README FILE:
Advanced Machine Learning Project - AI Generated Text Detector by Serin Park, Alex Lin, and James Lee (sp5862, acl10011, sjl9802)

Description:
Our project for Advanced Topics: Deep Learning is to evaluate the results produced by AI text detectors. With our analysis, we hope to answer the questions of how well current AI text detectors can classify texts as either AI or human generated as well as how they can be improved. AI text detectors are a relatively new concept as AI is becoming more and more commonplace, but it is specifically used a lot in academic settings. We wanted to see if we could test these detectors in more everyday settings, so we looked at online chat forums such as Reddit and Stackoverflow, which are not always related to academics, and tested the performance of the detectors there.

We generated our own dataset by scraping websites such as Reddit and StackOverflow, and creating AI generated responses to the questions that were posed on there. We removed posts and comments that were too short or of an unusual format, and also cleaned the data by removing any links or files. AFter cleaning, we ended up with around 8000 Reddit comments and 6000 from StackExchange.

The dataset can be found at ‘dataTogetherReddit.npy’.

In our file, you can find the dataset creation, cleaning, and the implementation of the various models as well as their outputs. Everything is labeled accordingly in the notebook, or, if preferred, in the table of contents there are easy subheadings to go to the exact section of interest.

Results and observations:
We trained the networks using RNN, LSTM, GRU, and BERT. These all have different strengths and serve different purposes, that is why we decided to implement all of these models. We also created visuals for the dataset, a precision recall curve, and visualizations for hyperparameter tuning.
From these results, we can see that the AI detection models are accurate at differentiating between human and AI generated texts, which is a good sign as there are many further implementations that can be made with these results.