To help debug programs, here are some tips. The first 2 are an expansion of what was mentioned in the Join Assignment, the 3rd shows how get information during execution. The others are also helpful notes that students have suggested.

1. Try the following from the command line:

>cat testfile1 | ./<mapper.py> | sort

This will show you if the mapper script is working ok outside map/reduce, and not using what is in the HDFS file system.

2. Also try:

> cat testfile* | ./<mapper.py> | sort | ./<reducer.py>

If these commands work you should see the all the <word, total count>'s , and then you can try running in hadoop.

If these commands work but do not show the correct output, then check your code, test your logic, add in print statements to see what is happening, try making a new test file with just a few words.

If these commands work but then your map/reduce job fails, then it becomes more tricky to debug. Perhaps something about data being partitioned is throwing off your logic, I added a reading material, which shows how you can debug a map/reduce program. I suspect you won't need to go there for wordcount, but maybe for join assignment.

3. Here's is some code techniques for debuging Python streaming.

Programs using map/reduce is notoriously difficult to debug,

especially because Hadoop is managing the execution and standard error/output files