Lab Five

MSDS Summer 2021

- Submit code via GitHub Classroom using Markdown Cells to **clearly** indicate which code answers which question and to answer short answer questions.
- Please put your name at the top of the Assignment in a Markdown cell.
- Failure to do either of the above will result in points deducted.

1. Create a train/validation split for the IMDB Sentiment Dataset. Train four models and compare results:

(a.) A CBOW model with an embedding layer,

(b.) an RNN model with an embedding layer,

(c.) a CBOW model using a pretrained word embedding,

(d.) and an RNN model using a pretrained word embedding.

Remember to turn off the gradient for the pretrained word embedding layer! You'll have to create a new DataSet for the pretrained word embedding model so that the correct indices are output. You have a couple of options for dealing with tokens unknown to the pretrained embedding:

(a.) Send them all to the zero index, or

(b.) add them to the embedding, randomly initializing the weights. You can then finetune the embedding after a few rounds of training the model (see the cbow notebook from your MSDS 630 course).

2. Use the gensim Word2Vec to train an embedding on the IMDB Dataset. Use this to train either an RNN or CBOW model and compare to the previous models. (This may take a long time; so if you don't finish this it's okay!)

BONUS: Check out the dataset that I scraped here for fun: `https://www.kaggle.com/michaelruddy/new-york-times-recipe-comments`. Train a Word2Vec word embedding on this dataset, and do the following:

(a.) Compare the word embedding you get to a pretrained word embedding. What sorts of words are close to "wine" in the recipe comment embedding versus the general embedding?

(b.) Why might you want such a word embedding rather than one trained on a more general corpus?