otherwise, just do plus one smoothing.
            if word in vocab:
                word_class_1_freq = vocab[word][1] + 1
                word_class_0_freq = vocab[word][0] + 1
            else:
                word_class_1_freq = 0 + 1
                word_class_0_freq = 0 + 1
            # Summation of the log ratios of word probabilities for each class.
            ln_word_spam_word_not_spam += math.log((word_class_1_freq * 1.0 /
                                                    (class_1_vocab + len(vocab))) /
                                                   (word_class_0_freq * 1.0 /
                                                    (class_0_vocab + len(vocab))))

        # The final caculation of the log odds ratio of class. If this ratio is
        # greater than zero, we have class 1, otherwise, class 0.
        ln_doc_spam_not_spam = ln_spam_not_spam + ln_word_spam_word_not_spam
        if ln_doc_spam_not_spam > 0:
            print '%s\t%s\t%s' % (document[0], document[1], 1)
        else:
```

```
print '%s\t%s\t%s' % (document[0], document[1], 0)
```

```
Writing 2.4_reducer.py
```

2.8.3 Problem 2.4 - Run Hadoop MapReduce Multiple Word Classifier with User Input Word

Passing three words - assistance, viagra, enlragementWithATypo - as inputs to the mapper.

```
In [276]: !hadoop jar /usr/local/Cellar/hadoop/2.7.0/libexec/share/hadoop/tools/lib/hadoop-streaming-2.
```

```
15/09/15 18:04:52 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
15/09/15 18:04:53 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.se
15/09/15 18:04:53 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
15/09/15 18:04:53 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessi
15/09/15 18:04:53 INFO mapred.FileInputFormat: Total input paths to process : 1
15/09/15 18:04:53 INFO mapreduce.JobSubmitter: number of splits:1
15/09/15 18:04:54 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1759235215_0001
15/09/15 18:04:54 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
15/09/15 18:04:54 INFO mapred.LocalJobRunner: OutputCommitter set in config null
15/09/15 18:04:54 INFO mapreduce.Job: Running job: job_local1759235215_0001
15/09/15 18:04:54 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
15/09/15 18:04:54 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
15/09/15 18:04:54 INFO mapred.LocalJobRunner: Waiting for map tasks
15/09/15 18:04:54 INFO mapred.LocalJobRunner: Starting task: attempt_local1759235215_0001_m_000000_0
15/09/15 18:04:54 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
15/09/15 18:04:54 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
15/09/15 18:04:54 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
15/09/15 18:04:54 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/gtumuluri/enronemail
15/09/15 18:04:54 INFO mapred.MapTask: numReduceTasks: 1
15/09/15 18:04:54 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
15/09/15 18:04:54 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
15/09/15 18:04:54 INFO mapred.MapTask: soft limit at 83886080
15/09/15 18:04:54 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
15/09/15 18:04:54 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
15/09/15 18:04:54 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Map
15/09/15 18:04:54 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/gtumuluri/Documents/BerkeleyMIDS/S
15/09/15 18:04:54 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.ta
15/09/15 18:04:54 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce
15/09/15 18:04:54 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.m
15/09/15 18:04:54 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.j
15/09/15 18:04:54 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce
15/09/15 18:04:54 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use map
15/09/15 18:04:54 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.
15/09/15 18:04:54 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.jo
15/09/15 18:04:54 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.us
15/09/15 18:04:54 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapredu
15/09/15 18:04:54 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.t
15/09/15 18:04:54 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapr
15/09/15 18:04:54 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:04:54 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:04:54 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:04:54 INFO streaming.PipeMapRed: MRErrorThread done
15/09/15 18:04:54 INFO streaming.PipeMapRed: Records R/W=101/1
15/09/15 18:04:54 INFO streaming.PipeMapRed: mapRedFinished
15/09/15 18:04:54 INFO mapred.LocalJobRunner:
```

```
15/09/15 18:04:54 INFO mapred.MapTask: Starting flush of map output
15/09/15 18:04:54 INFO mapred.MapTask: Spilling map output
15/09/15 18:04:54 INFO mapred.MapTask: bufstart = 0; bufend = 1171; bufvoid = 104857600
15/09/15 18:04:54 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26213940(104855760); leng
15/09/15 18:04:54 INFO mapred.MapTask: Finished spill 0
15/09/15 18:04:54 INFO mapred.Task: Task:attempt_local1759235215_0001_m_000000_0 is done. And is in the p
15/09/15 18:04:54 INFO mapred.LocalJobRunner: Records R/W=101/1
15/09/15 18:04:54 INFO mapred.Task: Task 'attempt_local1759235215_0001_m_000000_0' done.
15/09/15 18:04:54 INFO mapred.LocalJobRunner: Finishing task: attempt_local1759235215_0001_m_000000_0
15/09/15 18:04:54 INFO mapred.LocalJobRunner: map task executor complete.
15/09/15 18:04:54 INFO mapred.LocalJobRunner: Waiting for reduce tasks
15/09/15 18:04:54 INFO mapred.LocalJobRunner: Starting task: attempt_local1759235215_0001_r_000000_0
15/09/15 18:04:55 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
15/09/15 18:04:55 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
15/09/15 18:04:55 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
15/09/15 18:04:55 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
15/09/15 18:04:55 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleL
15/09/15 18:04:55 INFO reduce.EventFetcher: attempt_local1759235215_0001_r_000000_0 Thread started: Event
15/09/15 18:04:55 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1
15/09/15 18:04:55 INFO reduce.InMemoryMapOutput: Read 1403 bytes from map-output for attempt_local175923
15/09/15 18:04:55 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 1403, inMemoryM
15/09/15 18:04:55 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
15/09/15 18:04:55 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 18:04:55 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-
15/09/15 18:04:55 INFO mapred.Merger: Merging 1 sorted segments
15/09/15 18:04:55 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
15/09/15 18:04:55 INFO reduce.MergeManagerImpl: Merged 1 segments, 1403 bytes to disk to satisfy reduce
15/09/15 18:04:55 INFO reduce.MergeManagerImpl: Merging 1 files, 1407 bytes from disk
15/09/15 18:04:55 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
15/09/15 18:04:55 INFO mapred.Merger: Merging 1 sorted segments
15/09/15 18:04:55 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size:
15/09/15 18:04:55 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 18:04:55 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/gtumuluri/Documents/BerkeleyMIDS/S
15/09/15 18:04:55 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapredu
15/09/15 18:04:55 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce
15/09/15 18:04:55 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:04:55 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:04:55 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:04:55 INFO streaming.PipeMapRed: Records R/W=115/1
15/09/15 18:04:55 INFO streaming.PipeMapRed: MRErrorThread done
15/09/15 18:04:55 INFO streaming.PipeMapRed: mapRedFinished
15/09/15 18:04:55 INFO mapreduce.Job: Job job_local1759235215_0001 running in uber mode : false
15/09/15 18:04:55 INFO mapreduce.Job:  map 100% reduce 0%
15/09/15 18:04:55 INFO mapred.Task: Task:attempt_local1759235215_0001_r_000000_0 is done. And is in the p
15/09/15 18:04:55 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 18:04:55 INFO mapred.Task: Task attempt_local1759235215_0001_r_000000_0 is allowed to commit now
15/09/15 18:04:55 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1759235215_0001_r_
15/09/15 18:04:55 INFO mapred.LocalJobRunner: Records R/W=115/1 > reduce
15/09/15 18:04:55 INFO mapred.Task: Task 'attempt_local1759235215_0001_r_000000_0' done.
15/09/15 18:04:55 INFO mapred.LocalJobRunner: Finishing task: attempt_local1759235215_0001_r_000000_0
15/09/15 18:04:55 INFO mapred.LocalJobRunner: reduce task executor complete.
15/09/15 18:04:56 INFO mapreduce.Job:  map 100% reduce 100%
15/09/15 18:04:56 INFO mapreduce.Job: Job job_local1759235215_0001 completed successfully
15/09/15 18:04:56 INFO mapreduce.Job: Counters: 35
```

```
        File System Counters
                FILE: Number of bytes read=214920
                FILE: Number of bytes written=803775
                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=2672
                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=115
                Map output bytes=1171
                Map output materialized bytes=1407
                Input split bytes=106
                Combine input records=0
                Combine output records=0
                Reduce input groups=4
                Reduce shuffle bytes=1407
                Reduce input records=115
                Reduce output records=100
                Spilled Records=230
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=6
                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=2672
15/09/15 18:04:56 INFO streaming.StreamJob: Output directory: classMultiWordFreqOutput
```

2.8.4 Problem 2.4 - Show Output of Multiple Word Classifier

Passing 'assistance' as the only vocabulary word to classify by.

```
In [277]: !hdfs dfs -cat classMultiWordFreqOutput/part-00000
```

```
15/09/15 18:05:02 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
0001.1999-12-10.farmer        0        0
0001.1999-12-10.kaminski        0        0
0001.2000-01-17.beck        0        0
0001.2000-06-06.lokay        0        0
0001.2001-02-07.kitchen        0        0
0001.2001-04-02.williams        0        0
```

```
0002.1999-12-13.farmer        0        0
0002.2001-02-07.kitchen        0        0
0002.2001-05-25.SA_and_HP        1        0
0002.2003-12-18.GP        1        0
0002.2004-08-01.BG        1        0
0003.1999-12-10.kaminski        0        0
0003.1999-12-14.farmer        0        0
0003.2000-01-17.beck        0        0
0003.2001-02-08.kitchen        0        0
0003.2003-12-18.GP        1        0
0003.2004-08-01.BG        1        0
0004.1999-12-10.kaminski        0        0
0004.1999-12-14.farmer        0        0
0004.2001-04-02.williams        0        0
0004.2001-06-12.SA_and_HP        1        0
0004.2004-08-01.BG        1        0
0005.1999-12-12.kaminski        0        0
0005.1999-12-14.farmer        0        0
0005.2000-06-06.lokay        0        0
0005.2001-02-08.kitchen        0        0
0005.2001-06-23.SA_and_HP        1        0
0005.2003-12-18.GP        1        0
0006.1999-12-13.kaminski        0        0
0006.2001-02-08.kitchen        0        0
0006.2001-04-03.williams        0        0
0006.2001-06-25.SA_and_HP        1        0
0006.2003-12-18.GP        1        0
0006.2004-08-01.BG        1        0
0007.1999-12-13.kaminski        0        0
0007.1999-12-14.farmer        0        0
0007.2000-01-17.beck        0        0
0007.2001-02-09.kitchen        0        0
0007.2003-12-18.GP        1        0
0007.2004-08-01.BG        1        0
0008.2001-02-09.kitchen        0        0
0008.2001-06-12.SA_and_HP        1        0
0008.2001-06-25.SA_and_HP        1        0
0008.2003-12-18.GP        1        0
0008.2004-08-01.BG        1        0
0009.1999-12-13.kaminski        0        0
0009.1999-12-14.farmer        0        0
0009.2000-06-07.lokay        0        0
0009.2001-02-09.kitchen        0        0
0009.2001-06-26.SA_and_HP        1        0
0009.2003-12-18.GP        1        0
0010.1999-12-14.farmer        0        0
0010.1999-12-14.kaminski        0        0
0010.2001-02-09.kitchen        0        0
0010.2001-06-28.SA_and_HP        1        0
0010.2003-12-18.GP        1        0
0010.2004-08-01.BG        1        0
0011.1999-12-14.farmer        0        0
0011.2001-06-28.SA_and_HP        1        0
0011.2001-06-29.SA_and_HP        1        0
```

```
0011.2003-12-18.GP        1        0
0011.2004-08-01.BG        1        0
0012.1999-12-14.farmer        0        0
0012.1999-12-14.kaminski        0        0
0012.2000-01-17.beck        0        0
0012.2000-06-08.lokay        0        0
0012.2001-02-09.kitchen        0        0
0012.2003-12-19.GP        1        0
0013.1999-12-14.farmer        0        0
0013.1999-12-14.kaminski        0        0
0013.2001-04-03.williams        0        0
0013.2001-06-30.SA_and_HP        1        0
0013.2004-08-01.BG        1        0
0014.1999-12-14.kaminski        0        0
0014.1999-12-15.farmer        0        0
0014.2001-02-12.kitchen        0        0
0014.2001-07-04.SA_and_HP        1        0
0014.2003-12-19.GP        1        0
0014.2004-08-01.BG        1        0
0015.1999-12-14.kaminski        0        0
0015.1999-12-15.farmer        0        0
0015.2000-06-09.lokay        0        0
0015.2001-02-12.kitchen        0        0
0015.2001-07-05.SA_and_HP        1        0
0015.2003-12-19.GP        1        0
0016.1999-12-15.farmer        0        0
0016.2001-02-12.kitchen        0        0
0016.2001-07-05.SA_and_HP        1        0
0016.2001-07-06.SA_and_HP        1        0
0016.2003-12-19.GP        1        0
0016.2004-08-01.BG        1        0
0017.1999-12-14.kaminski        0        0
0017.2000-01-17.beck        0        0
0017.2001-04-03.williams        0        0
0017.2003-12-18.GP        1        0
0017.2004-08-01.BG        1        0
0017.2004-08-02.BG        1        0
0018.1999-12-14.kaminski        0        0
0018.2001-07-13.SA_and_HP        1        0
0018.2003-12-18.GP        1        0
```

2.9 Problem 2.5

2.9.1 Problem 2.5 - Mapper for Full Naive Bayes Classification

In [279]:
```
%%writefile 2.5_mapper.py
#!/usr/bin/python
import sys
import string

transtable = string.maketrans("","")

# Read input from the standard input
for line in sys.stdin:
    line = line.strip()
```

```
    items = line.split('\t')

    # If there is no content (as in subject/body in the data), skip
    if len(items) < 3:
        continue
    if items[1] != '0' and items[1] != '1':
        continue

    # Output a special word/keyword to allow reducer
    # to count the number of times a given class occurs.
    # Class is the second field in the data, so output
    # that by appending it to the 'class_' keyword string
    # and a count of 1 for each occurrence.
    print '%s\t%s' % ('class_' + items[1], 1)
    if len(items) == 3:
        content = items[2]
    if len(items) == 4:
        content = items[2] + ' ' + items[3]
    content = content.split()

    # For each word in content, see if the word is same as user
    # chosen word, and then output the word and class to which
    # the document the word occurred in belongs to. This way, the
    # reducer can compute class frequencies for a given word.
    for word in content:
        # Remove punctuation
        word = word.translate(transtable, string.punctuation)
        print '%s\t%s' % (word, items[1])
```

```
Overwriting 2.5_mapper.py
```

2.9.2 Problem 2.5 - Reducer for Full Naive Bayes Classification

```
In [280]: %%writefile 2.5_reducer.py
          #!/usr/bin/python
          import sys
          import math
          import string

          transtable = string.maketrans("","")

          # input comes from STDIN (standard input)

          # Placeholders for the vocabulary, frequencies
          # Dictionary is of form {vocab_word: {0: x, 1:y}} where
          # 0 and 1 are classes, and x and y are number of occurrences
          # of vocab word in respective classes.
          vocab = {}
          class0_freq = 0
          class1_freq = 0

          # Read each line from standard in and keep adding
          # class 0 and class 1 occurrences of the word into
          # the dictionary.
          for line in sys.stdin:
```

```
    words = line.strip('')
    words = line.split()
    if len(words) != 2:
        continue
    vocab.setdefault(words[0], {0: 0, 1:0})
    if int(words[1]):
        vocab[words[0]][1] += 1
    else:
        vocab[words[0]][0] += 1

# Class frequencies come in special keywords from the mapper.
# Extract them and remove them from the dictionary.
class_0_freq = vocab['class_0'][1]
class_1_freq = vocab['class_1'][1]
vocab.pop('class_0')
vocab.pop('class_1')

# Compute class probabilities
class_0_prob = class_0_freq * 1.0 / (class_0_freq + class_1_freq)
class_1_prob = class_1_freq * 1.0 / (class_0_freq + class_1_freq)

# Comput size of the vocabulary for each class from the compiled
# dictionary above.
class_0_vocab = 0
class_1_vocab = 0
for key in vocab:
    class_0_vocab += vocab[key][0]
    class_1_vocab += vocab[key][1]

# The probability math implemented below to predict class given a document.
# P(Spam | Document) > P(Not Spam | Document)
# => ln(P(Spam | Document) / P(Not Spam | Document)) > 0
#
# So, we caclulate this value and the apply the above rule.
# ln(P(Spam | Document) / P(Not Spam | Document)) =
#   ln(P(Spam) / P(Not Spam)) + SUM(wi) {ln(P(word | Spam)/P(word | Not Spam))}

# P(Spam)/P(Not Spam) is always constant. Caclulate and store away.
ln_spam_not_spam = math.log(class_1_prob / class_0_prob)

# Read each document and compute the prediction using the algorithm above.
with open('enronemail_1h.txt') as infile:
    for document in infile:
        document = document.strip()
        document = document.split('\t')

        # If the document does not have subject/body fields, move on.
        if len(document) < 3 or len(document) > 4:
            continue

        # If it has the subject and body, catenate the two, otherwise use
        # the one available as the whole document.
        if len(document) == 4:
            content = document[2] + ' ' + document[3]
```

```
        else:
            content = document[2]

        # For each word in the document, compute the probability that the
        # word belongs to Spam/Not Spam classes.
        content = content.split()
        ln_word_spam_word_not_spam = 0
        for word in content:
            word = word.translate(transtable, string.punctuation)

            # If the word is in vocabulary, grab its frequency (plus one smoothing),
            # otherwise, just do plus one smoothing.
            if word in vocab:
                word_class_1_freq = vocab[word][1] + 1
                word_class_0_freq = vocab[word][0] + 1
            else:
                word_class_1_freq = 0 + 1
                word_class_0_freq = 0 + 1
            # Summation of the log ratios of word probabilities for each class.
            ln_word_spam_word_not_spam += math.log((word_class_1_freq * 1.0 /
                                                    (class_1_vocab + len(vocab))) /
                                                   (word_class_0_freq * 1.0 /
                                                    (class_0_vocab + len(vocab))))

        # The final caculation of the log odds ratio of class. If this ratio is
        # greater than zero, we have class 1, otherwise, class 0.
        ln_doc_spam_not_spam = ln_spam_not_spam + ln_word_spam_word_not_spam
        if ln_doc_spam_not_spam > 0:
            print '%s\t%s\t%s' % (document[0], document[1], 1)
        else:
            print '%s\t%s\t%s' % (document[0], document[1], 0)
```

```
Writing 2.5_reducer.py
```

2.9.3 Problem 2.5 - Run Hadoop MapReduce Full Naive Bayes Classifier

No words are passed as arguments in this case as we will use the full vocabulary to train and test the model.

```
In [281]: !hadoop jar /usr/local/Cellar/hadoop/2.7.0/libexec/share/hadoop/tools/lib/hadoop-streaming-2.
```

```
15/09/15 18:06:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
15/09/15 18:06:06 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.ses
15/09/15 18:06:06 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
15/09/15 18:06:06 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessi
15/09/15 18:06:06 INFO mapred.FileInputFormat: Total input paths to process : 1
15/09/15 18:06:06 INFO mapreduce.JobSubmitter: number of splits:1
15/09/15 18:06:07 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1826367675_0001
15/09/15 18:06:07 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
15/09/15 18:06:07 INFO mapred.LocalJobRunner: OutputCommitter set in config null
15/09/15 18:06:07 INFO mapreduce.Job: Running job: job_local1826367675_0001
15/09/15 18:06:07 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
15/09/15 18:06:07 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
15/09/15 18:06:07 INFO mapred.LocalJobRunner: Waiting for map tasks
15/09/15 18:06:07 INFO mapred.LocalJobRunner: Starting task: attempt_local1826367675_0001_m_000000_0
15/09/15 18:06:07 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
```

```
15/09/15 18:06:07 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
15/09/15 18:06:07 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
15/09/15 18:06:07 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/gtumuluri/enronemail
15/09/15 18:06:07 INFO mapred.MapTask: numReduceTasks: 1
15/09/15 18:06:07 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
15/09/15 18:06:07 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
15/09/15 18:06:07 INFO mapred.MapTask: soft limit at 83886080
15/09/15 18:06:07 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
15/09/15 18:06:07 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
15/09/15 18:06:07 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Map
15/09/15 18:06:07 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/gtumuluri/Documents/BerkeleyMIDS/S
15/09/15 18:06:07 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.ta
15/09/15 18:06:07 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce
15/09/15 18:06:07 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.m
15/09/15 18:06:07 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.j
15/09/15 18:06:07 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce
15/09/15 18:06:07 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use map
15/09/15 18:06:07 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce
15/09/15 18:06:07 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.jo
15/09/15 18:06:07 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.us
15/09/15 18:06:07 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapredu
15/09/15 18:06:07 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.t
15/09/15 18:06:07 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapr
15/09/15 18:06:07 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:06:07 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:06:07 INFO streaming.PipeMapRed: Records R/W=73/1
15/09/15 18:06:07 INFO streaming.PipeMapRed: R/W/S=100/14975/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:06:07 INFO streaming.PipeMapRed: MRErrorThread done
15/09/15 18:06:07 INFO streaming.PipeMapRed: mapRedFinished
15/09/15 18:06:07 INFO mapred.LocalJobRunner:
15/09/15 18:06:07 INFO mapred.MapTask: Starting flush of map output
15/09/15 18:06:07 INFO mapred.MapTask: Spilling map output
15/09/15 18:06:07 INFO mapred.MapTask: bufstart = 0; bufend = 249750; bufvoid = 104857600
15/09/15 18:06:07 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26083984(104335936); leng
15/09/15 18:06:08 INFO mapred.MapTask: Finished spill 0
15/09/15 18:06:08 INFO mapred.Task: Task:attempt_local1826367675_0001_m_000000_0 is done. And is in the p
15/09/15 18:06:08 INFO mapred.LocalJobRunner: Records R/W=73/1
15/09/15 18:06:08 INFO mapred.Task: Task 'attempt_local1826367675_0001_m_000000_0' done.
15/09/15 18:06:08 INFO mapred.LocalJobRunner: Finishing task: attempt_local1826367675_0001_m_000000_0
15/09/15 18:06:08 INFO mapred.LocalJobRunner: map task executor complete.
15/09/15 18:06:08 INFO mapred.LocalJobRunner: Waiting for reduce tasks
15/09/15 18:06:08 INFO mapred.LocalJobRunner: Starting task: attempt_local1826367675_0001_r_000000_0
15/09/15 18:06:08 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
15/09/15 18:06:08 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
15/09/15 18:06:08 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
15/09/15 18:06:08 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
15/09/15 18:06:08 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLi
15/09/15 18:06:08 INFO reduce.EventFetcher: attempt_local1826367675_0001_r_000000_0 Thread started: Event
15/09/15 18:06:08 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1
15/09/15 18:06:08 INFO reduce.InMemoryMapOutput: Read 314960 bytes from map-output for attempt_local1826
15/09/15 18:06:08 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 314960, inMemor
15/09/15 18:06:08 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
15/09/15 18:06:08 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 18:06:08 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-
```

```
15/09/15 18:06:08 INFO mapred.Merger: Merging 1 sorted segments
15/09/15 18:06:08 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 3
15/09/15 18:06:08 INFO mapreduce.Job: Job job_local1826367675_0001 running in uber mode : false
15/09/15 18:06:08 INFO mapreduce.Job:  map 100% reduce 0%
15/09/15 18:06:08 INFO reduce.MergeManagerImpl: Merged 1 segments, 314960 bytes to disk to satisfy reduc
15/09/15 18:06:08 INFO reduce.MergeManagerImpl: Merging 1 files, 314964 bytes from disk
15/09/15 18:06:08 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
15/09/15 18:06:08 INFO mapred.Merger: Merging 1 sorted segments
15/09/15 18:06:08 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 3
15/09/15 18:06:08 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 18:06:08 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/gtumuluri/Documents/BerkeleyMIDS/S
15/09/15 18:06:08 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapredu
15/09/15 18:06:08 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce
15/09/15 18:06:08 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:06:08 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:06:08 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:06:08 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:06:08 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:06:08 INFO streaming.PipeMapRed: Records R/W=32604/1
15/09/15 18:06:08 INFO streaming.PipeMapRed: MRErrorThread done
15/09/15 18:06:08 INFO streaming.PipeMapRed: mapRedFinished
15/09/15 18:06:08 INFO mapred.Task: Task:attempt_local1826367675_0001_r_000000_0 is done. And is in the p
15/09/15 18:06:08 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 18:06:08 INFO mapred.Task: Task attempt_local1826367675_0001_r_000000_0 is allowed to commit now
15/09/15 18:06:08 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1826367675_0001_r_0
15/09/15 18:06:08 INFO mapred.LocalJobRunner: Records R/W=32604/1 > reduce
15/09/15 18:06:08 INFO mapred.Task: Task 'attempt_local1826367675_0001_r_000000_0' done.
15/09/15 18:06:08 INFO mapred.LocalJobRunner: Finishing task: attempt_local1826367675_0001_r_000000_0
15/09/15 18:06:08 INFO mapred.LocalJobRunner: reduce task executor complete.
15/09/15 18:06:09 INFO mapreduce.Job:  map 100% reduce 100%
15/09/15 18:06:09 INFO mapreduce.Job: Job job_local1826367675_0001 completed successfully
15/09/15 18:06:09 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=842034
                FILE: Number of bytes written=1744290
                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=2672
                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=32604
                Map output bytes=249750
                Map output materialized bytes=314964
                Input split bytes=106
                Combine input records=0
                Combine output records=0
                Reduce input groups=5744
                Reduce shuffle bytes=314964
                Reduce input records=32604
```

```
                Reduce output records=100
                Spilled Records=65208
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=9
                Total committed heap usage (bytes)=488636416
        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=2672
15/09/15 18:06:09 INFO streaming.StreamJob: Output directory: fullClassificationOutput
```

2.9.4 Problem 2.5 - Show Output of Full Naive Bayes Classifier

No words are passed as argument. Training and prediction uses full vocabulary.

In [282]:
```
!hdfs dfs -cat fullClassificationOutput/part-00000
```

```
15/09/15 18:06:12 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
0001.1999-12-10.farmer        0        0
0001.1999-12-10.kaminski        0        0
0001.2000-01-17.beck        0        0
0001.2000-06-06.lokay        0        0
0001.2001-02-07.kitchen        0        0
0001.2001-04-02.williams        0        0
0002.1999-12-13.farmer        0        0
0002.2001-02-07.kitchen        0        0
0002.2001-05-25.SA_and_HP        1        1
0002.2003-12-18.GP        1        1
0002.2004-08-01.BG        1        1
0003.1999-12-10.kaminski        0        0
0003.1999-12-14.farmer        0        0
0003.2000-01-17.beck        0        0
0003.2001-02-08.kitchen        0        0
0003.2003-12-18.GP        1        1
0003.2004-08-01.BG        1        1
0004.1999-12-10.kaminski        0        0
0004.1999-12-14.farmer        0        0
0004.2001-04-02.williams        0        0
0004.2001-06-12.SA_and_HP        1        1
0004.2004-08-01.BG        1        1
0005.1999-12-12.kaminski        0        0
0005.1999-12-14.farmer        0        0
0005.2000-06-06.lokay        0        0
0005.2001-02-08.kitchen        0        0
0005.2001-06-23.SA_and_HP        1        1
0005.2003-12-18.GP        1        1
```

```
0006.1999-12-13.kaminski        0        0
0006.2001-02-08.kitchen        0        0
0006.2001-04-03.williams        0        0
0006.2001-06-25.SA_and_HP        1        1
0006.2003-12-18.GP        1        1
0006.2004-08-01.BG        1        1
0007.1999-12-13.kaminski        0        0
0007.1999-12-14.farmer        0        0
0007.2000-01-17.beck        0        0
0007.2001-02-09.kitchen        0        0
0007.2003-12-18.GP        1        1
0007.2004-08-01.BG        1        1
0008.2001-02-09.kitchen        0        0
0008.2001-06-12.SA_and_HP        1        1
0008.2001-06-25.SA_and_HP        1        1
0008.2003-12-18.GP        1        1
0008.2004-08-01.BG        1        1
0009.1999-12-13.kaminski        0        0
0009.1999-12-14.farmer        0        0
0009.2000-06-07.lokay        0        0
0009.2001-02-09.kitchen        0        0
0009.2001-06-26.SA_and_HP        1        1
0009.2003-12-18.GP        1        1
0010.1999-12-14.farmer        0        0
0010.1999-12-14.kaminski        0        0
0010.2001-02-09.kitchen        0        0
0010.2001-06-28.SA_and_HP        1        1
0010.2003-12-18.GP        1        1
0010.2004-08-01.BG        1        1
0011.1999-12-14.farmer        0        0
0011.2001-06-28.SA_and_HP        1        1
0011.2001-06-29.SA_and_HP        1        1
0011.2003-12-18.GP        1        1
0011.2004-08-01.BG        1        1
0012.1999-12-14.farmer        0        0
0012.1999-12-14.kaminski        0        0
0012.2000-01-17.beck        0        0
0012.2000-06-08.lokay        0        0
0012.2001-02-09.kitchen        0        0
0012.2003-12-19.GP        1        1
0013.1999-12-14.farmer        0        0
0013.1999-12-14.kaminski        0        0
0013.2001-04-03.williams        0        0
0013.2001-06-30.SA_and_HP        1        1
0013.2004-08-01.BG        1        1
0014.1999-12-14.kaminski        0        0
0014.1999-12-15.farmer        0        0
0014.2001-02-12.kitchen        0        0
0014.2001-07-04.SA_and_HP        1        1
0014.2003-12-19.GP        1        1
0014.2004-08-01.BG        1        1
0015.1999-12-14.kaminski        0        0
0015.1999-12-15.farmer        0        0
0015.2000-06-09.lokay        0        0
```

```
0015.2001-02-12.kitchen        0        0
0015.2001-07-05.SA_and_HP        1        1
0015.2003-12-19.GP        1        1
0016.1999-12-15.farmer        0        0
0016.2001-02-12.kitchen        0        0
0016.2001-07-05.SA_and_HP        1        1
0016.2001-07-06.SA_and_HP        1        1
0016.2003-12-19.GP        1        1
0016.2004-08-01.BG        1        1
0017.1999-12-14.kaminski        0        0
0017.2000-01-17.beck        0        0
0017.2001-04-03.williams        0        0
0017.2003-12-18.GP        1        1
0017.2004-08-01.BG        1        1
0017.2004-08-02.BG        1        1
0018.1999-12-14.kaminski        0        0
0018.2001-07-13.SA_and_HP        1        1
0018.2003-12-18.GP        1        1
```

2.9.5 Problem 2.5b - Reducer for Full Naive Bayes Classification REMOVE Infrequent Words

In this section, we re-do the classification of full Naive Bayes by dropping words that occur less than three times in the data set.

In [283]:
```
%%writefile 2.5b_reducer.py
#!/usr/bin/python
import sys
import math
import string

transtable = string.maketrans("","")

# input comes from STDIN (standard input)

# Placeholders for the vocabulary, frequencies
# Dictionary is of form {vocab_word: {0: x, 1:y}} where
# 0 and 1 are classes, and x and y are number of occurrences
# of vocab word in respective classes.
vocab = {}
class0_freq = 0
class1_freq = 0

# Read each line from standard in and keep adding
# class 0 and class 1 occurrences of the word into
# the dictionary.
for line in sys.stdin:
    words = line.strip('')
    words = line.split()
    if len(words) != 2:
        continue
    vocab.setdefault(words[0], {0: 0, 1:0})
    if int(words[1]):
        vocab[words[0]][1] += 1
    else:
        vocab[words[0]][0] += 1
```

```
# FIGURE OUT WHICH WORDS OCCUR WITH A FREQ OF LESS THAN 3
# AND REMOVE THEM FROM THE VOCABULARY.
exclude_list = []
for key in vocab:
    if sum(vocab[key].values()) < 3:
        exclude_list.append(key)
for word in exclude_list:
    vocab.pop(word)

# Class frequencies come in special keywords from the mapper.
# Extract them and remove them from the dictionary.
class_0_freq = vocab['class_0'][1]
class_1_freq = vocab['class_1'][1]
vocab.pop('class_0')
vocab.pop('class_1')

# Compute class probabilities
class_0_prob = class_0_freq * 1.0 / (class_0_freq + class_1_freq)
class_1_prob = class_1_freq * 1.0 / (class_0_freq + class_1_freq)

# Comput size of the vocabulary for each class from the compiled
# dictionary above.
class_0_vocab = 0
class_1_vocab = 0
for key in vocab:
    class_0_vocab += vocab[key][0]
    class_1_vocab += vocab[key][1]

# The probability math implemented below to predict class given a document.
# P(Spam | Document) > P(Not Spam | Document)
# => ln(P(Spam | Document) / P(Not Spam | Document)) > 0
#
# So, we caclulate this value and the apply the above rule.
# ln(P(Spam | Document) / P(Not Spam | Document)) =
#   ln(P(Spam) / P(Not Spam)) + SUM(wi) {ln(P(word | Spam)/P(word | Not Spam))}

# P(Spam)/P(Not Spam) is always constant. Caclulate and store away.
ln_spam_not_spam = math.log(class_1_prob / class_0_prob)

# Read each document and compute the prediction using the algorithm above.
with open('enronemail_1h.txt') as infile:
    for document in infile:
        document = document.strip()
        document = document.split('\t')

        # If the document does not have subject/body fields, move on.
        if len(document) < 3 or len(document) > 4:
            continue

        # If it has the subject and body, catenate the two, otherwise use
        # the one available as the whole document.
        if len(document) == 4:
            content = document[2] + ' ' + document[3]
```

```
        else:
            content = document[2]

        # For each word in the document, compute the probability that the
        # word belongs to Spam/Not Spam classes.
        content = content.split()
        ln_word_spam_word_not_spam = 0
        for word in content:
            word = word.translate(transtable, string.punctuation)

            # If the word is in vocabulary, grab its frequency (plus one smoothing),
            # otherwise, just do plus one smoothing.
            if word in vocab:
                word_class_1_freq = vocab[word][1] + 1
                word_class_0_freq = vocab[word][0] + 1
            else:
                word_class_1_freq = 0 + 1
                word_class_0_freq = 0 + 1
            # Summation of the log ratios of word probabilities for each class.
            ln_word_spam_word_not_spam += math.log((word_class_1_freq * 1.0 /
                                                    (class_1_vocab + len(vocab))) /
                                                   (word_class_0_freq * 1.0 /
                                                    (class_0_vocab + len(vocab))))

        # The final caculation of the log odds ratio of class. If this ratio is
        # greater than zero, we have class 1, otherwise, class 0.
        ln_doc_spam_not_spam = ln_spam_not_spam + ln_word_spam_word_not_spam
        if ln_doc_spam_not_spam > 0:
            print '%s\t%s\t%s' % (document[0], document[1], 1)
        else:
            print '%s\t%s\t%s' % (document[0], document[1], 0)
```

```
Writing 2.5b_reducer.py
```

2.9.6 Problem 2.5b - Run Hadoop MapReduce Full Naive Bayes Classifier REMOVE Infrequent Words

No words are passed as arguments in this case as we will use the full vocabulary to train and test the model.

```
In [284]: !hadoop jar /usr/local/Cellar/hadoop/2.7.0/libexec/share/hadoop/tools/lib/hadoop-streaming-2.
```

```
15/09/15 18:07:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
15/09/15 18:07:06 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.ses
15/09/15 18:07:06 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
15/09/15 18:07:06 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessi
15/09/15 18:07:06 INFO mapred.FileInputFormat: Total input paths to process : 1
15/09/15 18:07:06 INFO mapreduce.JobSubmitter: number of splits:1
15/09/15 18:07:06 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1207432264_0001
15/09/15 18:07:07 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
15/09/15 18:07:07 INFO mapred.LocalJobRunner: OutputCommitter set in config null
15/09/15 18:07:07 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputComm
15/09/15 18:07:07 INFO mapreduce.Job: Running job: job_local1207432264_0001
15/09/15 18:07:07 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
15/09/15 18:07:07 INFO mapred.LocalJobRunner: Waiting for map tasks
15/09/15 18:07:07 INFO mapred.LocalJobRunner: Starting task: attempt_local1207432264_0001_m_000000_0
```

```
15/09/15 18:07:07 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
15/09/15 18:07:07 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
15/09/15 18:07:07 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
15/09/15 18:07:07 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/gtumuluri/enronemail
15/09/15 18:07:07 INFO mapred.MapTask: numReduceTasks: 1
15/09/15 18:07:07 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
15/09/15 18:07:07 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
15/09/15 18:07:07 INFO mapred.MapTask: soft limit at 83886080
15/09/15 18:07:07 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
15/09/15 18:07:07 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
15/09/15 18:07:07 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$Map
15/09/15 18:07:07 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/gtumuluri/Documents/BerkeleyMIDS/S
15/09/15 18:07:07 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.ta
15/09/15 18:07:07 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce
15/09/15 18:07:07 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.m
15/09/15 18:07:07 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.j
15/09/15 18:07:07 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce
15/09/15 18:07:07 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use map
15/09/15 18:07:07 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce
15/09/15 18:07:07 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.jo
15/09/15 18:07:07 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.us
15/09/15 18:07:07 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapredu
15/09/15 18:07:07 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.t
15/09/15 18:07:07 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapr
15/09/15 18:07:07 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:07:07 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:07:07 INFO streaming.PipeMapRed: Records R/W=73/1
15/09/15 18:07:07 INFO streaming.PipeMapRed: R/W/S=100/10619/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:07:07 INFO streaming.PipeMapRed: MRErrorThread done
15/09/15 18:07:07 INFO streaming.PipeMapRed: mapRedFinished
15/09/15 18:07:07 INFO mapred.LocalJobRunner:
15/09/15 18:07:07 INFO mapred.MapTask: Starting flush of map output
15/09/15 18:07:07 INFO mapred.MapTask: Spilling map output
15/09/15 18:07:07 INFO mapred.MapTask: bufstart = 0; bufend = 249750; bufvoid = 104857600
15/09/15 18:07:07 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26083984(104335936); leng
15/09/15 18:07:07 INFO mapred.MapTask: Finished spill 0
15/09/15 18:07:07 INFO mapred.Task: Task:attempt_local1207432264_0001_m_000000_0 is done. And is in the p
15/09/15 18:07:07 INFO mapred.LocalJobRunner: Records R/W=73/1
15/09/15 18:07:07 INFO mapred.Task: Task 'attempt_local1207432264_0001_m_000000_0' done.
15/09/15 18:07:07 INFO mapred.LocalJobRunner: Finishing task: attempt_local1207432264_0001_m_000000_0
15/09/15 18:07:07 INFO mapred.LocalJobRunner: map task executor complete.
15/09/15 18:07:07 INFO mapred.LocalJobRunner: Waiting for reduce tasks
15/09/15 18:07:07 INFO mapred.LocalJobRunner: Starting task: attempt_local1207432264_0001_r_000000_0
15/09/15 18:07:07 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
15/09/15 18:07:07 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only
15/09/15 18:07:07 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
15/09/15 18:07:07 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task
15/09/15 18:07:07 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=334338464, maxSingleShuffleLi
15/09/15 18:07:07 INFO reduce.EventFetcher: attempt_local1207432264_0001_r_000000_0 Thread started: Event
15/09/15 18:07:07 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1
15/09/15 18:07:08 INFO reduce.InMemoryMapOutput: Read 314960 bytes from map-output for attempt_local1207
15/09/15 18:07:08 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 314960, inMemor
15/09/15 18:07:08 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
15/09/15 18:07:08 INFO mapred.LocalJobRunner: 1 / 1 copied.
```

```
15/09/15 18:07:08 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on
15/09/15 18:07:08 INFO mapred.Merger: Merging 1 sorted segments
15/09/15 18:07:08 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 3
15/09/15 18:07:08 INFO mapreduce.Job: Job job_local1207432264_0001 running in uber mode : false
15/09/15 18:07:08 INFO mapreduce.Job:  map 100% reduce 0%
15/09/15 18:07:08 INFO reduce.MergeManagerImpl: Merged 1 segments, 314960 bytes to disk to satisfy redu
15/09/15 18:07:08 INFO reduce.MergeManagerImpl: Merging 1 files, 314964 bytes from disk
15/09/15 18:07:08 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
15/09/15 18:07:08 INFO mapred.Merger: Merging 1 sorted segments
15/09/15 18:07:08 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 3
15/09/15 18:07:08 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 18:07:08 INFO streaming.PipeMapRed: PipeMapRed exec [/Users/gtumuluri/Documents/BerkeleyMIDS/S
15/09/15 18:07:08 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapredu
15/09/15 18:07:08 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduc
15/09/15 18:07:08 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:07:08 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:07:08 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:07:08 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:07:08 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 18:07:08 INFO streaming.PipeMapRed: MRErrorThread done
15/09/15 18:07:08 INFO streaming.PipeMapRed: Records R/W=32604/1
15/09/15 18:07:08 INFO streaming.PipeMapRed: mapRedFinished
15/09/15 18:07:08 INFO mapred.Task: Task:attempt_local1207432264_0001_r_000000_0 is done. And is in the p
15/09/15 18:07:08 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 18:07:08 INFO mapred.Task: Task attempt_local1207432264_0001_r_000000_0 is allowed to commit now
15/09/15 18:07:08 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1207432264_0001_r_0
15/09/15 18:07:08 INFO mapred.LocalJobRunner: Records R/W=32604/1 > reduce
15/09/15 18:07:08 INFO mapred.Task: Task 'attempt_local1207432264_0001_r_000000_0' done.
15/09/15 18:07:08 INFO mapred.LocalJobRunner: Finishing task: attempt_local1207432264_0001_r_000000_0
15/09/15 18:07:08 INFO mapred.LocalJobRunner: reduce task executor complete.
15/09/15 18:07:09 INFO mapreduce.Job:  map 100% reduce 100%
15/09/15 18:07:09 INFO mapreduce.Job: Job job_local1207432264_0001 completed successfully
15/09/15 18:07:09 INFO mapreduce.Job: Counters: 35
        File System Counters
                FILE: Number of bytes read=842034
                FILE: Number of bytes written=1744346
                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=2672
                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=32604
                Map output bytes=249750
                Map output materialized bytes=314964
                Input split bytes=106
                Combine input records=0
                Combine output records=0
                Reduce input groups=5744
                Reduce shuffle bytes=314964
```

```
                Reduce input records=32604
                Reduce output records=100
                Spilled Records=65208
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=9
                Total committed heap usage (bytes)=491782144
        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=2672
15/09/15 18:07:09 INFO streaming.StreamJob: Output directory: fullClassificationOutputExcludeInfreq
```

2.9.7 Problem 2.5b - Show Output of Full Naive Bayes Classifier REMOVE Infrequent Words

No words are passed as argument. Training and prediction uses full vocabulary AFTER excluding any words occurring 3 times or less in the documents.

```
In [285]: !hdfs dfs -cat fullClassificationOutputExcludeInfreq/part-00000
```

```
15/09/15 18:07:12 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
0001.1999-12-10.farmer        0        0
0001.1999-12-10.kaminski        0        0
0001.2000-01-17.beck        0        0
0001.2000-06-06.lokay        0        0
0001.2001-02-07.kitchen        0        0
0001.2001-04-02.williams        0        0
0002.1999-12-13.farmer        0        0
0002.2001-02-07.kitchen        0        0
0002.2001-05-25.SA_and_HP        1        1
0002.2003-12-18.GP        1        1
0002.2004-08-01.BG        1        1
0003.1999-12-10.kaminski        0        0
0003.1999-12-14.farmer        0        0
0003.2000-01-17.beck        0        0
0003.2001-02-08.kitchen        0        0
0003.2003-12-18.GP        1        1
0003.2004-08-01.BG        1        1
0004.1999-12-10.kaminski        0        0
0004.1999-12-14.farmer        0        0
0004.2001-04-02.williams        0        0
0004.2001-06-12.SA_and_HP        1        1
0004.2004-08-01.BG        1        1
0005.1999-12-12.kaminski        0        0
0005.1999-12-14.farmer        0        0
0005.2000-06-06.lokay        0        0
0005.2001-02-08.kitchen        0        0
```

```
0005.2001-06-23.SA_and_HP        1         1
0005.2003-12-18.GP        1         1
0006.1999-12-13.kaminski         0         0
0006.2001-02-08.kitchen         0         0
0006.2001-04-03.williams         0         0
0006.2001-06-25.SA_and_HP        1         1
0006.2003-12-18.GP        1         1
0006.2004-08-01.BG        1         1
0007.1999-12-13.kaminski         0         0
0007.1999-12-14.farmer         0         0
0007.2000-01-17.beck         0         0
0007.2001-02-09.kitchen         0         0
0007.2003-12-18.GP        1         1
0007.2004-08-01.BG        1         1
0008.2001-02-09.kitchen         0         0
0008.2001-06-12.SA_and_HP        1         1
0008.2001-06-25.SA_and_HP        1         1
0008.2003-12-18.GP        1         1
0008.2004-08-01.BG        1         1
0009.1999-12-13.kaminski         0         0
0009.1999-12-14.farmer         0         0
0009.2000-06-07.lokay         0         0
0009.2001-02-09.kitchen         0         0
0009.2001-06-26.SA_and_HP        1         1
0009.2003-12-18.GP        1         1
0010.1999-12-14.farmer         0         0
0010.1999-12-14.kaminski         0         0
0010.2001-02-09.kitchen         0         0
0010.2001-06-28.SA_and_HP        1         1
0010.2003-12-18.GP        1         0
0010.2004-08-01.BG        1         1
0011.1999-12-14.farmer         0         0
0011.2001-06-28.SA_and_HP        1         1
0011.2001-06-29.SA_and_HP        1         1
0011.2003-12-18.GP        1         1
0011.2004-08-01.BG        1         1
0012.1999-12-14.farmer         0         0
0012.1999-12-14.kaminski         0         0
0012.2000-01-17.beck         0         0
0012.2000-06-08.lokay         0         0
0012.2001-02-09.kitchen         0         0
0012.2003-12-19.GP        1         1
0013.1999-12-14.farmer         0         0
0013.1999-12-14.kaminski         0         0
0013.2001-04-03.williams         0         0
0013.2001-06-30.SA_and_HP        1         1
0013.2004-08-01.BG        1         1
0014.1999-12-14.kaminski         0         0
0014.1999-12-15.farmer         0         0
0014.2001-02-12.kitchen         0         0
0014.2001-07-04.SA_and_HP        1         1
0014.2003-12-19.GP        1         1
0014.2004-08-01.BG        1         1
0015.1999-12-14.kaminski         0         0
```

```
0015.1999-12-15.farmer        0        0
0015.2000-06-09.lokay         0        0
0015.2001-02-12.kitchen       0        0
0015.2001-07-05.SA_and_HP     1        1
0015.2003-12-19.GP            1        1
0016.1999-12-15.farmer        0        0
0016.2001-02-12.kitchen       0        0
0016.2001-07-05.SA_and_HP     1        1
0016.2001-07-06.SA_and_HP     1        1
0016.2003-12-19.GP            1        1
0016.2004-08-01.BG            1        1
0017.1999-12-14.kaminski      0        0
0017.2000-01-17.beck          0        0
0017.2001-04-03.williams      0        0
0017.2003-12-18.GP            1        1
0017.2004-08-01.BG            1        0
0017.2004-08-02.BG            1        1
0018.1999-12-14.kaminski      0        0
0018.2001-07-13.SA_and_HP     1        1
0018.2003-12-18.GP            1        1
```

2.10 Problem 1.6

2.10.1 Problem 1.6 - SKLearn Bernoulli and Multinomial Naive Bayes

In this case, we use the built-in SKLearn classifier training models and do prediction on the same data as above.

In [286]:
```
########################################
######## PROBLEM NUMBER 1.6 ##########
########################################

# SKLearn benchmark with NaiveBayes
from sklearn.naive_bayes import BernoulliNB
from sklearn.naive_bayes import MultinomialNB
from sklearn.feature_extraction.text import CountVectorizer
import pandas as pd

# Read the data set as a pandas dataframe and add a new column that
# combines the text of subject line and email body. Also, drop any
# rows that have NAs in key content and class columns.
enron = pd.read_csv('enronemail_1h.txt', sep = '\t', header = None)
enron = enron.dropna(subset = [1, 2, 3])
enron.loc[:, 'content'] = enron.loc[:, 2] + ' ' + enron.loc[:, 3]

# Extract columns into text content and labels. Remember, we will train
# and test on the same exact data - there is separate 'test' data.
train_labels = enron.loc[:, 1]
train_content = enron.loc[:, 'content']
test_labels = enron.loc[:, 1]
test_content = enron.loc[:, 'content']

# Transform text into features for training
count_vect = CountVectorizer()
train_features = count_vect.fit_transform(train_content)
```

```
        test_features = count_vect.transform(train_content)

        # Train a Bernoulli Naive Bayes model with defaults and measure prediction
        # accuracy on the same data. Print the output.
        bern = BernoulliNB()
        bern.fit(train_features, train_labels)
        predictions = bern.predict(test_features)
        accuracy = float(len([i for i, j in
                              zip(predictions, test_labels)
                              if i == j])) / len(test_labels)
        print "Bernoulli Naive Bayes Accuracy: " + str(round(accuracy, 2))

        # Train a Multinomial Naive Bayes model with defaults and measure prediction
        # accuracy on the same data. Print the output.
        mult = MultinomialNB()
        mult.fit(train_features, train_labels)
        predictions = mult.predict(test_features)
        accuracy = float(len([i for i, j in zip(predictions, test_labels) if i == j])) / len(test_labe
        print "Multinomial Naive Bayes Accuracy: " + str(round(accuracy, 2))
```

```
Bernoulli Naive Bayes Accuracy: 0.77
Multinomial Naive Bayes Accuracy: 1.0
```

2.10.2 Stop Yarn and HDFS

```
In [287]: !/usr/local/Cellar/hadoop/2.7.0/sbin/stop-yarn.sh
          !/usr/local/Cellar/hadoop/2.7.0/sbin/stop-dfs.sh
```

```
stopping yarn daemons
stopping resourcemanager
localhost: stopping nodemanager
no proxyserver to stop
15/09/15 18:07:45 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
Stopping namenodes on [localhost]
localhost: stopping namenode
localhost: stopping datanode
Stopping secondary namenodes [0.0.0.0]
0.0.0.0: stopping secondarynamenode
15/09/15 18:08:06 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
```

```
In [ ]:
```