Recurrent Neural Networks
Homework #1
Due: 2025/4/6

- Overview:
  The "AI_Human.csv" file is a dataset containing text generated by both AI and humans, along with labels. In this task, please preprocess the text data first, then use an LSTM model for the classification task of distinguishing between AI-generated and human-generated text. Aim to maximize accuracy as much as possible. For details of the dataset, see the Kaggle website:
  https://www.kaggle.com/datasets/shanegerami/ai-vs-human-text

- What to do:
  1. Load the dataset and preprocess the text (e.g., tokenize, embed).
     P.S. Due to the large size of the dataset, it is advisable to first divide it into smaller subsets for testing. This ensures that the model is trained stably before switching to the complete dataset.
  2. Train an LSTM model, which can be combined with other models to improve prediction accuracy.
  3. You must split the dataset into training, validation, and testing sets and strive to maximize the accuracy of the testing set.
  4. Create a Github account if you do not have one.
  5. Upload the report (.pdf) and all program files (.ipynb) to Github.
  6. Use your Github URL as the answer to the homework.

- Assignment Evaluation:
  1. Code (40%)
  2. Model Performance (30%)
  3. Report (30%)