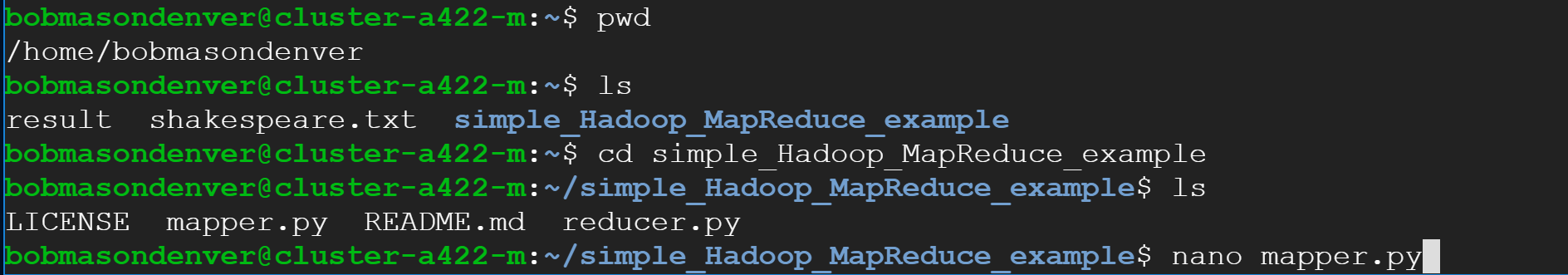
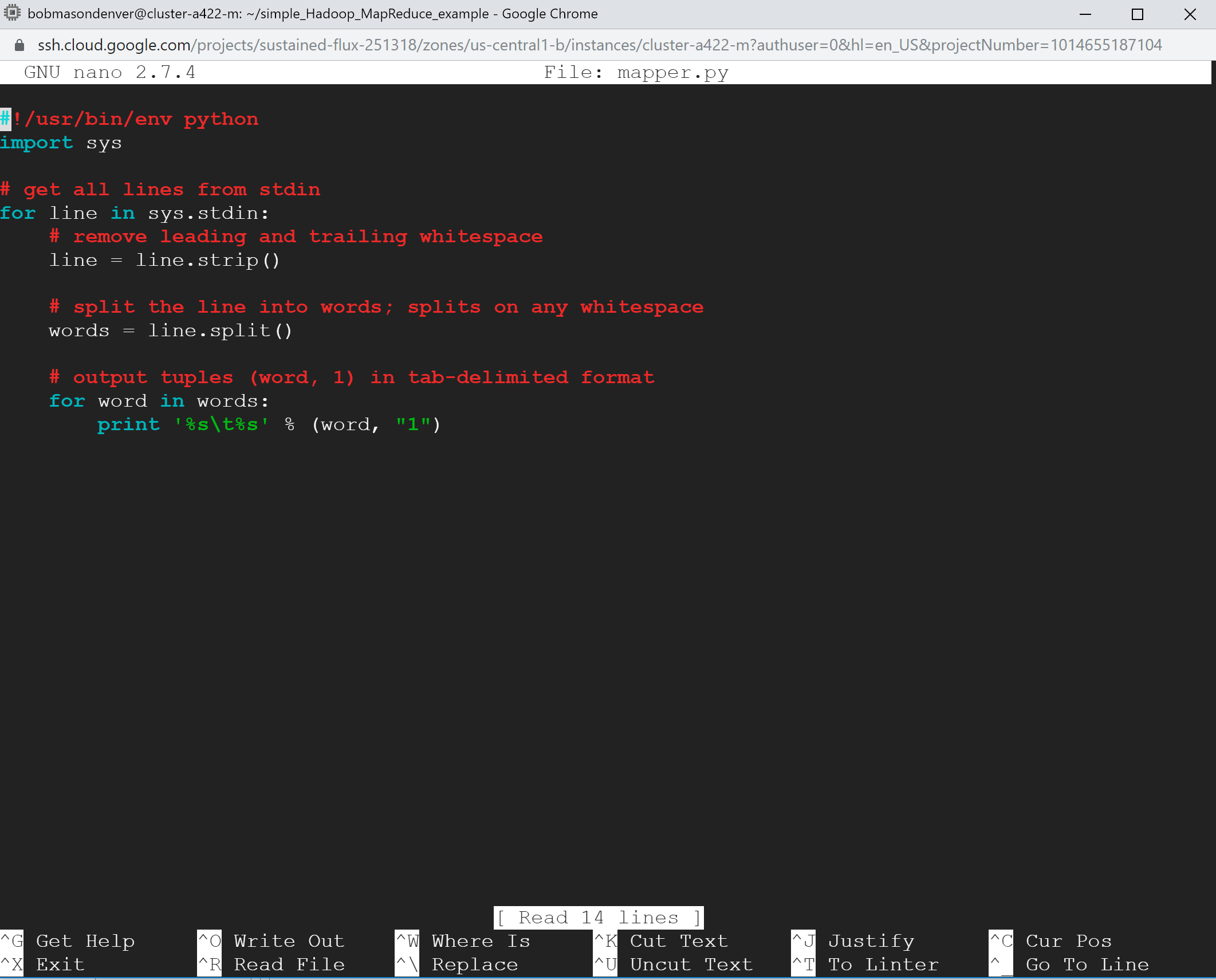
MSDS610 Data Engineering

Week 2 Lab Supplemental – Part 2

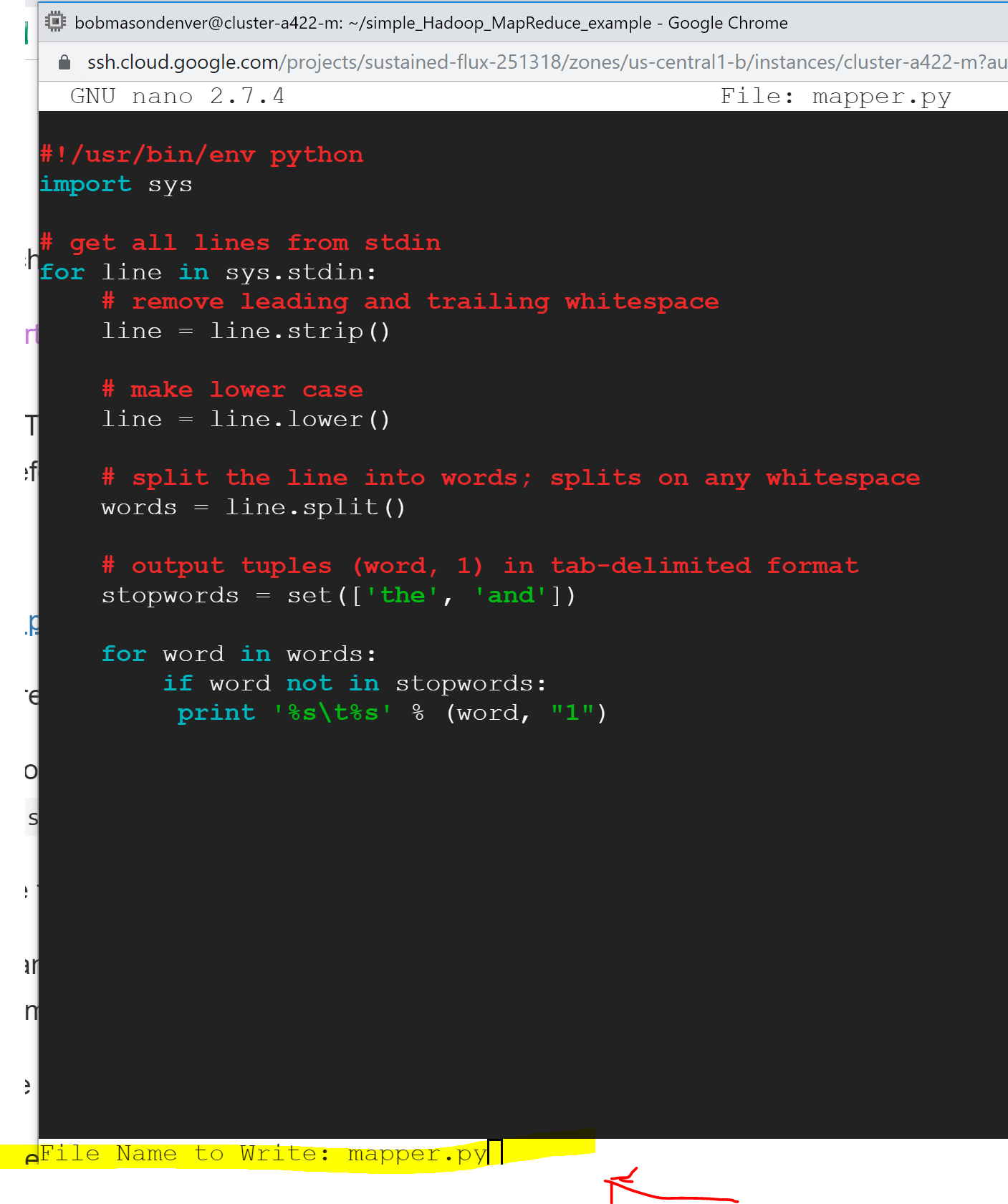
as of 9/6/2019

**Next, we want to remove stopwords. Stopwords are common words (e.g. the, I, and) that usually don't tell us much about the context of text.  Change (cd) into the directory for the simple haddoop MapReduce example and then use the nano editor to change the python script as shown below.**

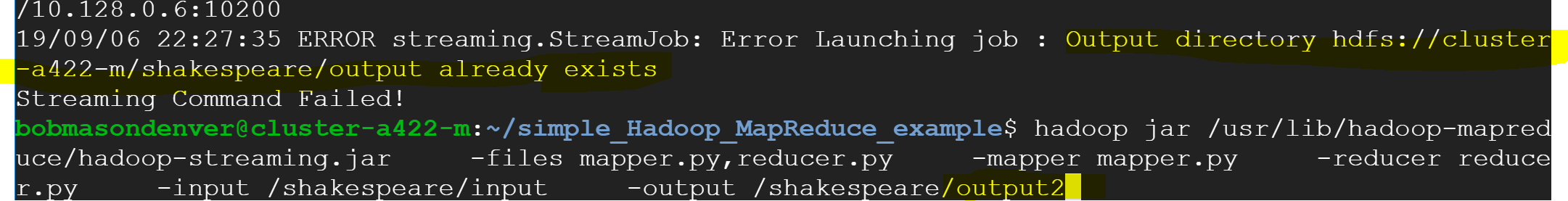


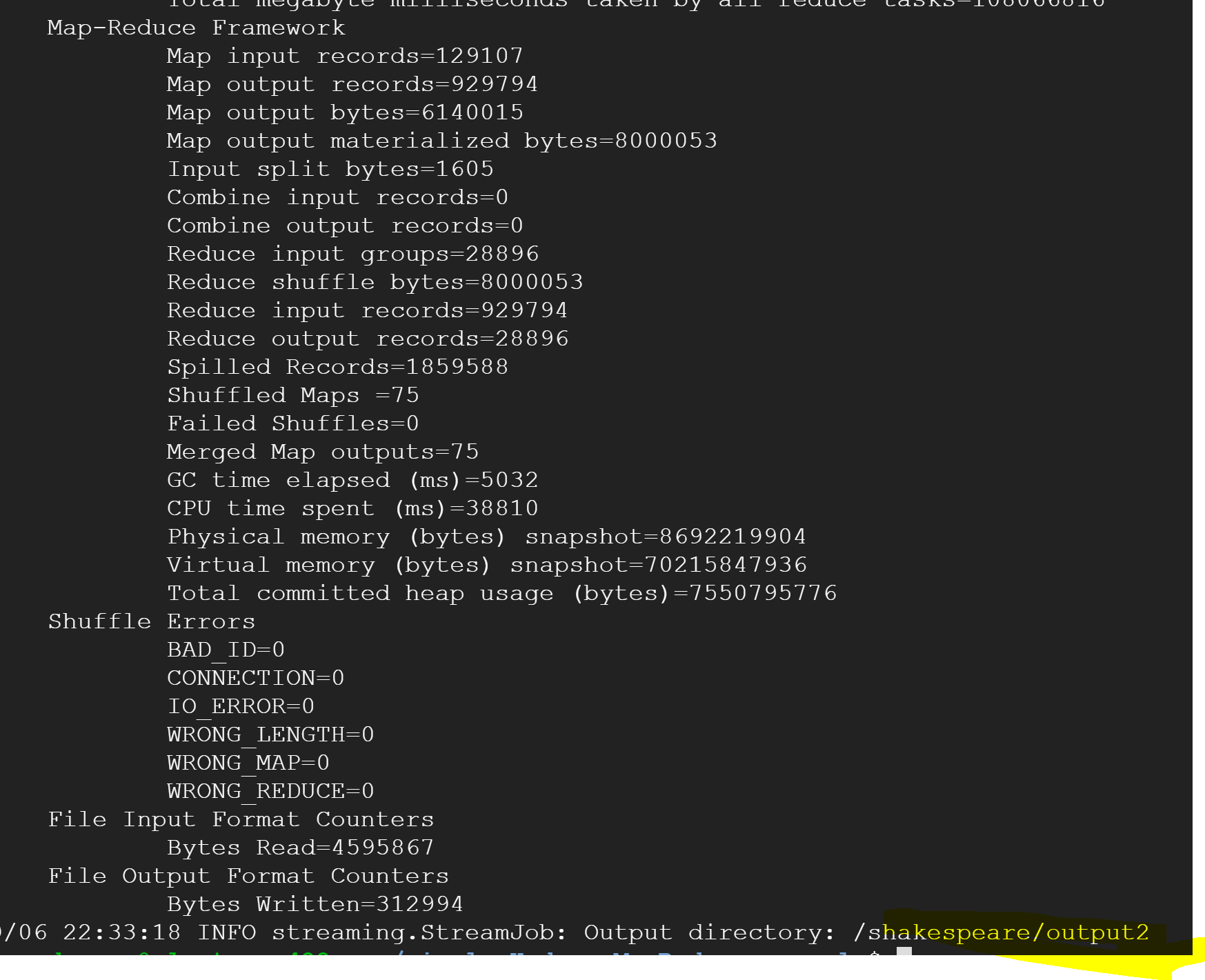
**When nano opens the mapper.py file initially, it will look like the following.**

**Add your new python code. You can save your changes by using Control O (notice the ^o at the bottom of the screen). You will be prompted for the file name and press enter to keep it the same. When you are done editing the file, use Control X (^X). Note: The example below does not exclude punctuation. I will let you add that code yourself.**

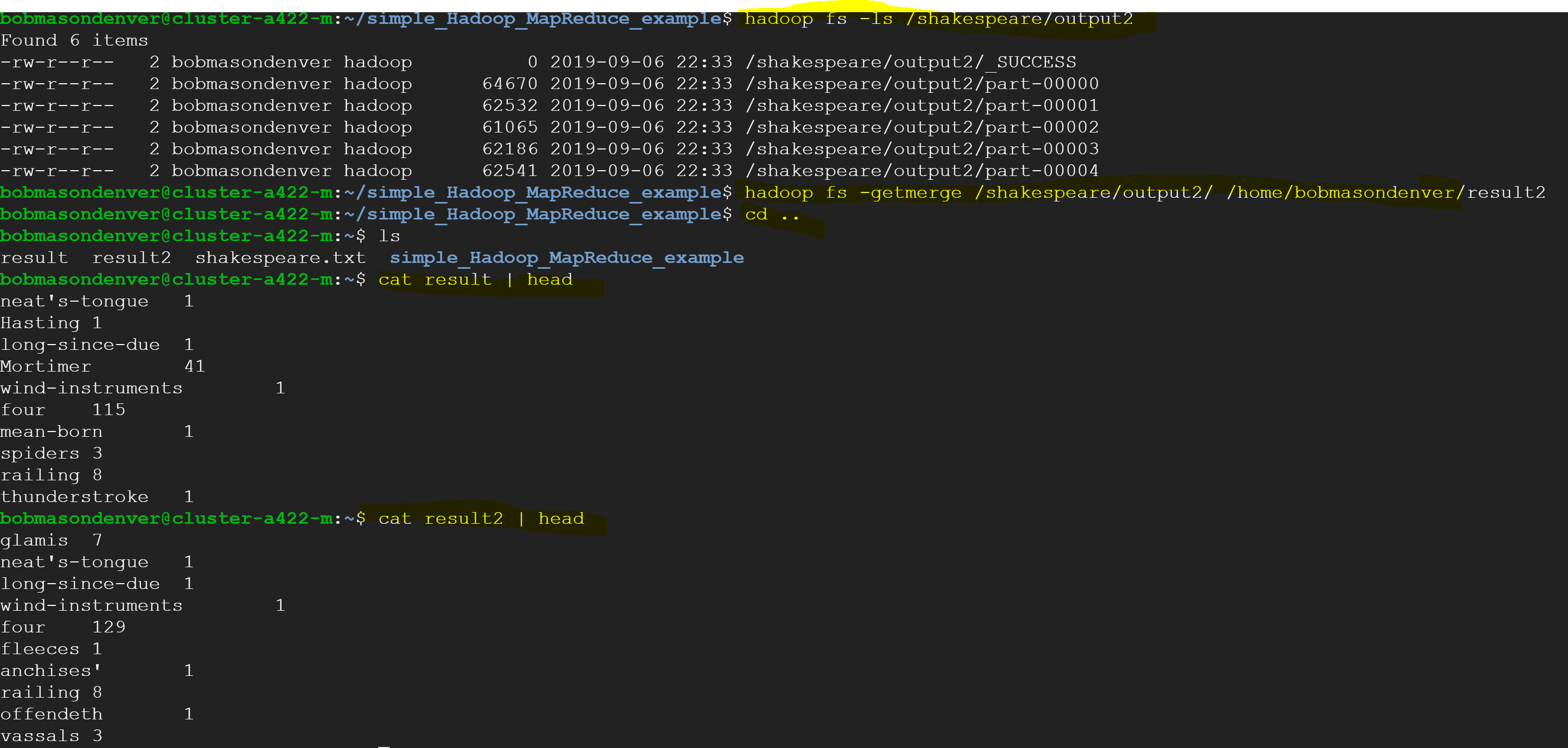


**Try running the map reduce job again. If you get an error that the output file already exists as shown below, try adding another character to the output file name.**

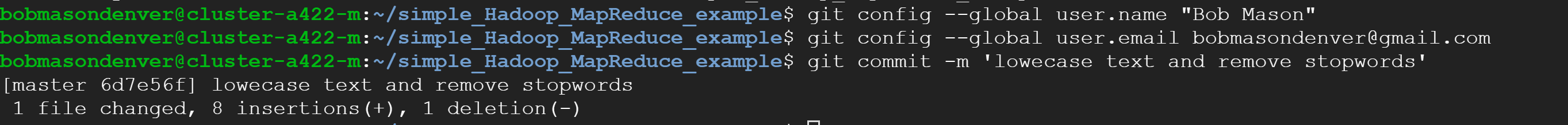




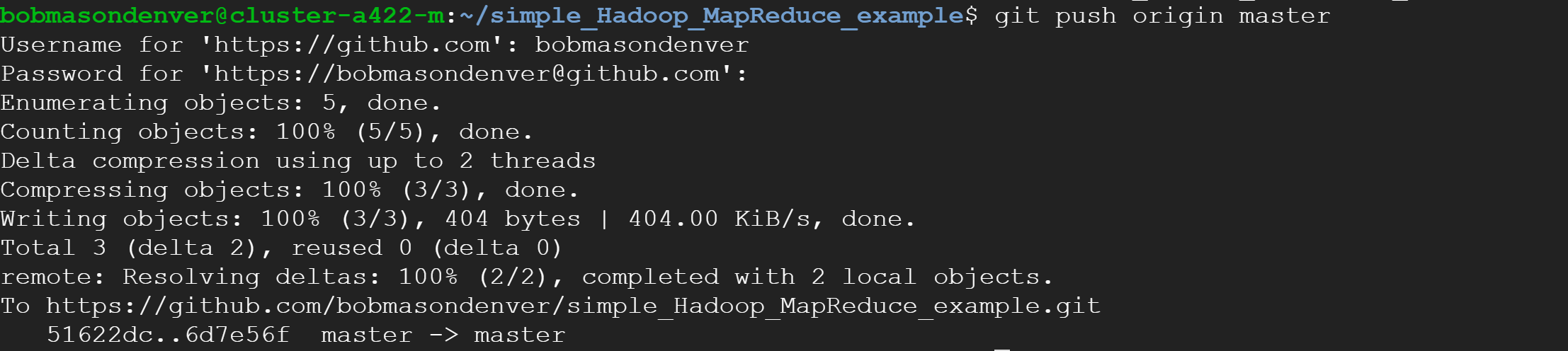
**Notice that the output was written to an output2 directory. Check on the files, merge from the new output into a result2 file. Compare that the files are different using cat. The results I had still have punctuation, therefore your results will be different.**



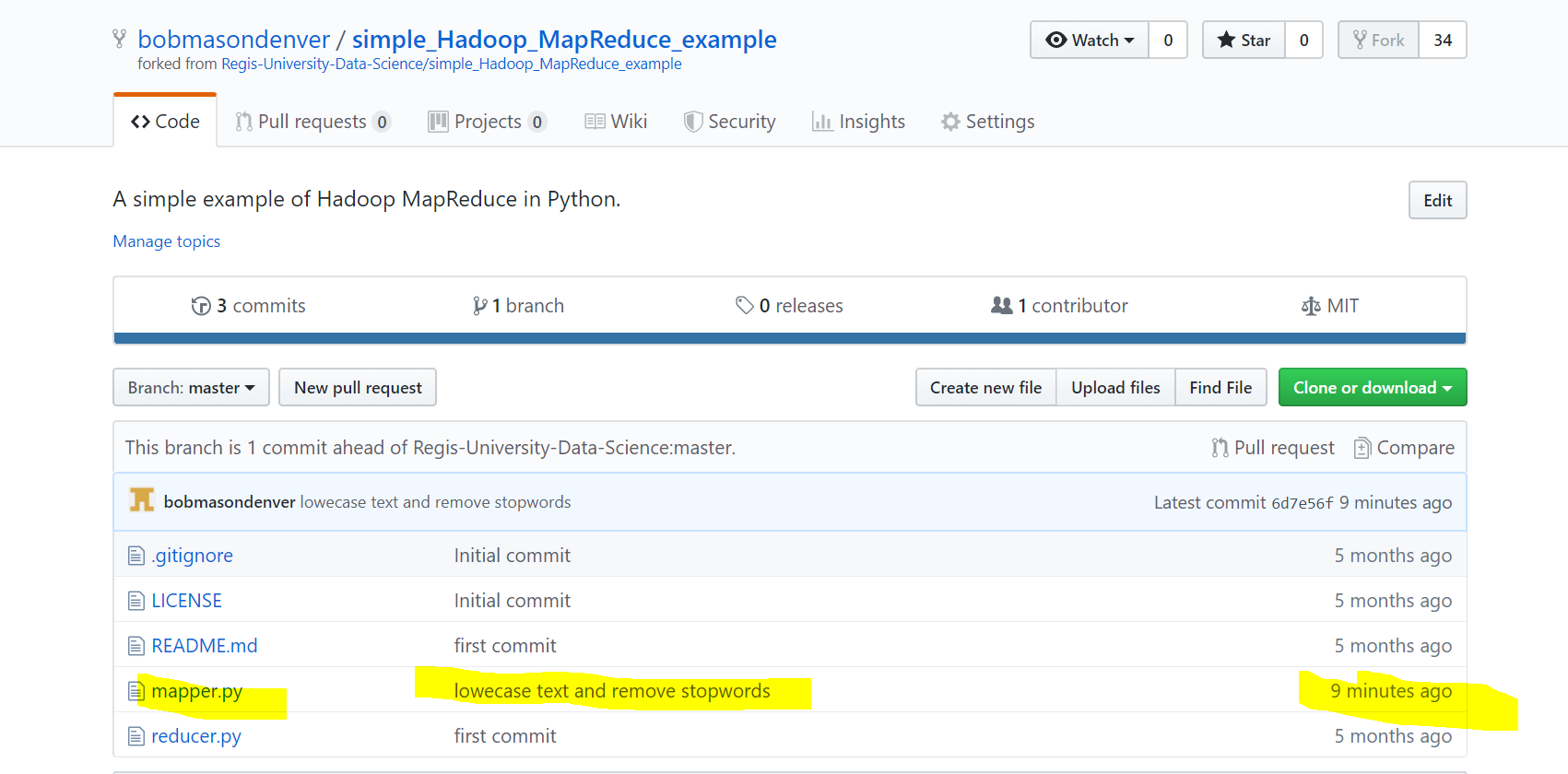
**The last step is to update our github repository. Before I could commit, I had to set my name and email address using the git config user.name and git config user.email commands. Then, I was able to use the commit command.**



**Push the files to the repository. You will need to enter your ID and password.**



**You can verify in github that your changes have been uploaded (as shown below).**



**\*\* include a screen shot of your github updates**

**This is the end of Lab 2 Part 2 Supplement with screen shots. Please submit your final lab report to the drop box using a file name with your last name and lab 2.**