DATSCIW261 ASSIGNMENT 2

MIDS UC Berkeley, Machine Learning at Scale

AUTHOR : Rajesh Thallam
EMAIL : rajesh.thallam@ischool.berkeley.edu
WEEK : 2
DATE : 15-Sep-15

HW2.0

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

What is a race condition in the context of parallel computation? Give an example.
Race condition is *consequence of simultaneous access of a shared data resource when two or more asynchronous (parallel) threads attempt to access and modify a shared resource*. Since the application is unknown of the order in which the threads access and modify the resource, the output is ambiguous. One of the ways to avoid the race condition is using mutex which basically allows for acquiring and releasing lock on the shared resource.
One of the common example I have encountered is multiple threads attempting to increment value of global variable. Imagine a global variable p accessed by two threads A and B to increment value by +1 using ++ (increment) operation. Increment operator performs three steps (i) read variable (ii) increment value and (iii) store variable. So increment is not an atomic operation.

```
# global variable p with current value
p = 18

# THREAD A
p++
# value will be 19

# THREAD B at same time as THREAD A or just little after
p++
# it still sees p as 18 and attempts to increment p to 19
```

At the end of the operation, we see that the value of p in both threads is 19 instead of 19 (A) and 20 (B).

What is MapReduce? How does it differ from Hadoop?
MapReduce is a functional programming design pattern accepting functions as arguments. This programming paradigm allows parallel data processing of embarrassingly parallel data problems. The map part of the progam chunks incoming data in parallel as defined by the number of mappers. Then the reduce part folds or combines the results of mappers to generate final result of the problem.

Hadoop is a framework built on MapReduce programming paradigm (data processing) and Hadoop file system (data storage) to solve the large data set problems in an embarrassingly parallel way by moving MapReduce program near to the data storage to process the data. The framework provides a distributed data handling capability combined with distributed computation by concealing system level details to the programmer. The framework also accomodates necessary fault tolerance and resiliency built into the application.

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

In [1]:
```
# this simple example calculates word counts in given strings
import itertools

# define mapper to split word and count as 1
def mapper(key, value):
    return [(word,1) for word in value.split()]

# define reducer to sum counts of a given word
def reducer(key, values):
    return (key, sum(values))

# tie map and reduce phases
def map_reduce(lines, mapper,reducer):
    map_out = []

    # call mapper
    for (key,value) in lines.items():
        map_out.extend(mapper(key, value))

    # partition mapper output
    groups = {}
    for key, group in itertools.groupby(sorted(map_out), lambda x: x[0]):
        groups[key] = list([y for x, y in group])

    # reduce phase to output counts
    return [reducer(key, groups[key]) for key in groups]

# feed input and call map reduce
lines = {}
lines["1"] = "foo bar foo bar foo bar foo foo foo bax lines line"
lines["2"] = "hello world this is foo bar"
map_reduce(lines, mapper, reducer)
```

Out[1]:
```
[('bar', 4),
 ('this', 1),
 ('is', 1),
 ('lines', 1),
 ('bax', 1),
 ('world', 1),
 ('line', 1),
 ('foo', 7),
 ('hello', 1)]
```

Preparation for HW2_*

In [2]:
```
# stop hadoop
!ssh hduser@rtubuntu /usr/local/hadoop/sbin/stop-yarn.sh
!ssh hduser@rtubuntu /usr/local/hadoop/sbin/stop-dfs.sh
```

```
stopping yarn daemons
no resourcemanager to stop
localhost: no nodemanager to stop
no proxyserver to stop
Stopping namenodes on [localhost]
localhost: no namenode to stop
localhost: no datanode to stop
Stopping secondary namenodes [0.0.0.0]
0.0.0.0: stopping secondarynamenode
```

In [3]:
```
# start hadoop
!ssh hduser@rtubuntu /usr/local/hadoop/sbin/start-yarn.sh
!ssh hduser@rtubuntu /usr/local/hadoop/sbin/start-dfs.sh
```

```
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-hduser-resourcemanager-rtubuntu.out
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-hduser-nodemanager-rtubuntu.out
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-hduser-namenode-rtubuntu.out
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hduser-datanode-rtubuntu.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-hduser-secondarynamenode-rtubuntu.out
```

In [4]:
```
# create necessary directories
!hdfs dfs -mkdir /hw2
```

```
15/09/15 01:30:15 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
classes where applicable
mkdir: `/hw2': File exists
```

HW2.1

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

Write code to generate N random records of the form (integer, "NA"). Let N = 10,000.
Write the python Hadoop streaming map-reduce job to perform this sort.

What happens if you have multiple reducers? Do you need additional steps? Explain.
When there are multiple reducers, each reducer will sort the data chunks sent to each reducer from the partition phase of mapreduce. The default partitioning uses hash code mod number of reducers i.e. if there are 5 reducers then there will be 5 output files, each sorted with overlapping ranges. In order to avoid the overlapping ranges we either need one reducer or make the partitioner more aware of the nature of the keys. For example, make partitioner to direct all of the keys within a range (say 1 to 2000) to the same partition. Thus, there will be multiple files in the output but all the files will have sorted data without overlapping ranges.

Generate Input

`gen_in_hw2_1.py` script generates input file for the mapreduce program to generate 10000 random numbers.

In [5]:
```
%%writefile gen_in_hw2_1.py
#!/usr/bin/python
import random

N = 10000
# used random.sample to avoid replacement of same numbers
r = random.sample(range(N), N)

for n in r:
    print "{0} {1}".format(n, "NA")
```

```
Overwriting gen_in_hw2_1.py
```

In [6]:
```
!./gen_in_hw2_1.py > hw2_1.txt
!head hw2_1.txt
```

```
9662 NA
4030 NA
6587 NA
9595 NA
9528 NA
6418 NA
2197 NA
5853 NA
9532 NA
4327 NA
```

Mapper

This is an identity mapper as hadoop streaming needs atleast one mapper. This mapper just prints the input

In [26]:
```
%%writefile mapper.py
#!/usr/bin/env python
import sys

for line in sys.stdin:
    print "%s" % (line.strip())
```

```
Overwriting mapper.py
```

Reducer

This is an identity reducer as the intention is sort the mapper output as is and the shuffle/sort phase is handled by the hadoop streaming (or hadoop framework)

In [27]:
```
%%writefile reducer.py
#!/usr/bin/env python
import sys

for line in sys.stdin:
    print "%s" % (line.strip())
```

```
Overwriting reducer.py
```

Preparing to run the job

In [9]:
```
# Use chmod for permissions
!chmod a+x mapper.py
!chmod a+x reducer.py
```

In [ ]:
```
!hdfs dfs -mkdir /hw2/hw2_1
!hdfs dfs -mkdir /hw2/hw2_1/src
!hdfs dfs -put ./hw2_1.txt /hw2/hw2_1/src
```

Driver Function

Driver function calls the hadoop streaming job after purging previously generated target files (to avoid the `'File Already Exists'` error). Few points to notice

- **used KeyFieldBasedComparator and key.comparator.options to sort the data from the mapper. This is provided by the Hadoop Streaming jar**
- **number of mappers is set to 10**
- **number of reducers is set to 1**
- **output first few lines from the output of the job**

In [28]:
```
# HW 2.1: execute hadoop streaming job to generate and sort
#         10K random integers
def hw2_1():
    # cleanup target directory
    !hdfs dfs -rm -R /hw2/hw2_1/tgt

    !echo "sample input data"
    !hdfs dfs -cat /hw2/hw2_1/src/hw2_1.txt | head

    # run map reduce job
    !hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar \
    -D mapred.output.key.comparator.class=org.apache.hadoop.mapred.lib.KeyFieldBasedComparator \
    -D mapred.text.key.comparator.options=-k1,1n \
    -Dmapreduce.job.maps=10 \
    -Dmapreduce.job.reduces=1 \
    -files mapper.py,reducer.py \
    -mapper mapper.py \
    -reducer reducer.py \
    -input /hw2/hw2_1/src/hw2_1.txt \
    -output /hw2/hw2_1/tgt

    print "\n"
    !echo "partial output data"
    !hdfs dfs -cat /hw2/hw2_1/tgt/part-00000 | head

hw2_1()
```

```
15/09/15 01:42:16 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
classes where applicable
15/09/15 01:42:17 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interva
l = 0 minutes.
Deleted /hw2/hw2_1/tgt
sample input data
15/09/15 01:42:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
classes where applicable
1956 NA
2198 NA
2266 NA
2762 NA
6692 NA
1838 NA
953 NA
1389 NA
4361 NA
9687 NA
cat: Unable to write to output stream.
15/09/15 01:42:22 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
classes where applicable
15/09/15 01:42:22 INFO Configuration.deprecation: mapred.output.key.comparator.class is deprecated. Instead, use mapreduce
.job.output.key.comparator.class
15/09/15 01:42:22 INFO Configuration.deprecation: mapred.text.key.comparator.options is deprecated. Instead, use mapreduce
.partition.keycomparator.options
15/09/15 01:42:23 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
15/09/15 01:42:23 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
15/09/15 01:42:23 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already ini
tialized
15/09/15 01:42:23 INFO mapred.FileInputFormat: Total input paths to process : 1
15/09/15 01:42:23 INFO mapreduce.JobSubmitter: number of splits:1
15/09/15 01:42:24 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1069108839_0001
15/09/15 01:42:24 INFO mapred.LocalDistributedCacheManager: Localized file:/media/sf_shared/GitHub/MIDS-W261-MACHINE-LEARN
ING-AT-SCALE/week2/hw2/mapper.py as file:/app/hadoop/tmp/mapred/local/1442306544393/mapper.py
15/09/15 01:42:24 INFO mapred.LocalDistributedCacheManager: Localized file:/media/sf_shared/GitHub/MIDS-W261-MACHINE-LEARN
ING-AT-SCALE/week2/hw2/reducer.py as file:/app/hadoop/tmp/mapred/local/1442306544394/reducer.py
15/09/15 01:42:24 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
15/09/15 01:42:24 INFO mapreduce.Job: Running job: job_local1069108839_0001
15/09/15 01:42:24 INFO mapred.LocalJobRunner: OutputCommitter set in config null
```

```
15/09/15 01:42:24 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
15/09/15 01:42:24 INFO mapred.LocalJobRunner: Waiting for map tasks
15/09/15 01:42:24 INFO mapred.LocalJobRunner: Starting task: attempt_local1069108839_0001_m_000000_0
15/09/15 01:42:25 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
15/09/15 01:42:25 INFO mapred.MapTask: Processing split: hdfs://localhost:54310/hw2/hw2_1/src/hw2_1.txt:0+78890
15/09/15 01:42:25 INFO mapred.MapTask: numReduceTasks: 1
15/09/15 01:42:25 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
15/09/15 01:42:25 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
15/09/15 01:42:25 INFO mapred.MapTask: soft limit at 83886080
15/09/15 01:42:25 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
15/09/15 01:42:25 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
15/09/15 01:42:25 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
15/09/15 01:42:25 INFO streaming.PipeMapRed: PipeMapRed exec [/media/sf_shared/GitHub/MIDS-W261-MACHINE-LEARNING-AT-SCALE/
week2/hw2/./mapper.py]
15/09/15 01:42:25 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
15/09/15 01:42:25 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
15/09/15 01:42:25 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
15/09/15 01:42:25 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
15/09/15 01:42:25 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
15/09/15 01:42:25 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output
.dir
15/09/15 01:42:25 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
15/09/15 01:42:25 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
15/09/15 01:42:25 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
15/09/15 01:42:25 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
15/09/15 01:42:25 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
15/09/15 01:42:25 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partiti
on
15/09/15 01:42:25 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:42:25 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:42:25 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:42:25 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:42:25 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:42:25 INFO streaming.PipeMapRed: Records R/W=10000/1
15/09/15 01:42:25 INFO streaming.PipeMapRed: MRErrorThread done
15/09/15 01:42:25 INFO streaming.PipeMapRed: mapRedFinished
15/09/15 01:42:25 INFO mapred.LocalJobRunner:
15/09/15 01:42:25 INFO mapred.MapTask: Starting flush of map output
15/09/15 01:42:25 INFO mapred.MapTask: Spilling map output
15/09/15 01:42:25 INFO mapred.MapTask: bufstart = 0; bufend = 88890; bufvoid = 104857600
15/09/15 01:42:25 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26174400(104697600); length = 39997/6553600
15/09/15 01:42:25 INFO mapreduce.Job: Job job_local1069108839_0001 running in uber mode : false
15/09/15 01:42:25 INFO mapreduce.Job:  map 0% reduce 0%
15/09/15 01:42:26 INFO mapred.MapTask: Finished spill 0
15/09/15 01:42:26 INFO mapred.Task: Task:attempt_local1069108839_0001_m_000000_0 is done. And is in the process of committ
ing
15/09/15 01:42:26 INFO mapred.LocalJobRunner: Records R/W=10000/1
15/09/15 01:42:26 INFO mapred.Task: Task 'attempt_local1069108839_0001_m_000000_0' done.
15/09/15 01:42:26 INFO mapred.LocalJobRunner: Finishing task: attempt_local1069108839_0001_m_000000_0
15/09/15 01:42:26 INFO mapred.LocalJobRunner: map task executor complete.
15/09/15 01:42:26 INFO mapred.LocalJobRunner: Waiting for reduce tasks
15/09/15 01:42:26 INFO mapred.LocalJobRunner: Starting task: attempt_local1069108839_0001_r_000000_0
15/09/15 01:42:26 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
15/09/15 01:42:26 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@7e9
bdd4f
15/09/15 01:42:26 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=363285696, maxSingleShuffleLimit=90821424, merg
eThreshold=239768576, ioSortFactor=10, memToMemMergeOutputsThreshold=10
15/09/15 01:42:26 INFO reduce.EventFetcher: attempt_local1069108839_0001_r_000000_0 Thread started: EventFetcher for fetch
ing Map Completion Events
15/09/15 01:42:26 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1069108839_0001_m_0
00000_0 decomp: 108892 len: 108896 to MEMORY
15/09/15 01:42:26 INFO reduce.InMemoryMapOutput: Read 108892 bytes from map-output for attempt_local1069108839_0001_m_0000
00_0
15/09/15 01:42:26 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 108892, inMemoryMapOutputs.size()
-> 1, commitMemory -> 0, usedMemory ->108892
15/09/15 01:42:26 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
15/09/15 01:42:26 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 01:42:26 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
15/09/15 01:42:26 INFO mapred.Merger: Merging 1 sorted segments
15/09/15 01:42:26 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 108885 bytes
15/09/15 01:42:26 INFO reduce.MergeManagerImpl: Merged 1 segments, 108892 bytes to disk to satisfy reduce memory limit
15/09/15 01:42:26 INFO reduce.MergeManagerImpl: Merging 1 files, 108896 bytes from disk
15/09/15 01:42:26 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
15/09/15 01:42:26 INFO mapred.Merger: Merging 1 sorted segments
15/09/15 01:42:26 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 108885 bytes
15/09/15 01:42:26 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 01:42:26 INFO streaming.PipeMapRed: PipeMapRed exec [/media/sf_shared/GitHub/MIDS-W261-MACHINE-LEARNING-AT-SCALE/
week2/hw2/./reducer.py]
15/09/15 01:42:26 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
15/09/15 01:42:26 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
15/09/15 01:42:26 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:42:26 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:42:26 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:42:26 INFO mapreduce.Job:  map 100% reduce 0%
15/09/15 01:42:26 INFO streaming.PipeMapRed: R/W/S=1000/0/0 in:NA [rec/s] out:NA [rec/s]
```

```
15/09/15 01:42:26 INFO streaming.PipeMapRed: R/W/S=10000/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:42:26 INFO streaming.PipeMapRed: Records R/W=10000/1
15/09/15 01:42:27 INFO streaming.PipeMapRed: MRErrorThread done
15/09/15 01:42:27 INFO streaming.PipeMapRed: mapRedFinished
15/09/15 01:42:27 INFO mapred.Task: Task:attempt_local1069108839_0001_r_000000_0 is done. And is in the process of committ
ing
15/09/15 01:42:27 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 01:42:27 INFO mapred.Task: Task attempt_local1069108839_0001_r_000000_0 is allowed to commit now
15/09/15 01:42:27 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1069108839_0001_r_000000_0' to hdfs:
//localhost:54310/hw2/hw2_1/tgt/_temporary/0/task_local1069108839_0001_r_000000
15/09/15 01:42:27 INFO mapred.LocalJobRunner: Records R/W=10000/1 > reduce
15/09/15 01:42:27 INFO mapred.Task: Task 'attempt_local1069108839_0001_r_000000_0' done.
15/09/15 01:42:27 INFO mapred.LocalJobRunner: Finishing task: attempt_local1069108839_0001_r_000000_0
15/09/15 01:42:27 INFO mapred.LocalJobRunner: reduce task executor complete.
15/09/15 01:42:27 INFO mapreduce.Job:  map 100% reduce 100%
15/09/15 01:42:27 INFO mapreduce.Job: Job job_local1069108839_0001 completed successfully
15/09/15 01:42:27 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=428454
                FILE: Number of bytes written=1051908
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=157780
                HDFS: Number of bytes written=88890
                HDFS: Number of read operations=13
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Map-Reduce Framework
                Map input records=10000
                Map output records=10000
                Map output bytes=88890
                Map output materialized bytes=108896
                Input split bytes=98
                Combine input records=0
                Combine output records=0
                Reduce input groups=10000
                Reduce shuffle bytes=108896
                Reduce input records=10000
                Reduce output records=10000
                Spilled Records=20000
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=55
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=335683584
        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=78890
        File Output Format Counters
                Bytes Written=88890
15/09/15 01:42:27 INFO streaming.StreamJob: Output directory: /hw2/hw2_1/tgt


partial output data
15/09/15 01:42:30 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
classes where applicable
0 NA
1 NA
2 NA
3 NA
4 NA
5 NA
6 NA
7 NA
8 NA
9 NA
cat: Unable to write to output stream.
```

HW2.2

Using the Enron data from HW1 and Hadoop MapReduce streaming, write mapper/reducer pair that will determine the number of occurrences of a single, user-specified word. Examine the word "assistance" and report your results. To do so, make sure that

- mapper.py counts all occurrences of a single word, and
- reducer.py collates the counts of the single word.

Assumptions

1. **For this problem, both email body and subject is considered for classification**
2. **Removed punctuations, special characters from email content**

Mapper

In [29]:
```
%%writefile mapper.py
#!/usr/bin/python
import traceback
import sys
import re

# read input parameters
find_word = sys.argv[1]

try:
    for email in sys.stdin:
        # split email by tab (\t)
        mail = email.split('\t')

        # handle missing email content
        if len(mail) == 3:
            mail.append(mail[2])
            mail[2] = ""
        assert len(mail) == 4

        # email id
        email_id = mail[0]
        # email content - remove special characters and punctuations
        content = re.sub('[^A-Za-z0-9\s]+', '', mail[2] + " " +  mail[3])

        # find word with counts
        for word in content.split():
            if word == find_word:
                print '{}\t{}'.format(word, 1)
except Exception:
    traceback.print_exc()
```

```
Overwriting mapper.py
```

In [30]:
```
%%writefile reducer.py
#!/usr/bin/python
import traceback
import sys

try:
    word_counts = {}

    # read each map output
    for line in sys.stdin:
        # parse mapper output
        word, count = line.strip('\n').split('\t')

        try:
            word_counts[word] += int(count)
        except:
            word_counts[word] = int(count)

    print word_counts
except Exception:
    traceback.print_exc()
```

```
Overwriting reducer.py
```

Preparing to run the job

In [ ]:
```
# move source file to hdfs
!hdfs dfs -mkdir /hw2/hw2_2
!hdfs dfs -mkdir /hw2/hw2_2/src
!hdfs dfs -put ./enronemail_1h.txt /hw2/hw2_2/src
```

Driver Function

In [15]:
```
# HW 2.2  Mapper/reducer pair to determine the number of occurrences
#         of a single, user-specified word

def hw2_2(word):
    # cleanup target directory
    !hdfs dfs -rm -R /hw2/hw2_2/tgt

    # run map reduce job
    !hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar \
    -Dmapreduce.job.maps=10 \
    -Dmapreduce.job.reduces=1 \
    -files mapper.py,reducer.py \
    -mapper 'mapper.py {word}' \
    -reducer reducer.py \
    -input /hw2/hw2_2/src/enronemail_1h.txt \
    -output /hw2/hw2_2/tgt

    print "\nOUTPUT"
    # display count on the screen
    print "output from mapper/reducer to determine the number of occurrences of word assistance"
    !hdfs dfs -cat /hw2/hw2_2/tgt/part-00000

    # CROSSCHECK
    print "\nCROSSCHECK"
    print "output from command line mapper/reducer"
    ! grep assistance enronemail_1h.txt | awk -F'\t' '{print $3, $4}' | grep -o assistance | wc -l

hw2_2("assistance")
```

```
15/09/15 01:30:45 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
classes where applicable
15/09/15 01:30:46 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interva
l = 0 minutes.
Deleted /hw2/hw2_2/tgt
15/09/15 01:30:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
classes where applicable
15/09/15 01:30:49 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
15/09/15 01:30:49 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
15/09/15 01:30:49 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already ini
tialized
15/09/15 01:30:49 INFO mapred.FileInputFormat: Total input paths to process : 1
15/09/15 01:30:50 INFO mapreduce.JobSubmitter: number of splits:1
15/09/15 01:30:50 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1125170023_0001
15/09/15 01:30:51 INFO mapred.LocalDistributedCacheManager: Localized file:/media/sf_shared/GitHub/MIDS-W261-MACHINE-LEARN
ING-AT-SCALE/week2/hw2/mapper.py as file:/app/hadoop/tmp/mapred/local/1442305850488/mapper.py
15/09/15 01:30:51 INFO mapred.LocalDistributedCacheManager: Localized file:/media/sf_shared/GitHub/MIDS-W261-MACHINE-LEARN
ING-AT-SCALE/week2/hw2/reducer.py as file:/app/hadoop/tmp/mapred/local/1442305850489/reducer.py
15/09/15 01:30:51 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
15/09/15 01:30:51 INFO mapreduce.Job: Running job: job_local1125170023_0001
15/09/15 01:30:51 INFO mapred.LocalJobRunner: OutputCommitter set in config null
15/09/15 01:30:51 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
15/09/15 01:30:51 INFO mapred.LocalJobRunner: Waiting for map tasks
15/09/15 01:30:51 INFO mapred.LocalJobRunner: Starting task: attempt_local1125170023_0001_m_000000_0
15/09/15 01:30:51 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
15/09/15 01:30:51 INFO mapred.MapTask: Processing split: hdfs://localhost:54310/hw2/hw2_2/src/enronemail_1h.txt:0+203979
15/09/15 01:30:51 INFO mapred.MapTask: numReduceTasks: 1
15/09/15 01:30:51 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
15/09/15 01:30:51 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
15/09/15 01:30:51 INFO mapred.MapTask: soft limit at 83886080
15/09/15 01:30:51 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
15/09/15 01:30:51 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
15/09/15 01:30:51 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
15/09/15 01:30:51 INFO streaming.PipeMapRed: PipeMapRed exec [/media/sf_shared/GitHub/MIDS-W261-MACHINE-LEARNING-AT-SCALE/
week2/hw2/./mapper.py, assistance]
15/09/15 01:30:51 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
15/09/15 01:30:51 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
15/09/15 01:30:51 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
15/09/15 01:30:51 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
15/09/15 01:30:51 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
15/09/15 01:30:51 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output
.dir
15/09/15 01:30:51 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
15/09/15 01:30:51 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
15/09/15 01:30:51 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
15/09/15 01:30:51 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
15/09/15 01:30:51 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
```

```
15/09/15 01:30:51 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partiti
on
15/09/15 01:30:51 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:30:51 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:30:51 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:30:51 INFO streaming.PipeMapRed: Records R/W=100/1
15/09/15 01:30:51 INFO streaming.PipeMapRed: MRErrorThread done
15/09/15 01:30:51 INFO streaming.PipeMapRed: mapRedFinished
15/09/15 01:30:52 INFO mapred.LocalJobRunner:
15/09/15 01:30:52 INFO mapred.MapTask: Starting flush of map output
15/09/15 01:30:52 INFO mapred.MapTask: Spilling map output
15/09/15 01:30:52 INFO mapred.MapTask: bufstart = 0; bufend = 130; bufvoid = 104857600
15/09/15 01:30:52 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214360(104857440); length = 37/6553600
15/09/15 01:30:52 INFO mapred.MapTask: Finished spill 0
15/09/15 01:30:52 INFO mapred.Task: Task:attempt_local1125170023_0001_m_000000_0 is done. And is in the process of committ
ing
15/09/15 01:30:52 INFO mapred.LocalJobRunner: Records R/W=100/1
15/09/15 01:30:52 INFO mapred.Task: Task 'attempt_local1125170023_0001_m_000000_0' done.
15/09/15 01:30:52 INFO mapred.LocalJobRunner: Finishing task: attempt_local1125170023_0001_m_000000_0
15/09/15 01:30:52 INFO mapred.LocalJobRunner: map task executor complete.
15/09/15 01:30:52 INFO mapred.LocalJobRunner: Waiting for reduce tasks
15/09/15 01:30:52 INFO mapred.LocalJobRunner: Starting task: attempt_local1125170023_0001_r_000000_0
15/09/15 01:30:52 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
15/09/15 01:30:52 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@76f
444bc
15/09/15 01:30:52 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=363285696, maxSingleShuffleLimit=90821424, merg
eThreshold=239768576, ioSortFactor=10, memToMemMergeOutputsThreshold=10
15/09/15 01:30:52 INFO reduce.EventFetcher: attempt_local1125170023_0001_r_000000_0 Thread started: EventFetcher for fetch
ing Map Completion Events
15/09/15 01:30:52 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1125170023_0001_m_0
00000_0 decomp: 152 len: 156 to MEMORY
15/09/15 01:30:52 INFO reduce.InMemoryMapOutput: Read 152 bytes from map-output for attempt_local1125170023_0001_m_000000_
0
15/09/15 01:30:52 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 152, inMemoryMapOutputs.size() ->
1, commitMemory -> 0, usedMemory ->152
15/09/15 01:30:52 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
15/09/15 01:30:52 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 01:30:52 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
15/09/15 01:30:52 INFO mapred.Merger: Merging 1 sorted segments
15/09/15 01:30:52 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 139 bytes
15/09/15 01:30:52 INFO mapreduce.Job: Job job_local1125170023_0001 running in uber mode : false
15/09/15 01:30:52 INFO mapreduce.Job:  map 100% reduce 0%
15/09/15 01:30:52 INFO reduce.MergeManagerImpl: Merged 1 segments, 152 bytes to disk to satisfy reduce memory limit
15/09/15 01:30:52 INFO reduce.MergeManagerImpl: Merging 1 files, 156 bytes from disk
15/09/15 01:30:52 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
15/09/15 01:30:52 INFO mapred.Merger: Merging 1 sorted segments
15/09/15 01:30:52 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 139 bytes
15/09/15 01:30:52 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 01:30:52 INFO streaming.PipeMapRed: PipeMapRed exec [/media/sf_shared/GitHub/MIDS-W261-MACHINE-LEARNING-AT-SCALE/
week2/hw2/./reducer.py]
15/09/15 01:30:52 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
15/09/15 01:30:52 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
15/09/15 01:30:52 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:30:52 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:30:52 INFO streaming.PipeMapRed: Records R/W=10/1
15/09/15 01:30:52 INFO streaming.PipeMapRed: MRErrorThread done
15/09/15 01:30:52 INFO streaming.PipeMapRed: mapRedFinished
15/09/15 01:30:53 INFO mapred.Task: Task:attempt_local1125170023_0001_r_000000_0 is done. And is in the process of committ
ing
15/09/15 01:30:53 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 01:30:53 INFO mapred.Task: Task attempt_local1125170023_0001_r_000000_0 is allowed to commit now
15/09/15 01:30:53 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1125170023_0001_r_000000_0' to hdfs:
//localhost:54310/hw2/hw2_2/tgt/_temporary/0/task_local1125170023_0001_r_000000
15/09/15 01:30:53 INFO mapred.LocalJobRunner: Records R/W=10/1 > reduce
15/09/15 01:30:53 INFO mapred.Task: Task 'attempt_local1125170023_0001_r_000000_0' done.
15/09/15 01:30:53 INFO mapred.LocalJobRunner: Finishing task: attempt_local1125170023_0001_r_000000_0
15/09/15 01:30:53 INFO mapred.LocalJobRunner: reduce task executor complete.
15/09/15 01:30:54 INFO mapreduce.Job:  map 100% reduce 100%
15/09/15 01:30:54 INFO mapreduce.Job: Job job_local1125170023_0001 completed successfully
15/09/15 01:30:54 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=212964
                FILE: Number of bytes written=725476
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=407958
                HDFS: Number of bytes written=20
                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=10
                Map output bytes=130
```

```
                Map output materialized bytes=156
                Input split bytes=106
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=156
                Reduce input records=10
                Reduce output records=1
                Spilled Records=20
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=51
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=335683584
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=203979
        File Output Format Counters
                Bytes Written=20
15/09/15 01:30:54 INFO streaming.StreamJob: Output directory: /hw2/hw2_2/tgt

OUTPUT
output from mapper/reducer to determine the number of occurrences of word assistance
15/09/15 01:30:56 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
classes where applicable
{'assistance': 10}

CROSSCHECK
output from command line mapper/reducer
10
```

HW2.3

Using the Enron data from HW1 and Hadoop MapReduce that will classify the email messages by a single, user-specified word. Examine the word “assistance” and report your results. To do so, make sure that:

- mapper.py
- reducer.py that performs a single word multinomial Naive Bayes classification

Assumptions

1. **Based on the instructions on LMS, only email body is considered for classification**
2. **The mapper takes care of classifiation based on user specified single word, multiple words or all words (*)**
3. **The reducer would require additional logic to handle special requirements when all words are used in the classifier**

Mapper

I chose mapper output to contain following fields for eeah email in the input data set

- **email_id (as key)**
- **spam/ham indicator**
- **total number of words in eah email**
- **number of ocurrences of each word in the vocab**

In [16]:
```
%%writefile mapper.py
#!/usr/bin/python
import traceback
import sys
import re

from collections import Counter

# read input parameters
find_words = sys.argv[1:]

try:
    search_all = 0

    # custom logic to handle all words (*)
    if find_words[0] == "*":
        search_all = 1
        word_list = []
    else:
        word_list = find_words

    for email in sys.stdin:
        # split email by tab (\t)
        mail = email.split('\t')

        # handle missing email content
        if len(mail) == 3:
            mail.append(mail[2])
            mail[2] = ""
        assert len(mail) == 4

        # email id
        email_id = mail[0]
        # spam/ham binary indicator
        is_spam = mail[1]
        # email content - remove special characters and punctuations
        #content = re.sub('[^A-Za-z0-9\s]+', '', mail[2] + " " +  mail[3])
        content = re.sub('[^A-Za-z0-9\s]+', '', mail[3])
        # count number of words
        content_wc = len(content.split())

        # find words with counts - works for single word or list of words
        # custom logic to handle all words (*)
        if search_all == 1:
            hits = Counter(content.split())
        else:
            find_words = re.compile("|".join(r"\b%s\b" % w for w in word_list))
            hits = Counter(re.findall(find_words, content))

        hits = {k: v for k, v in hits.iteritems()}

        # emit tuple delimited by |
        # (email id, spam ind, content word count, word hit counts)
        print "{} | {} | {} | {}".format(email_id, is_spam, content_wc, hits)
except Exception:
    traceback.print_exc()
```

```
Overwriting mapper.py
```

Reducer

Reducer does all the magic of training the classifier and predictions. The program preserves the output of mappers as a list after reading from standard in to use the mapper output as input for training and prediction. Based on the search term the program dynamically sets the vocabulary size. The output of the reducer is each email id with actual spam/ham indicator with prediction followed by accuracy.

NOTE Even if a search term is not available in the training data set, vocabulatory includes the missing search term for calculations during Laplace smoothing.

In [17]:
```
%%writefile reducer.py
#!/usr/bin/python
import traceback
import math
import sys
import ast

from collections import Counter

# read input parameters
find_words = sys.argv[1:]

# vocab
vocab = find_words

try:
    spam_count = 0
```

```
ham_count = 0
spam_all_wc = 0
ham_all_wc = 0
spam_term_wc = {}
ham_term_wc = {}
pr_word_given_spam = {}
pr_word_given_ham = {}

# read each mapper output to loop during the prediction phase
# after training the model
map_output = []
for line in sys.stdin:
    map_output.append(line)

for email in map_output:
    # parse mapper output
    mail = email.split(" | ")
    # read spam/ham indicator, content word count,
    is_spam = int(mail[1])
    content_wc = int(mail[2])
    hits = ast.literal_eval(mail[3])

    # capture counts required for naive bayes probabilities
    if is_spam:
        # spam mail count
        spam_count += 1
        # term count when spam
        spam_term_wc = dict(Counter(hits) + Counter(spam_term_wc))
        # all word count when spam
        spam_all_wc += content_wc
    else:
        # ham email count
        ham_count += 1
        # term count when ham
        ham_term_wc = dict(Counter(hits) + Counter(ham_term_wc))
        # all word count when ham
        ham_all_wc += content_wc

vocab = dict(Counter(vocab) + Counter(spam_term_wc) + Counter(ham_term_wc))
V = len(vocab) * 1.0
print "vocab size = {}".format(V)

# calculate priors
pr_spam_prior = (1.0 * spam_count) / (spam_count + ham_count)
pr_ham_prior = (1.0 - pr_spam_prior)
pr_spam_prior = math.log10(pr_spam_prior)
pr_ham_prior = math.log10(pr_ham_prior)

# calculate conditional probabilites with laplace smoothing = 1
# pr_word_given_class = ( count(w, c) + 1 ) / (count(c) + 1 * |V|)
for word in vocab:
    pr_word_given_spam[word] = math.log10((spam_term_wc.get(word, 0) + 1.0) / (spam_all_wc + V))
    pr_word_given_ham[word] = math.log10((ham_term_wc.get(word, 0) + 1.0) / (ham_all_wc + V))

print "/*log probabilities*/"
print "pr_spam_prior = {}".format(pr_spam_prior)
print "pr_ham_prior = {}".format(pr_ham_prior)

print "\n"
print "{0: <50} | {1} | {2}".format("ID", "TRUTH", "CLASS")
print "{0: <50}-+-{1}-+-{2}".format("-" * 50, "-" * 7, "-" * 10)

# spam/ham prediction using Multinomial Naive Bayes priors and conditional probabilities
accuracy = []

for email in map_output:
    # initialize
    word_count = 0
    pred_is_spam = 0
    pr_spam = pr_spam_prior
    pr_ham = pr_ham_prior

    # parse mapper output
    mail = email.split(" | ")
    email_id = mail[0]
    is_spam = int(mail[1])
    hits = ast.literal_eval(mail[3])

    # number of search words
    word_count = sum(hits.values())

    # probability for each class for a given email
    # argmax [ log P(C) + sum( P(Wi|C) ) ]
    for word in vocab:
        pr_spam += (pr_word_given_spam.get(word, 0) * hits.get(word, 0))
        pr_ham += (pr_word_given_ham.get(word, 0) * hits.get(word, 0))
```

```
        # predict based on maximum likelihood
        if pr_spam > pr_ham:
            pred_is_spam = 1

        # calculate accuracy
        accuracy.append(pred_is_spam==is_spam)

        print '{0:<50} | {1:<7} | {2:<10}'.format(email_id, is_spam, pred_is_spam)

    print "\n"
    print "/*accuracy*/"
    print "accuracy = {:.2f}".format(sum(accuracy) / float(len(accuracy)))

except Exception:
    traceback.print_exc()
```

```
Overwriting reducer.py
```

Preparing to run the job

In [ ]:
```
!hdfs dfs -mkdir /hw2/hw2_3
```

Driver Function

In [19]:
```
# HW 2.3  Mapper/reducer pair to classify the email messages by a single,
#         user-specified word using the Naive Bayes Formulation
def hw2_3(word):
    # cleanup target directory
    !hdfs dfs -rm -R /hw2/hw2_3/tgt

    # run map reduce job
    !hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar \
    -Dmapreduce.job.maps=10 \
    -Dmapreduce.job.reduces=1 \
    -files mapper.py,reducer.py \
    -mapper 'mapper.py {word}' \
    -reducer 'reducer.py {word}' \
    -input /hw2/hw2_2/src/enronemail_1h.txt \
    -output /hw2/hw2_3/tgt

    print "\nOUTPUT"
    # display accuracy on the console
    print "Accuracy of the Naive Bayes classifier with single word '{}'\n".format(word)
    !hdfs dfs -cat /hw2/hw2_3/tgt/part-00000

hw2_3("assistance")
```

```
15/09/15 01:31:02 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
classes where applicable
15/09/15 01:31:03 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interva
l = 0 minutes.
Deleted /hw2/hw2_3/tgt
15/09/15 01:31:04 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
classes where applicable
15/09/15 01:31:05 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
15/09/15 01:31:05 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
15/09/15 01:31:05 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already ini
tialized
15/09/15 01:31:06 INFO mapred.FileInputFormat: Total input paths to process : 1
15/09/15 01:31:06 INFO mapreduce.JobSubmitter: number of splits:1
15/09/15 01:31:06 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local699239845_0001
15/09/15 01:31:06 INFO mapred.LocalDistributedCacheManager: Localized file:/media/sf_shared/GitHub/MIDS-W261-MACHINE-LEARN
ING-AT-SCALE/week2/hw2/mapper.py as file:/app/hadoop/tmp/mapred/local/1442305866597/mapper.py
15/09/15 01:31:06 INFO mapred.LocalDistributedCacheManager: Localized file:/media/sf_shared/GitHub/MIDS-W261-MACHINE-LEARN
ING-AT-SCALE/week2/hw2/reducer.py as file:/app/hadoop/tmp/mapred/local/1442305866598/reducer.py
15/09/15 01:31:07 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
15/09/15 01:31:07 INFO mapred.LocalJobRunner: OutputCommitter set in config null
15/09/15 01:31:07 INFO mapreduce.Job: Running job: job_local699239845_0001
15/09/15 01:31:07 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
15/09/15 01:31:07 INFO mapred.LocalJobRunner: Waiting for map tasks
15/09/15 01:31:07 INFO mapred.LocalJobRunner: Starting task: attempt_local699239845_0001_m_000000_0
15/09/15 01:31:07 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
15/09/15 01:31:07 INFO mapred.MapTask: Processing split: hdfs://localhost:54310/hw2/hw2_2/src/enronemail_1h.txt:0+203979
15/09/15 01:31:07 INFO mapred.MapTask: numReduceTasks: 1
15/09/15 01:31:07 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
15/09/15 01:31:07 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
15/09/15 01:31:07 INFO mapred.MapTask: soft limit at 83886080
15/09/15 01:31:07 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
15/09/15 01:31:07 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
15/09/15 01:31:07 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
15/09/15 01:31:07 INFO streaming.PipeMapRed: PipeMapRed exec [/media/sf_shared/GitHub/MIDS-W261-MACHINE-LEARNING-AT-SCALE/
```

```
week2/hw2/./mapper.py, assistance]
15/09/15 01:31:07 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
15/09/15 01:31:07 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
15/09/15 01:31:07 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
15/09/15 01:31:07 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
15/09/15 01:31:07 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
15/09/15 01:31:07 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output
.dir
15/09/15 01:31:07 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
15/09/15 01:31:07 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
15/09/15 01:31:07 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
15/09/15 01:31:07 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
15/09/15 01:31:07 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
15/09/15 01:31:07 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partiti
on
15/09/15 01:31:07 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:31:07 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:31:07 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:31:07 INFO streaming.PipeMapRed: Records R/W=100/1
15/09/15 01:31:07 INFO streaming.PipeMapRed: MRErrorThread done
15/09/15 01:31:07 INFO streaming.PipeMapRed: mapRedFinished
15/09/15 01:31:07 INFO mapred.LocalJobRunner:
15/09/15 01:31:07 INFO mapred.MapTask: Starting flush of map output
15/09/15 01:31:07 INFO mapred.MapTask: Spilling map output
15/09/15 01:31:07 INFO mapred.MapTask: bufstart = 0; bufend = 3956; bufvoid = 104857600
15/09/15 01:31:07 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214000(104856000); length = 397/6553600
15/09/15 01:31:07 INFO mapred.MapTask: Finished spill 0
15/09/15 01:31:08 INFO mapreduce.Job: Job job_local699239845_0001 running in uber mode : false
15/09/15 01:31:08 INFO mapreduce.Job:  map 0% reduce 0%
15/09/15 01:31:08 INFO mapred.Task: Task:attempt_local699239845_0001_m_000000_0 is done. And is in the process of committi
ng
15/09/15 01:31:08 INFO mapred.LocalJobRunner: Records R/W=100/1
15/09/15 01:31:08 INFO mapred.Task: Task 'attempt_local699239845_0001_m_000000_0' done.
15/09/15 01:31:08 INFO mapred.LocalJobRunner: Finishing task: attempt_local699239845_0001_m_000000_0
15/09/15 01:31:08 INFO mapred.LocalJobRunner: map task executor complete.
15/09/15 01:31:08 INFO mapred.LocalJobRunner: Waiting for reduce tasks
15/09/15 01:31:08 INFO mapred.LocalJobRunner: Starting task: attempt_local699239845_0001_r_000000_0
15/09/15 01:31:08 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
15/09/15 01:31:08 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@7db
5b6ec
15/09/15 01:31:08 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=363285696, maxSingleShuffleLimit=90821424, merg
eThreshold=239768576, ioSortFactor=10, memToMemMergeOutputsThreshold=10
15/09/15 01:31:08 INFO reduce.EventFetcher: attempt_local699239845_0001_r_000000_0 Thread started: EventFetcher for fetchi
ng Map Completion Events
15/09/15 01:31:08 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local699239845_0001_m_00
0000_0 decomp: 4158 len: 4162 to MEMORY
15/09/15 01:31:08 INFO reduce.InMemoryMapOutput: Read 4158 bytes from map-output for attempt_local699239845_0001_m_000000_
0
15/09/15 01:31:08 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 4158, inMemoryMapOutputs.size() -
> 1, commitMemory -> 0, usedMemory ->4158
15/09/15 01:31:08 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
15/09/15 01:31:08 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 01:31:08 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
15/09/15 01:31:08 INFO mapred.Merger: Merging 1 sorted segments
15/09/15 01:31:08 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 4120 bytes
15/09/15 01:31:08 INFO reduce.MergeManagerImpl: Merged 1 segments, 4158 bytes to disk to satisfy reduce memory limit
15/09/15 01:31:08 INFO reduce.MergeManagerImpl: Merging 1 files, 4162 bytes from disk
15/09/15 01:31:08 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
15/09/15 01:31:08 INFO mapred.Merger: Merging 1 sorted segments
15/09/15 01:31:08 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 4120 bytes
15/09/15 01:31:08 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 01:31:08 INFO streaming.PipeMapRed: PipeMapRed exec [/media/sf_shared/GitHub/MIDS-W261-MACHINE-LEARNING-AT-SCALE/
week2/hw2/./reducer.py, assistance]
15/09/15 01:31:08 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
15/09/15 01:31:08 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
15/09/15 01:31:08 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:31:08 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:31:08 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:31:08 INFO streaming.PipeMapRed: Records R/W=100/1
15/09/15 01:31:08 INFO streaming.PipeMapRed: MRErrorThread done
15/09/15 01:31:08 INFO streaming.PipeMapRed: mapRedFinished
15/09/15 01:31:09 INFO mapred.Task: Task:attempt_local699239845_0001_r_000000_0 is done. And is in the process of committi
ng
15/09/15 01:31:09 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 01:31:09 INFO mapred.Task: Task attempt_local699239845_0001_r_000000_0 is allowed to commit now
15/09/15 01:31:09 INFO output.FileOutputCommitter: Saved output of task 'attempt_local699239845_0001_r_000000_0' to hdfs:/
/localhost:54310/hw2/hw2_3/tgt/_temporary/0/task_local699239845_0001_r_000000
15/09/15 01:31:09 INFO mapred.LocalJobRunner: Records R/W=100/1 > reduce
15/09/15 01:31:09 INFO mapred.Task: Task 'attempt_local699239845_0001_r_000000_0' done.
15/09/15 01:31:09 INFO mapred.LocalJobRunner: Finishing task: attempt_local699239845_0001_r_000000_0
15/09/15 01:31:09 INFO mapred.LocalJobRunner: reduce task executor complete.
15/09/15 01:31:09 INFO mapreduce.Job:  map 100% reduce 100%
15/09/15 01:31:10 INFO mapreduce.Job: Job job_local699239845_0001 completed successfully
15/09/15 01:31:10 INFO mapreduce.Job: Counters: 38
        File System Counters
```

```
                FILE: Number of bytes read=229170
                FILE: Number of bytes written=743032
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=407958
                HDFS: Number of bytes written=7788
                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=3956
                Map output materialized bytes=4162
                Input split bytes=106
                Combine input records=0
                Combine output records=0
                Reduce input groups=100
                Reduce shuffle bytes=4162
                Reduce input records=100
                Reduce output records=112
                Spilled Records=200
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=40
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=335683584
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=203979
        File Output Format Counters
                Bytes Written=7788
15/09/15 01:31:10 INFO streaming.StreamJob: Output directory: /hw2/hw2_3/tgt

OUTPUT
Accuracy of the Naive Bayes classifier with single word 'assistance'

15/09/15 01:31:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
classes where applicable
vocab size = 1.0
/*log probabilities*/
pr_spam_prior = -0.356547323514
pr_ham_prior = -0.251811972994


ID                                                  | TRUTH | CLASS
----------------------------------------------------+---------+-----------
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       | 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       | 0
0003.2004-08-01.BG                                  | 1       | 0
0004.1999-12-10.kaminski                            | 0       | 1
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       | 1
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	1
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	1
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	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	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	1
0018.2003-12-18.GP	1	1

```
/*accuracy*/
accuracy = 0.60
```

HW2.4

Using the Enron data from HW1 and in the Hadoop MapReduce framework, write a mapper/reducer pair that will classify the email messages using multinomial Naive Bayes Classifier using a list of one or more user-specified words. Examine the words "assistance", "valium", and "enlargementWithATypo" and report your results. To do so, make sure that

- mapper.py counts all occurrences of a list of words, and
- reducer.py
that performs a multiple-word multinomial Naive Bayes classification via the chosen list

Preparing to run the job

In [ ]:
```
!hdfs dfs -mkdir /hw2/hw2_4
```

Driver Function

In [21]:
```
# HW 2.4  Mapper/reducer pair to classify the email messages by a
#         list of multiple word using the multinomial Naive Bayes
#         classification

def hw2_4(word):
    # cleanup target directory
    !hdfs dfs -rm -R /hw2/hw2_4/tgt

    # run map reduce job
    !hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar \
    -Dmapreduce.job.maps=10 \
    -Dmapreduce.job.reduces=1 \
    -files mapper.py,reducer.py \
    -mapper 'mapper.py {word}' \
    -reducer 'reducer.py {word}' \
    -input /hw2/hw2_2/src/enronemail_1h.txt \
    -output /hw2/hw2_4/tgt

    print "\nOUTPUT"
    # display accuracy on the console
    print "Accuracy of the Naive Bayes classifier with single word '{}'\n".format(word)
    !hdfs dfs -cat /hw2/hw2_4/tgt/part-00000

hw2_4("assistance valium enlargementWithATypo")
```

```
15/09/15 01:31:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
classes where applicable
15/09/15 01:31:20 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interva
l = 0 minutes.
Deleted /hw2/hw2_4/tgt
15/09/15 01:31:21 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
classes where applicable
15/09/15 01:31:22 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
15/09/15 01:31:22 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
15/09/15 01:31:22 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already ini
tialized
15/09/15 01:31:23 INFO mapred.FileInputFormat: Total input paths to process : 1
15/09/15 01:31:23 INFO mapreduce.JobSubmitter: number of splits:1
15/09/15 01:31:23 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1149815606_0001
15/09/15 01:31:24 INFO mapred.LocalDistributedCacheManager: Localized file:/media/sf_shared/GitHub/MIDS-W261-MACHINE-LEARN
ING-AT-SCALE/week2/hw2/mapper.py as file:/app/hadoop/tmp/mapred/local/1442305883804/mapper.py
15/09/15 01:31:24 INFO mapred.LocalDistributedCacheManager: Localized file:/media/sf_shared/GitHub/MIDS-W261-MACHINE-LEARN
ING-AT-SCALE/week2/hw2/reducer.py as file:/app/hadoop/tmp/mapred/local/1442305883805/reducer.py
15/09/15 01:31:24 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
15/09/15 01:31:24 INFO mapreduce.Job: Running job: job_local1149815606_0001
15/09/15 01:31:24 INFO mapred.LocalJobRunner: OutputCommitter set in config null
15/09/15 01:31:24 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
15/09/15 01:31:24 INFO mapred.LocalJobRunner: Waiting for map tasks
15/09/15 01:31:24 INFO mapred.LocalJobRunner: Starting task: attempt_local1149815606_0001_m_000000_0
15/09/15 01:31:24 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
15/09/15 01:31:24 INFO mapred.MapTask: Processing split: hdfs://localhost:54310/hw2/hw2_2/src/enronemail_1h.txt:0+203979
15/09/15 01:31:24 INFO mapred.MapTask: numReduceTasks: 1
15/09/15 01:31:24 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
15/09/15 01:31:24 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
15/09/15 01:31:24 INFO mapred.MapTask: soft limit at 83886080
15/09/15 01:31:24 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
15/09/15 01:31:24 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
15/09/15 01:31:24 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
15/09/15 01:31:24 INFO streaming.PipeMapRed: PipeMapRed exec [/media/sf_shared/GitHub/MIDS-W261-MACHINE-LEARNING-AT-SCALE/
week2/hw2/./mapper.py, assistance, valium, enlargementWithATypo]
15/09/15 01:31:24 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
15/09/15 01:31:24 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
15/09/15 01:31:24 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
```

```
15/09/15 01:31:24 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
15/09/15 01:31:24 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
15/09/15 01:31:24 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output
.dir
15/09/15 01:31:24 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
15/09/15 01:31:24 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
15/09/15 01:31:24 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
15/09/15 01:31:24 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
15/09/15 01:31:24 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
15/09/15 01:31:24 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partiti
on
15/09/15 01:31:24 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:31:24 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:31:24 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:31:24 INFO streaming.PipeMapRed: Records R/W=100/1
15/09/15 01:31:24 INFO streaming.PipeMapRed: MRErrorThread done
15/09/15 01:31:24 INFO streaming.PipeMapRed: mapRedFinished
15/09/15 01:31:24 INFO mapred.LocalJobRunner:
15/09/15 01:31:24 INFO mapred.MapTask: Starting flush of map output
15/09/15 01:31:24 INFO mapred.MapTask: Spilling map output
15/09/15 01:31:24 INFO mapred.MapTask: bufstart = 0; bufend = 3978; bufvoid = 104857600
15/09/15 01:31:24 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214000(104856000); length = 397/6553600
15/09/15 01:31:24 INFO mapred.MapTask: Finished spill 0
15/09/15 01:31:24 INFO mapred.Task: Task:attempt_local1149815606_0001_m_000000_0 is done. And is in the process of committ
ing
15/09/15 01:31:24 INFO mapred.LocalJobRunner: Records R/W=100/1
15/09/15 01:31:24 INFO mapred.Task: Task 'attempt_local1149815606_0001_m_000000_0' done.
15/09/15 01:31:24 INFO mapred.LocalJobRunner: Finishing task: attempt_local1149815606_0001_m_000000_0
15/09/15 01:31:24 INFO mapred.LocalJobRunner: map task executor complete.
15/09/15 01:31:24 INFO mapred.LocalJobRunner: Waiting for reduce tasks
15/09/15 01:31:24 INFO mapred.LocalJobRunner: Starting task: attempt_local1149815606_0001_r_000000_0
15/09/15 01:31:25 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
15/09/15 01:31:25 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@411
a0a31
15/09/15 01:31:25 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=363285696, maxSingleShuffleLimit=90821424, merg
eThreshold=239768576, ioSortFactor=10, memToMemMergeOutputsThreshold=10
15/09/15 01:31:25 INFO reduce.EventFetcher: attempt_local1149815606_0001_r_000000_0 Thread started: EventFetcher for fetch
ing Map Completion Events
15/09/15 01:31:25 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1149815606_0001_m_0
00000_0 decomp: 4180 len: 4184 to MEMORY
15/09/15 01:31:25 INFO reduce.InMemoryMapOutput: Read 4180 bytes from map-output for attempt_local1149815606_0001_m_000000
_0
15/09/15 01:31:25 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 4180, inMemoryMapOutputs.size() -
> 1, commitMemory -> 0, usedMemory ->4180
15/09/15 01:31:25 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
15/09/15 01:31:25 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 01:31:25 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
15/09/15 01:31:25 INFO mapred.Merger: Merging 1 sorted segments
15/09/15 01:31:25 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 4142 bytes
15/09/15 01:31:25 INFO reduce.MergeManagerImpl: Merged 1 segments, 4180 bytes to disk to satisfy reduce memory limit
15/09/15 01:31:25 INFO reduce.MergeManagerImpl: Merging 1 files, 4184 bytes from disk
15/09/15 01:31:25 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
15/09/15 01:31:25 INFO mapred.Merger: Merging 1 sorted segments
15/09/15 01:31:25 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 4142 bytes
15/09/15 01:31:25 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 01:31:25 INFO streaming.PipeMapRed: PipeMapRed exec [/media/sf_shared/GitHub/MIDS-W261-MACHINE-LEARNING-AT-SCALE/
week2/hw2/./reducer.py, assistance, valium, enlargementWithATypo]
15/09/15 01:31:25 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
15/09/15 01:31:25 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
15/09/15 01:31:25 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:31:25 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:31:25 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:31:25 INFO streaming.PipeMapRed: Records R/W=100/1
15/09/15 01:31:25 INFO streaming.PipeMapRed: MRErrorThread done
15/09/15 01:31:25 INFO streaming.PipeMapRed: mapRedFinished
15/09/15 01:31:25 INFO mapreduce.Job: Job job_local1149815606_0001 running in uber mode : false
15/09/15 01:31:25 INFO mapreduce.Job:  map 100% reduce 0%
15/09/15 01:31:25 INFO mapred.Task: Task:attempt_local1149815606_0001_r_000000_0 is done. And is in the process of committ
ing
15/09/15 01:31:25 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 01:31:25 INFO mapred.Task: Task attempt_local1149815606_0001_r_000000_0 is allowed to commit now
15/09/15 01:31:25 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1149815606_0001_r_000000_0' to hdfs:
//localhost:54310/hw2/hw2_4/tgt/_temporary/0/task_local1149815606_0001_r_000000
15/09/15 01:31:25 INFO mapred.LocalJobRunner: Records R/W=100/1 > reduce
15/09/15 01:31:25 INFO mapred.Task: Task 'attempt_local1149815606_0001_r_000000_0' done.
15/09/15 01:31:25 INFO mapred.LocalJobRunner: Finishing task: attempt_local1149815606_0001_r_000000_0
15/09/15 01:31:25 INFO mapred.LocalJobRunner: reduce task executor complete.
15/09/15 01:31:26 INFO mapreduce.Job:  map 100% reduce 100%
15/09/15 01:31:26 INFO mapreduce.Job: Job job_local1149815606_0001 completed successfully
15/09/15 01:31:26 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=229214
                FILE: Number of bytes written=746102
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
```

```
                FILE: Number of write operations=0
                HDFS: Number of bytes read=407958
                HDFS: Number of bytes written=7788
                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=3978
                Map output materialized bytes=4184
                Input split bytes=106
                Combine input records=0
                Combine output records=0
                Reduce input groups=100
                Reduce shuffle bytes=4184
                Reduce input records=100
                Reduce output records=112
                Spilled Records=200
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=38
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=335683584
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=203979
        File Output Format Counters
                Bytes Written=7788
15/09/15 01:31:26 INFO streaming.StreamJob: Output directory: /hw2/hw2_4/tgt

OUTPUT
Accuracy of the Naive Bayes classifier with single word 'assistance valium enlargementWithATypo'

15/09/15 01:31:27 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
classes where applicable
vocab size = 3.0
/*log probabilities*/
pr_spam_prior = -0.356547323514
pr_ham_prior = -0.251811972994


ID                                                  | TRUTH | CLASS
----------------------------------------------------+---------+-----------
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       | 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       | 0
0003.2004-08-01.BG                                  | 1       | 0
0004.1999-12-10.kaminski                            | 0       | 1
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       | 1
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       | 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       | 0
0011.1999-12-14.farmer                      | 0       | 0
0011.2001-06-28.SA_and_HP                   | 1       | 1
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       | 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       | 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       | 1
0017.2004-08-02.BG                          | 1       | 0
0018.1999-12-14.kaminski                    | 0       | 0
0018.2001-07-13.SA_and_HP                   | 1       | 1
0018.2003-12-18.GP                          | 1       | 1


/*accuracy*/
accuracy = 0.62
```

HW2.5

Using the Enron data from HW1 an in the Hadoop MapReduce framework, write a mapper/reducer for a multinomial Naive Bayes Classifier that will classify the email messages using words present. Also drop words with a frequency of less than three (3). How does it affect the misclassification error of learnt naive multinomial Bayesian Classifiers on the training dataset.

Reducer

The reducer in this problem handles all the words in the emails. Additionally, classifier drops words with a frequency of less than three (3). The output of the reducer is each email id with actual spam/ham indicator with prediction followed by accuracy.

When the classifier drops words with frequency less than 3, I see there is NO change in accuracy though vocabularizy size reduces by ~60%.

In [22]:
```
%%writefile reducer.py
#!/usr/bin/python
import traceback
import math
import sys
import ast

from collections import Counter

# read input parameters
find_words = sys.argv[1:]

# vocab size
if find_words == "*":
    vocab = []
else:
    vocab = find_words

try:
    spam_count = 0
    ham_count = 0
    spam_all_wc = 0
    ham_all_wc = 0
    spam_term_wc = {}
    ham_term_wc = {}
    pr_word_given_spam = {}
    pr_word_given_ham = {}

    # read each mapper output to loop during the prediction phase
    # after training the model
    map_output = []
    for line in sys.stdin:
        map_output.append(line)

    for email in map_output:
        # parse mapper output
        mail = email.split(" | ")
        # read spam/ham indicator, content word count,
        is_spam = int(mail[1])
        content_wc = int(mail[2])
        hits = ast.literal_eval(mail[3])

        # capture counts required for naive bayes probabilities
        if is_spam:
            # spam mail count
            spam_count += 1
            # term count when spam
            spam_term_wc = dict(Counter(hits) + Counter(spam_term_wc))
            # all word count when spam
            spam_all_wc += content_wc
        else:
            # ham email count
            ham_count += 1
            # term count when ham
            ham_term_wc = dict(Counter(hits) + Counter(ham_term_wc))
            # all word count when ham
            ham_all_wc += content_wc

    vocab = dict(Counter(vocab) + Counter(spam_term_wc) + Counter(ham_term_wc))
    vocab = {k:v for (k,v) in vocab.items() if v >= 3}
    V = len(vocab) * 1.0
    print "vocab size = {}".format(V)

    # calculate priors
    pr_spam_prior = (1.0 * spam_count) / (spam_count + ham_count)
    pr_ham_prior = (1.0 - pr_spam_prior)
    pr_spam_prior = math.log10(pr_spam_prior)
    pr_ham_prior = math.log10(pr_ham_prior)

    # calculate conditional probabilites with laplace smoothing = 1
    # pr_word_given_class = ( count(w, c) + 1 ) / (count(c) + 1 * |V|)
    for word in vocab:
        #if (vocab[word] >= 3):
        pr_word_given_spam[word] = math.log10((spam_term_wc.get(word, 0) + 1.0) / (spam_all_wc + V))
        pr_word_given_ham[word] = math.log10((ham_term_wc.get(word, 0) + 1.0) / (ham_all_wc + V))

    print "/*log probabilities*/"
    print "pr_spam_prior = {}".format(pr_spam_prior)
```

```
    print "pr_ham_prior = {}".format(pr_ham_prior)

    print "\n"
    print "{0: <50} | {1} | {2}".format("email id", "actuals", "predictions")
    print "{0: <50}-+-{1}-+-{2}".format("-" * 50, "-" * 7, "-" * 10)

    # spam/ham prediction using Multinomial Naive Bayes priors and conditional probabilities
    accuracy = []

    for email in map_output:
        # initialize
        word_count = 0
        pred_is_spam = 0
        pr_spam = pr_spam_prior
        pr_ham = pr_ham_prior

        # parse mapper output
        mail = email.split(" | ")
        email_id = mail[0]
        is_spam = int(mail[1])
        hits = ast.literal_eval(mail[3])

        # number of search words
        word_count = sum(hits.values())

        # probability for each class for a given email
        # argmax [ log P(C) + sum( P(Wi|C) ) ]
        for word in vocab:
            pr_spam += (pr_word_given_spam.get(word, 0) * hits.get(word, 0))
            pr_ham += (pr_word_given_ham.get(word, 0) * hits.get(word, 0))

        # predict based on maximum likelihood
        if pr_spam > pr_ham:
            pred_is_spam = 1

        # calculate accuracy
        accuracy.append(pred_is_spam==is_spam)

        print '{0:<50} | {1:<7} | {2:<10}'.format(email_id, is_spam, pred_is_spam)

    print "\n"
    print "/*accuracy*/"
    print "accuracy = {:.2f}".format(sum(accuracy) / float(len(accuracy)))

except Exception:
    traceback.print_exc()
```

```
Overwriting reducer.py
```

Preparing to run the job

In [ ]:
```
!hdfs dfs -mkdir /hw2/hw2_5
```

Driver Function

In [24]:
```
# HW 2.5  Mapper/reducer pair to classify the email messages by a
#         all words present to perform a word-distribution-wide Naive
#         Bayes classification

def hw2_5(word):
    # cleanup target directory
    !hdfs dfs -rm -R /hw2/hw2_5/tgt

    # run map reduce job
    !hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar \
    -Dmapreduce.job.maps=10 \
    -Dmapreduce.job.reduces=1 \
    -files mapper.py,reducer.py \
    -mapper 'mapper.py {word}' \
    -reducer 'reducer.py {word}' \
    -input /hw2/hw2_2/src/enronemail_1h.txt \
    -output /hw2/hw2_5/tgt

    print "\nOUTPUT"
    # display accuracy on the console
    print "Accuracy of the Naive Bayes classifier with single word '{}'\n".format(word)
    !hdfs dfs -cat /hw2/hw2_5/tgt/part-00000

hw2_5("*")
```

```
15/09/15 01:31:33 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
classes where applicable
```

```
15/09/15 01:31:34 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interva
l = 0 minutes.
Deleted /hw2/hw2_5/tgt
15/09/15 01:31:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
classes where applicable
15/09/15 01:31:36 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
15/09/15 01:31:36 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
15/09/15 01:31:36 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already ini
tialized
15/09/15 01:31:37 INFO mapred.FileInputFormat: Total input paths to process : 1
15/09/15 01:31:37 INFO mapreduce.JobSubmitter: number of splits:1
15/09/15 01:31:37 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1701776022_0001
15/09/15 01:31:37 INFO mapred.LocalDistributedCacheManager: Localized file:/media/sf_shared/GitHub/MIDS-W261-MACHINE-LEARN
ING-AT-SCALE/week2/hw2/mapper.py as file:/app/hadoop/tmp/mapred/local/1442305897611/mapper.py
15/09/15 01:31:37 INFO mapred.LocalDistributedCacheManager: Localized file:/media/sf_shared/GitHub/MIDS-W261-MACHINE-LEARN
ING-AT-SCALE/week2/hw2/reducer.py as file:/app/hadoop/tmp/mapred/local/1442305897612/reducer.py
15/09/15 01:31:38 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
15/09/15 01:31:38 INFO mapreduce.Job: Running job: job_local1701776022_0001
15/09/15 01:31:38 INFO mapred.LocalJobRunner: OutputCommitter set in config null
15/09/15 01:31:38 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
15/09/15 01:31:38 INFO mapred.LocalJobRunner: Waiting for map tasks
15/09/15 01:31:38 INFO mapred.LocalJobRunner: Starting task: attempt_local1701776022_0001_m_000000_0
15/09/15 01:31:38 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
15/09/15 01:31:38 INFO mapred.MapTask: Processing split: hdfs://localhost:54310/hw2/hw2_2/src/enronemail_1h.txt:0+203979
15/09/15 01:31:38 INFO mapred.MapTask: numReduceTasks: 1
15/09/15 01:31:38 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
15/09/15 01:31:38 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
15/09/15 01:31:38 INFO mapred.MapTask: soft limit at 83886080
15/09/15 01:31:38 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
15/09/15 01:31:38 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
15/09/15 01:31:38 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
15/09/15 01:31:38 INFO streaming.PipeMapRed: PipeMapRed exec [/media/sf_shared/GitHub/MIDS-W261-MACHINE-LEARNING-AT-SCALE/
week2/hw2/./mapper.py, *]
15/09/15 01:31:38 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
15/09/15 01:31:38 INFO Configuration.deprecation: mapred.local.dir is deprecated. Instead, use mapreduce.cluster.local.dir
15/09/15 01:31:38 INFO Configuration.deprecation: map.input.file is deprecated. Instead, use mapreduce.map.input.file
15/09/15 01:31:38 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
15/09/15 01:31:38 INFO Configuration.deprecation: map.input.length is deprecated. Instead, use mapreduce.map.input.length
15/09/15 01:31:38 INFO Configuration.deprecation: mapred.work.output.dir is deprecated. Instead, use mapreduce.task.output
.dir
15/09/15 01:31:38 INFO Configuration.deprecation: map.input.start is deprecated. Instead, use mapreduce.map.input.start
15/09/15 01:31:38 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
15/09/15 01:31:38 INFO Configuration.deprecation: user.name is deprecated. Instead, use mapreduce.job.user.name
15/09/15 01:31:38 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
15/09/15 01:31:38 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
15/09/15 01:31:38 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partiti
on
15/09/15 01:31:38 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:31:38 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:31:38 INFO streaming.PipeMapRed: Records R/W=72/1
15/09/15 01:31:38 INFO streaming.PipeMapRed: R/W/S=100/40/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:31:38 INFO streaming.PipeMapRed: MRErrorThread done
15/09/15 01:31:38 INFO streaming.PipeMapRed: mapRedFinished
15/09/15 01:31:38 INFO mapred.LocalJobRunner:
15/09/15 01:31:38 INFO mapred.MapTask: Starting flush of map output
15/09/15 01:31:38 INFO mapred.MapTask: Spilling map output
15/09/15 01:31:38 INFO mapred.MapTask: bufstart = 0; bufend = 196335; bufvoid = 104857600
15/09/15 01:31:38 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214000(104856000); length = 397/6553600
15/09/15 01:31:38 INFO mapred.MapTask: Finished spill 0
15/09/15 01:31:38 INFO mapred.Task: Task:attempt_local1701776022_0001_m_000000_0 is done. And is in the process of committ
ing
15/09/15 01:31:38 INFO mapred.LocalJobRunner: Records R/W=72/1
15/09/15 01:31:38 INFO mapred.Task: Task 'attempt_local1701776022_0001_m_000000_0' done.
15/09/15 01:31:38 INFO mapred.LocalJobRunner: Finishing task: attempt_local1701776022_0001_m_000000_0
15/09/15 01:31:38 INFO mapred.LocalJobRunner: map task executor complete.
15/09/15 01:31:38 INFO mapred.LocalJobRunner: Waiting for reduce tasks
15/09/15 01:31:38 INFO mapred.LocalJobRunner: Starting task: attempt_local1701776022_0001_r_000000_0
15/09/15 01:31:38 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
15/09/15 01:31:38 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@59c
ea8fc
15/09/15 01:31:38 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=363285696, maxSingleShuffleLimit=90821424, merg
eThreshold=239768576, ioSortFactor=10, memToMemMergeOutputsThreshold=10
15/09/15 01:31:38 INFO reduce.EventFetcher: attempt_local1701776022_0001_r_000000_0 Thread started: EventFetcher for fetch
ing Map Completion Events
15/09/15 01:31:38 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1701776022_0001_m_0
00000_0 decomp: 196725 len: 196729 to MEMORY
15/09/15 01:31:38 INFO reduce.InMemoryMapOutput: Read 196725 bytes from map-output for attempt_local1701776022_0001_m_0000
00_0
15/09/15 01:31:38 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 196725, inMemoryMapOutputs.size()
-> 1, commitMemory -> 0, usedMemory ->196725
15/09/15 01:31:38 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
15/09/15 01:31:38 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 01:31:38 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
15/09/15 01:31:38 INFO mapred.Merger: Merging 1 sorted segments
15/09/15 01:31:38 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 196680 bytes
15/09/15 01:31:38 INFO reduce.MergeManagerImpl: Merged 1 segments, 196725 bytes to disk to satisfy reduce memory limit
```

```
15/09/15 01:31:38 INFO reduce.MergeManagerImpl: Merging 1 files, 196729 bytes from disk
15/09/15 01:31:38 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
15/09/15 01:31:38 INFO mapred.Merger: Merging 1 sorted segments
15/09/15 01:31:38 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 196680 bytes
15/09/15 01:31:38 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 01:31:38 INFO streaming.PipeMapRed: PipeMapRed exec [/media/sf_shared/GitHub/MIDS-W261-MACHINE-LEARNING-AT-SCALE/
week2/hw2/./reducer.py, *]
15/09/15 01:31:39 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
15/09/15 01:31:39 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
15/09/15 01:31:39 INFO mapreduce.Job: Job job_local1701776022_0001 running in uber mode : false
15/09/15 01:31:39 INFO mapreduce.Job:  map 100% reduce 0%
15/09/15 01:31:39 INFO streaming.PipeMapRed: R/W/S=1/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:31:39 INFO streaming.PipeMapRed: R/W/S=10/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:31:39 INFO streaming.PipeMapRed: R/W/S=100/0/0 in:NA [rec/s] out:NA [rec/s]
15/09/15 01:31:39 INFO streaming.PipeMapRed: Records R/W=100/1
15/09/15 01:31:39 INFO streaming.PipeMapRed: MRErrorThread done
15/09/15 01:31:39 INFO streaming.PipeMapRed: mapRedFinished
15/09/15 01:31:39 INFO mapred.Task: Task:attempt_local1701776022_0001_r_000000_0 is done. And is in the process of committ
ing
15/09/15 01:31:39 INFO mapred.LocalJobRunner: 1 / 1 copied.
15/09/15 01:31:39 INFO mapred.Task: Task attempt_local1701776022_0001_r_000000_0 is allowed to commit now
15/09/15 01:31:39 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1701776022_0001_r_000000_0' to hdfs:
//localhost:54310/hw2/hw2_5/tgt/_temporary/0/task_local1701776022_0001_r_000000
15/09/15 01:31:39 INFO mapred.LocalJobRunner: Records R/W=100/1 > reduce
15/09/15 01:31:39 INFO mapred.Task: Task 'attempt_local1701776022_0001_r_000000_0' done.
15/09/15 01:31:39 INFO mapred.LocalJobRunner: Finishing task: attempt_local1701776022_0001_r_000000_0
15/09/15 01:31:39 INFO mapred.LocalJobRunner: reduce task executor complete.
15/09/15 01:31:40 INFO mapreduce.Job:  map 100% reduce 100%
15/09/15 01:31:40 INFO mapreduce.Job: Job job_local1701776022_0001 completed successfully
15/09/15 01:31:40 INFO mapreduce.Job: Counters: 38
        File System Counters
                FILE: Number of bytes read=614610
                FILE: Number of bytes written=1323747
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=407958
                HDFS: Number of bytes written=7799
                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=196335
                Map output materialized bytes=196729
                Input split bytes=106
                Combine input records=0
                Combine output records=0
                Reduce input groups=100
                Reduce shuffle bytes=196729
                Reduce input records=100
                Reduce output records=112
                Spilled Records=200
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=47
                CPU time spent (ms)=0
                Physical memory (bytes) snapshot=0
                Virtual memory (bytes) snapshot=0
                Total committed heap usage (bytes)=335683584
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=203979
        File Output Format Counters
                Bytes Written=7799
15/09/15 01:31:40 INFO streaming.StreamJob: Output directory: /hw2/hw2_5/tgt

OUTPUT
Accuracy of the Naive Bayes classifier with single word '*'

15/09/15 01:31:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
classes where applicable
vocab size = 1802.0
/*log probabilities*/
pr_spam_prior = -0.356547323514
pr_ham_prior = -0.251811972994
```

email id	actuals	predictions
0001.1999-12-10.farmer	0	0
0001.1999-12-10.kaminski	0	1
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       | 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


/*accuracy*/
accuracy = 0.98
```

** -- END OF ASSIGNMENT 2 -- **