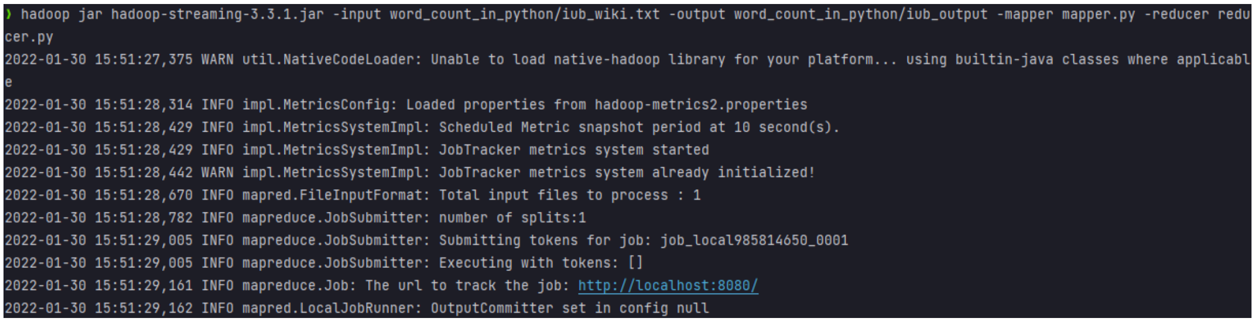
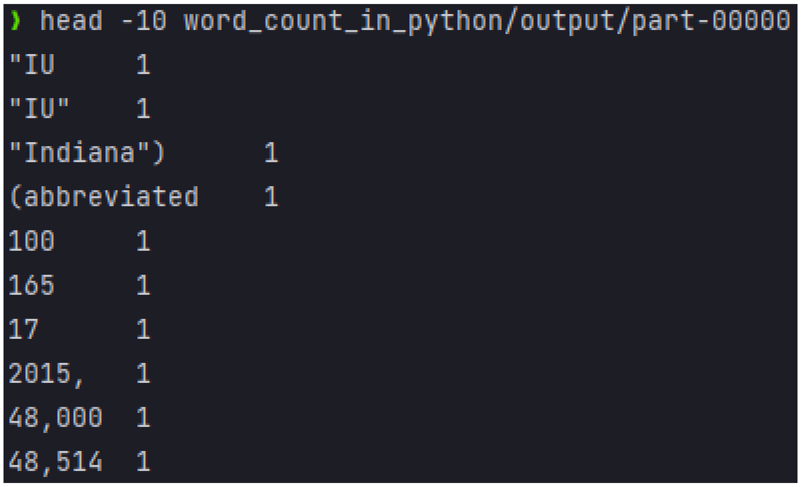
Scott Berry

Running Hadoop Streaming Command (full logging was way too much too screenshot)



```
❯ hadoop jar hadoop-streaming-3.3.1.jar -input word_count_in_python/iub_wiki.txt -output word_count_in_python/iub_output -mapper mapper.py -reducer redu
cer.py
2022-01-30 15:51:27,375 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicabl
e
2022-01-30 15:51:28,314 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2022-01-30 15:51:28,429 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2022-01-30 15:51:28,429 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2022-01-30 15:51:28,442 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2022-01-30 15:51:28,670 INFO mapred.FileInputFormat: Total input files to process : 1
2022-01-30 15:51:28,782 INFO mapreduce.JobSubmitter: number of splits:1
2022-01-30 15:51:29,005 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local985814650_0001
2022-01-30 15:51:29,005 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-01-30 15:51:29,161 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2022-01-30 15:51:29,162 INFO mapred.LocalJobRunner: OutputCommitter set in config null
```

With the output logging to the following file (Seems clear that cleaned data would yield more insightful results)



```
❯ head -10 word_count_in_python/output/part-00000
"IU     1
"IU"    1
"Indiana")      1
(abbreviated    1
100     1
165     1
17      1
2015,   1
48,000  1
48,514  1
```

The Hadoop Streaming command was enabled by creating a HDFS directory which I placed the IUB text file in. Along with providing the mapper and reducer python scripts as arguments, the output directory was created in the HDFS directory. With mapper, the IUB text was mapped into keys and values of individual words and counts of their occurrences. With reducer, the keys are aggregated and made unique with their corresponding count value. With this command and scripts, I was able to quickly create a count of every unique word in the IUB Wiki text file.