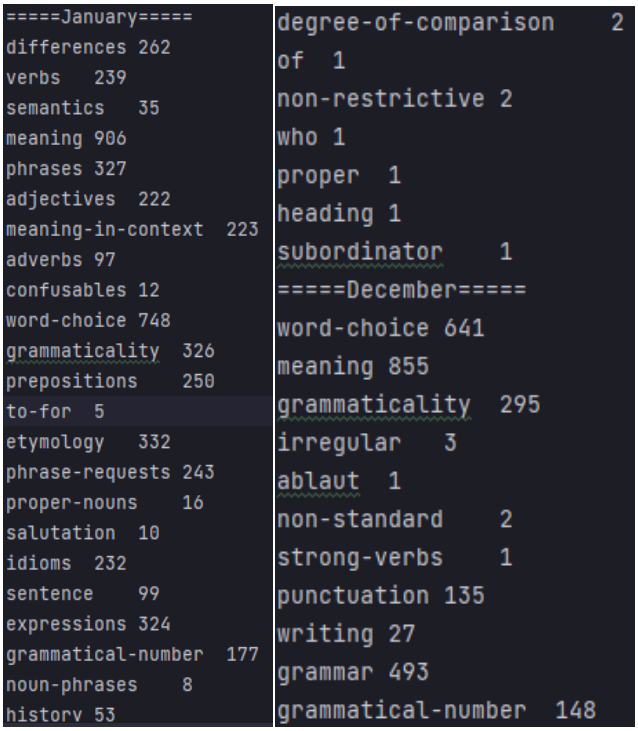
Scott Berry

Running MapReduce(way too much logging to include all):

```
› hadoop jar hadoop-streaming-3.3.1.jar -input Assignment_3/posts_stripped.csv -output Assignment_3/output -mapper Assignment_3/mapper.py -reducer Assignment_3/reducer.py
2022-02-27 10:45:33,356 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2022-02-27 10:45:33,566 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2022-02-27 10:45:33,631 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2022-02-27 10:45:33,631 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2022-02-27 10:45:33,639 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2022-02-27 10:45:33,820 INFO mapred.FileInputFormat: Total input files to process : 1
2022-02-27 10:45:33,916 INFO mapreduce.JobSubmitter: number of splits:7
2022-02-27 10:45:34,060 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1904969386_0001
2022-02-27 10:45:34,060 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-02-27 10:45:34,205 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2022-02-27 10:45:34,206 INFO mapreduce.Job: Running job: job_local1904969386_0001
2022-02-27 10:45:34,206 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2022-02-27 10:45:34,207 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2022-02-27 10:45:34,210 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2022-02-27 10:45:34,210 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2022-02-27 10:45:34,301 INFO mapred.LocalJobRunner: Waiting for map tasks
2022-02-27 10:45:34,302 INFO mapred.LocalJobRunner: Starting task: attempt_local1904969386_0001_m_000000_0
2022-02-27 10:45:34,319 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2022-02-27 10:45:34,319 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2022-02-27 10:45:34,326 INFO util.ProcfsBasedProcessTree: ProcfsBasedProcessTree currently is supported only on Linux.
2022-02-27 10:45:34,327 INFO mapred.Task:  Using ResourceCalculatorProcessTree : null
2022-02-27 10:45:34,333 INFO mapred.MapTask: Processing split: file:/Users/SRB3C8X/OneDrive - The Home Depot/DSOnRampHadoop/Assignment_3/posts_stripped.csv:0+33554432
```

```
2022-02-27 10:45:37,854 INFO mapred.LocalJobRunner: Finishing task: attempt_local1904969386_0001_r_000000_0
2022-02-27 10:45:37,854 INFO mapred.LocalJobRunner: reduce task executor complete.
2022-02-27 10:45:38,224 INFO mapreduce.Job:  map 100% reduce 100%
2022-02-27 10:45:38,224 INFO mapreduce.Job: Job job_local1904969386_0001 completed successfully
2022-02-27 10:45:38,231 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=1143193423
                FILE: Number of bytes written=63950140
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=252946
                Map output records=252945
                Map output bytes=8411622
                Map output materialized bytes=8917554
                Input split bytes=1008
                Combine input records=0
                Combine output records=0
                Reduce input groups=252433
                Reduce shuffle bytes=8917554
                Reduce input records=252945
                Reduce output records=8112
                Spilled Records=505890
                Shuffled Maps =7
                Failed Shuffles=0
                Merged Map outputs=7
                GC time elapsed (ms)=21
                Total committed heap usage (bytes)=2181038080
        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=209733213
        File Output Format Counters
                Bytes Written=116415
2022-02-27 10:45:38,231 INFO streaming.StreamJob: Output directory: Assignment_3/output
```

With the following output:



This Hadoop job is tasked with taking in StackOverflow posts and their metadata and compiling the counts of every tag's appearance over each month. This job successfully created an output of 8100 unique month:tag combinations and recorded the number of appearances of each tag in each month. The mapper.py logic takes in the csv of StackOverflow posts and tracks every post's creation date and tags to be sent to the reducer. The reducer then reads in every line and gets the month from the creation date using decision tree logic from the scraped month int as well as a list of tags from the tag field using text cleaning logic. Iterating over each tag, the tag count is added or incremented to a dictionary tracking all tag data for the corresponding month. The output is then iterated over for each month with every tag and count printed.