William Oliver

Kyle Gorman

LING 83800

8 March 2021

HW2 Write Up

In this second homework assignment, I built a Python command line tool to read data and split it into training, dev, and test data. My first challenge was running python code on the Windows subsystem for Linux with the shebang. Previously, I had run my code as `sudo pipenv run python file.py`, but, as Kyle mentioned in class, this is not ideal because you can damage your computer with a sudo command if you run the wrong code. Therefore, I planned to now run the code with the shebang and `./file.py` in the command line, but this would give me the error message `no such file or directory`. After I made sure this was not the result of a typo, I googled the problem and found that I needed to install `dos2unix` to run Windows files in Linux. Now, I run `dos2unix file.py`, `chmod -x file.py`, and `./file.py` to run python scripts.

After fixing this issue, the assignment went smooth. For a time, I was attempting to write and split the data using `islice` imported from `itertools`, but I changed my strategy when I was not successful and made a function, which I called `write_tags`, to write out the data.

My last problem was that, initially, my training data contained more than 80% of the total data. I was confused why this was before realizing that I had started to run the script a few times when it had a typo in it that would cause it to stop. I think it wrote the training data but crashed before it wrote the other files. I, then, deleted the existing training, dev, and test files and reran the code, and the word count and sizes of the files were correctly portioned.