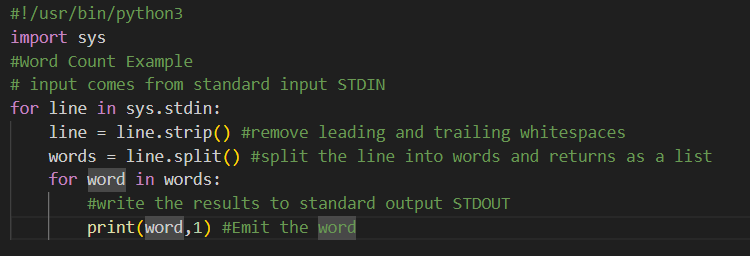
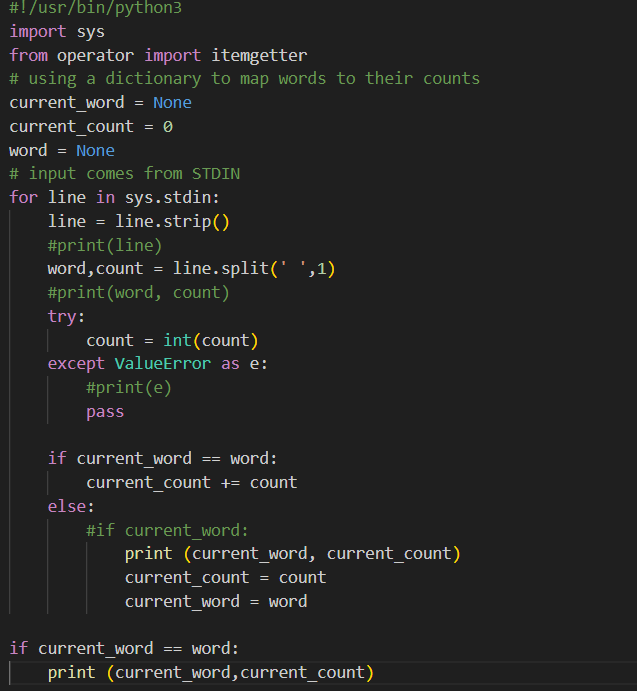
**Hadoop Streaming:**

1. download the mapper.py and reducer.py files



* This script reads lines of text, splits them into words, and outputs each word with a count of 1.



* This script aggregates the word counts, summing up the counts for each word.

2. test mapper:

cat streaming\_source\_data.txt | ./streaming\_mapper.py

* Output must be each word in the source file, with a 1 alongside it.
* words repeated in source file must be seen repeatedly in output, each having a 1 alongside

3. test reducer:

cat streaming\_source\_data.txt | ./streaming\_mapper.py | sort -k1,1 | ./streaming\_reducer.py

* Each word should be followed by its total count. Repeated words should be aggregated into a single entry with the total count.

4. Upload Test File to HDFS:

* Command: hdfs dfs -put streaming\_source\_data.txt /path/in/hdfs/
* Replace /path/in/hdfs/ with your target directory in HDFS.

5. Locate Hadoop Streaming JAR:

* Find the Hadoop streaming JAR file, typically named something like hadoop-streaming-\*.jar. Ensure it's accessible from your command line.

6. Run mapreduce job:

hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.10.2.jar \

-input /streaming\_source\_data.txt \

-output /stream\_out159 \

-mapper streaming\_mapper.py \

-reducer streaming\_reducer.py \

-file streaming\_mapper.py \

-file streaming\_reducer.py

* Adjust the paths as needed based on your Hadoop setup and file locations.
* Testing Mapper and Reducer: Ensure your mapper and reducer scripts are functioning as expected before deploying them on Hadoop.
* Upload to HDFS: Your data must be in HDFS for Hadoop to process it.
* Running MapReduce Job: This command uses Hadoop Streaming to run your mapper and reducer scripts on the input data stored in HDFS and outputs the result to the specified directory.