GAI Project 2.a Arithmetic text generation

- Topic: Arithmetic text generation
- If you have any questions, please e-mail to nckudm@gmail.com (mailto:nckudm@gmail.com).

Scoring Criteria

We provide a sample code of RNN text generation, which you can adjust to an arithmetic version.

1. Data (30 pts)
   We provide an example of arithmetic data on google drive: https://drive.google.com/file/d/1sGVtqG4J5w6kJtsXuLWVEqbD6TY1U2ui/view?usp=sharing (https://drive.google.com/file/d/1sGVtqG4J5w6kJtsXuLWVEqbD6TY1U2ui/view?usp=sharing)

   1. Load the data and split them into train and validation (10 pts). You can use **pandas**, **Dataset** and **Dataloader**.
   2. Generate your own version of data (10 pts), At least three two-digit addition and subtraction problems, and if you'd like, you can also try multiplication and division.
   3. Tokenize the text (10 pts). You can design your own tokenizer or use any API (if available).

2. Generation Models (30 pts)

   1. Model design (10 pts). You can use RNN, GRU or LSTM... whatever. (**at least one sequential model**)
   2. Train(finetune) the model (10 pts).
   3. Evaluate your model when you are training. (10 pts)

3. Analysis (40 pts)

   1. Model analysis (10 pts)
      Include your model design and the process of loss reduction and the results of validation and testing (if available).
   2. Dataset analysis(15 pts)
      What are the characteristics of the datasets you have generated, and what is your method or understanding of it? What adjustments did you make to the dataset? How did they impact your results? (At least 2 variations, 5pts per variation)

   3. Discussion (15 pts)
      During the training process, what impact does different learning rates have? What impact does different batch sizes have? What are the characteristics of the model you are using, and why do you think it is suitable(not suitable) for this task?

4. Bonus (10 pts)

   1. Compare the performance of multiple models and provide a brief analysis. (10 pts)

Submission

- Structure

  - Your should submit a `.zip` file with the name `{student_id}_GAI_Project2a` (eg. `F1234567_GAI_Project2a`). It should be unzipped into a directory with the same name, and the directory structure should be:

```
{student_id}_GAI_Project2a
├── main.py (.ipynb)   // the code you use to run the language models & generate figur
...               // include other scripts if you have more
├── requirements.txt // pip freeze > requirements.txt
└── report.pdf   // your report file, be sure it is .pdf
```

  - TA will not run your code for this project, but please make sure that you hand in the code that train the model(s) and executes the generation.
    - Make it readable **with comments**, lest we would need to refer to it under any circumstances.
    - If your code does not look like it can reproduce the results described in your report, we would consider a grade discount/ask you for a demo.

- Submission Deadline: 4/18 (Thursday) 9:00 a.m.

  - Note that the deadline is 9:00 **in the morning**.
  - Late submission within 1 week will get a 10% discount, and 3 weeks will get a 30% discount.
  - Submissions later than that will not be graded and you get a 0 as a result.

Appendix

We recommend you to run your code on Colab or Kaggle. Of course, if you have your own hardware resources, you can use them as well.

The old-newspapers dataset in sample code :
https://www.kaggle.com/datasets/alvations/old-newspapers
(https://www.kaggle.com/datasets/alvations/old-newspapers)