Victoria Page
Comp Ling Methods II
March 8, 2021

Homework 2 Write-Up

I didn't find this assignment too challenging, which suggests that I probably should have gone for more of the stretch goals, however I hadn't left myself enough time for that. As far as I can though, I did get it to pass the mypy/flake8/black tests though. The problems I encountered were what I would think are more "typical" issues such as just not applying the correct method or running into limitations of the language/specific method. I also couldn't remember a bunch of what I had previously learned about Python, but some strategic googling was able to help me with that. I'm not sure if it's over-sharing or overly detailed, but what follows below is a quasi stream of consciousness description of issues as I was running into them. Please feel free and do ignore if it is too much:

The main thing that I struggled with was maintaining the data type of each individual word as it was divided up into the data sets. What should have been a list turned into a string, and I wasn't able to solve this problem. I was able to get each word onto its own line with a newline character using writelines() istead of write(), which had initially been a problem as well. I'm not sure how to maintain the type as "list" instead of "str" as even writelines() was rejecting that data type.

One question I had was whether each argument needed to be defined one by one, or whether there was a way to do a bunch of them more succinctly. I had this question because except for the optional --seed tag, they were all simply argparse.add_argument(). I'm wondering if there's some way (using semi-colons or something?) to define all of those basic args in one command or if no, it really does have to all be individual. If in more advanced practice args are so complex that it ends up being better practice to take the time to define everything.

I was relying too much on the practicum at first, copying over the code we used in that. While it was indeed helpful to have a template for argparse, ultimately I wasn't modifying it enough to account for the fact that the homework assignment was inherently different in its goals. This actually was a good reminder that I can't rely on/simply copy pieces of code that I come across; they can be used but must indeed be modified or else they'll hurt me more than they help.

At some point I was wondering if it would be useful to add unittest testing that verifies all of the input data makes it to the train-dev-test sets. When I compared my numbers, I was consistently losing around 10,000 pieces of data. I don't know if that is acceptable or expected based on Conll2000, but it seems like more loss than we would prefer. Said testing could also verify that the data is evenly distributed along the 80-10-10 split, without having to use "word count" manually. I don't know if this wouldn't work though, as doesn't testing have to be performed before running the program on actual data? Would you need to test everything on a set you already have like Conll2000 before applying it to other data? Is this in actuality what we were doing in this assignment? Only less formally, not using unittest?