SplitLogsIntoBuckets

The 'split-logs-into-buckets' tool

Use masses/tenpass/split-logs-into-buckets to split up mass-check output log files. Often, you may need to select a subset of lines from a 200000-line log file; for example, if you want to test using a sample of 2000 lines. This is achieved by splitting the log into "buckets".

For example, here's a sample run extracting ~2100-line log files from a single 210442-line file:

```
wc -l /home/corpus-rsync/corpus/Obsolete/submit-2.60-GA-run1/ham-set0.log
 210442 /home/corpus-rsync/corpus/Obsolete/submit-2.60-GA-run1/ham-set0.log

./tenpass/split-log-into-buckets 10 \
    < /home/corpus-rsync/corpus/Obsolete/submit-2.60-GA-run1/ham-set0.log
mv split-1.log new
./tenpass/split-log-into-buckets 10 < new

mv split-*.log ../../logs/nonspam-jm/

wc -l ../../logs/nonspam-jm/*.log
   2104 ../../logs/nonspam-jm/split-1.log
   2103 ../../logs/nonspam-jm/split-10.log
   2106 ../../logs/nonspam-jm/split-2.log
   2103 ../../logs/nonspam-jm/split-3.log
   2102 ../../logs/nonspam-jm/split-4.log
   2105 ../../logs/nonspam-jm/split-5.log
   2102 ../../logs/nonspam-jm/split-6.log
   2103 ../../logs/nonspam-jm/split-7.log
   2103 ../../logs/nonspam-jm/split-8.log
   2104 ../../logs/nonspam-jm/split-9.log
```

One key point about split-logs-into-buckets – it selects lines in a round-robin fashion. So the first line goes into split-1.log, second line into split-2.log ... tenth line into split-10.log, eleventh into split-1.log, twelfth into split-2.log, and so on until the input runs out.

The command line argument is the number of buckets to create.