Homework Two

MSDS Summer 2022

- Submit code via Canvas using Markdown Cells to **clearly** indicate which code answers which question and to answer short answer questions.
- This will be graded for correctness, but **not** model performance.

For part of this homework, you will create a Recurrent Neural Network to evaluate whether is sarcastic or not using the dataset here:
`https://www.kaggle.com/rmisra/news-headlines-dataset-for-sarcasm-detection`

1. Build an RNN model using LSTM and/or GRU layers to determine if a headline is sarcastic or not. Pre-process the text using lemmatization and word tokenization. Do not use a pre-trained word embedding. Make sure you are doing a train/validation split and reporting the accuracy, precision, and recall at the end of each epoch. Train for at least 2 epochs.

- Note: You'll learn about GRUs and LSTMs on Thursday. For now build the pipeline/model using RNN layers, but then replace them w/ LSTM or GRU layers after Thursday.)

2. Repeat question 1, but this time use a more *granular* tokenization than word tokenization such as sub-word tokenization. Again do not use pre-trained embeddings. Did the model improve?

3. Repeat question 1 twice, but make the following changes:

(a) Use a pre-trained word embedding such as glove or word2vec.

(b) Now train a custom word embedding via word2vec on the dataset below:

`https://www.kaggle.com/therohk/million-headlines`

Did the model improve with either of these methods?

4. Use HuggingFace (pre-trained models, tokenization and pre-trained embeddings specifically designed for social media text) to create a pipeline for the task here (already has a train/test split):

`https://www.kaggle.com/datatattle/covid-19-nlp-text-classification`