Homework

Reading

- Raschka: Chapter 8 (ML & Sentiment Analysis), Chapter 16 (RNNs), Read Karpathy on RNNs, colah on LSTM, and why we might not want RNNs
- Optional prep for next week: Attention is All You Need

Coding

- Build a sentiment classifier (negative=0, positive=1). Use the Stanford/IMDB dataset from the sentiment embedding example. Model(s) should be neural nets.
    - First, create a model using either bag-of-words or tf-idf implementations. (One-hot or wordcount vectors as discussed in class). This will be your baseline.
    - Second, use embeddings as the input to a neural network. Importantly, figure out how you want to combine word vectors to document vectors (e.g., average of word vectors, tf-idf weighted avg, etc.) - please don't directly use doc2vec.
    - Finally, build either an LSTM or CNN using w2v embeddings
- Compare model quality. Report accuracy, precision, recall, and F1 scores. Discuss dimensionality of input to your models - what's different? Does one take longer/shorter to train?
- Find a few examples where your models disagree in their test data predictions. Which model got it right? Is the "wrong" inference reasonable/intuitive? Are there any patterns you see in what the models get wrong?
- You can use this sentiment embedding example for your embeddings if you want (also fine to use pre-trained embeddings).
- You can get extra credit comparing custom w2v embeddings vs pre-trained.