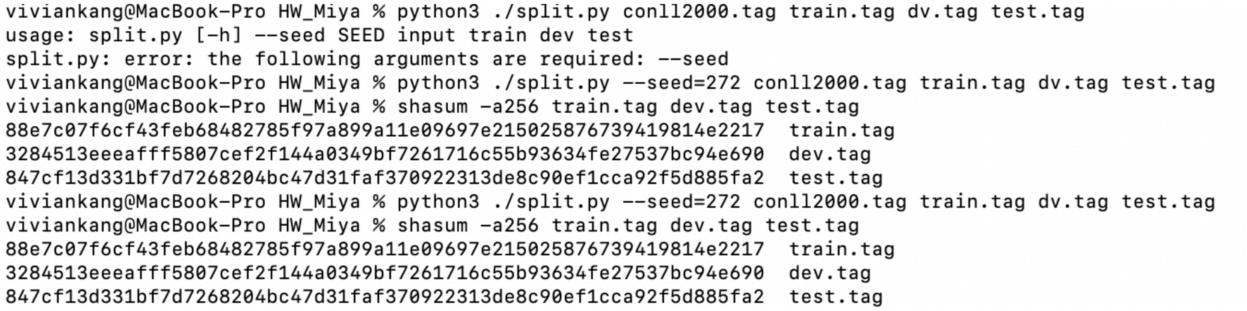
Summary

This python command-line tool is completed to read in tagging data and split it into training, development, and test dataset. Arguments are passed with argparse module; data are read through read_tags and written through write_tags. Randomization is achieved using random.shuffle and a random seed is given.

The result for step 3 is as below.

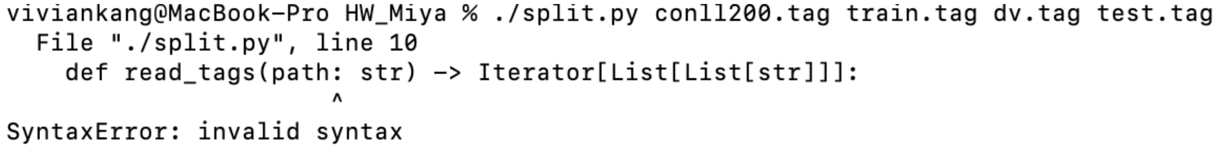


```
viviankang@MacBook-Pro HW_Miya % python3 ./split.py conll2000.tag train.tag dv.tag test.tag
usage: split.py [-h] --seed SEED input train dev test
split.py: error: the following arguments are required: --seed
viviankang@MacBook-Pro HW_Miya % python3 ./split.py --seed=272 conll2000.tag train.tag dv.tag test.tag
viviankang@MacBook-Pro HW_Miya % shasum -a256 train.tag dev.tag test.tag
88e7c07f6cf43feb68482785f97a899a11e09697e215025876739419814e2217  train.tag
3284513eeeafff5807cef2f144a0349bf7261716c55b93634fe27537bc94e690  dev.tag
847cf13d331bf7d7268204bc47d31faf370922313de8c90ef1cca92f5d885fa2  test.tag
viviankang@MacBook-Pro HW_Miya % python3 ./split.py --seed=272 conll2000.tag train.tag dv.tag test.tag
viviankang@MacBook-Pro HW_Miya % shasum -a256 train.tag dev.tag test.tag
88e7c07f6cf43feb68482785f97a899a11e09697e215025876739419814e2217  train.tag
3284513eeeafff5807cef2f144a0349bf7261716c55b93634fe27537bc94e690  dev.tag
847cf13d331bf7d7268204bc47d31faf370922313de8c90ef1cca92f5d885fa2  test.tag
```

The error hint will be presented if --seed is missing. The results of shasum for the same random seed is the same.

Challenges

I was unfamiliar with running python script in terminal so I forgot to use python3 to run the .py file. So, it leads to the error as below.



```
viviankang@MacBook-Pro HW_Miya % ./split.py conll200.tag train.tag dv.tag test.tag
  File "./split.py", line 10
    def read_tags(path: str) -> Iterator[List[List[str]]]:
                      ^
SyntaxError: invalid syntax
```

I tried to locate the error. I changed the main function in the file to make it run in IDE without using argparse. It worked so next I confirmed that my usage of argparse is correct. I also googled for the solution and there is similar problem whose solution is changed the python version to python3 by running with python3 xx.py. It suddenly occurred to me that I need to add python3 when I intend to use python3 to run .py file.

After that, everything is fine.