Predictive Analytics World / Deep Learning World
Exercises - The IMDB Movie Review Dataset

1. Prepare the IMDB movie review dataset for use by a Keras/TensorFlow program. Locate the zipped or text files and extract them. Write a Python program named make_data_files.py that combines the 50,000 files into a single training data file and a single test data file, filtered by the maximum length (in words) allowed for a review.

Use the keras.datasets format: 0 reserved for padding, 1 reserved for start-of-sequence, 2 for out-of-vocabulary, 3 reserved for future use. Index values should be based on frequency in the corpus, with 4 = most frequent word, 5 = second most frequent word, etc.

2. Write a Python program to create an LSTM sentiment analysis model. Limit the maximum review length to 50 words. Use the built-in Keras Embedding layer to generate word embeddings as part of the model rather than external embedding from tools such as gensim (custom embeddings) or GloVe (global English embeddings).

3. Modify your program to make a prediction for a new, previously unseen review such as, "I wish I could say I liked this movie but I can't." Alternatively, modify your program to save the trained model, and then write a new program named something like use_imdb.py that loads the saved model and then makes the prediction.

4. Which statement is most accurate?

a.) The key distinguishing characteristic of LSTM cells is that they maintain state.
b.) The key distinguishing characteristic of LSTM cells is that they are unsupervised systems.
c.) The key distinguishing characteristic of LSTM cells is that they are generative rather than predictive.

5. Which statement is most accurate?

a.) You can add an "Attention" component to LSTMs which is especially useful for seq-to-seq problems.
b.) You can add a "Confusion" component to LSTMs for seq-to-class problems.
c.) You can add a "Horizon" component to LSTMs for class-to-seq problems.

6. Which statement about peephole LSTM cells is most accurate?

a.) A peephole LSTM usually uses c(t-1) rather than h(t-1) for most connections.
b.) A peephole LSTM is also known as an Elman network.
c.) A peephole LSTM is also known as a Jordan network.

7. Which statement about gated recurrent unit (GRU) networks is most accurate?

a.) GRUs have fewer parameters than LSTMs because they don't have an output gate.
b.) GRUs have more parameters than LSTMs because they have two hidden cell states.
c.) GRUs and LSTMs are they same -- the term "GRU" was used prior to 2014, now "LSTM" is used.