GAI Project 2.b Text summarization

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

Scoring Criteria

We provide a sample code of T5 English text generation, which you can adjust to a Chinese text summarization version.

1. Data (30 pts)
   Text summarization dataset: https://huggingface.co/datasets/hugcyp/LCSTS (https://huggingface.co/datasets/hugcyp/LCSTS)

   1. Load the "train" and "validation" splits of data (15 pts). You can use **pandas**, **Dataset** and **Dataloader**.
   2. Tokenize the text (15 pts). You can design your own tokenizer or use any API (**recommended**).

2. Generation Models (30 pts)

   1. Model design (15 pts). Unlike Project 2.a, you should use transformer-based model. (Huggingface API)
   2. Train(finetune) the model (10 pts).
   3. Evaluate your model when you are training. (5 pts)

3. Report (40 pts)

   1. Model (20 pts)
      Intrioduce what model you have used in your code (10 pts). Compare the T5 model with GPT2 (10pts) and describe the differences between T5 and GPT2.
   2. Dataset (5 pts)
      Briefly describe your methods to process the data and how to input them into the model.
   3. Train (10 pts)
      Describe how do you train(tune) your model.
   4. Evaluation (5pts)
      Select evaluation metrics (BLEU, rouge, ...) and show the scores.

Submission

- Structure
    - Your should submit a `.zip` file with the name `{student_id}_GAI_Project2b` (eg. `F1234567_GAI_Project2b`). It should be unzipped into a directory with the same name, and the directory structure should be:

```
{student_id}_GAI_Project2b
├── 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.