Module 2

Programming Assignment

Programming assignments are due by 11:59:59 PM on the day of the next lecture. All programming assignments are graded for clarity and functionality. Your code should be well-organized and include helpful comments where appropriate. Please upload your completed assignment to canvas as a zip file.

In this assignment, you will implement the BPE tokenizer training algorithm. Please see tokenizer_trainer.py for details. Note that the tokenization training algorithm can take a long time to run- try testing on very small amounts of text. You do not need to train the tokenizer on anything in particular, just try to verify that it works on your own (when grading the assignment your code will be run against an unseen text dataset).

To assist in this assignment is a file that downloads text which you can use to test your implementation. Use of this data is completely optional.

- **download_data.py** - downloads 100 chapters from famous books and saves them in a single text file, data.txt

---

Your deliverables for this assignment are:

- A completed *tokenizer_trainer.py* script