Adam Noack
Intro AI
Hw6
22.4)

I used this dataset of spam mail: https://www.kaggle.com/uciml/sms-spam-collection-dataset. The dataset has 5572 samples and 747 of these were spam. The remaining samples were legitimate messages.

I initially planned on removing punctuation and converting all of the words in each data sample to lowercase, but I noticed upon looking at the data that many of the samples labelled spam used uppercase lettering and excess punctuation, so I kept the casing and punctuation. It also seemed like the presence of many unigrams in the samples such as free, urgent, winner, cash, prize were correlated with spam class. Bigrams or n-grams of any sort didn't seem to be heavily correlated with a particular output class. The message length didn't seem to be correlated with a particular output class. Neither the sender's address nor the time of message arrival were included in this corpus.

After loading the data into a jupyter notebook, I tokenized the samples. The words with the highest frequency were mapped to the lowest index values. I then trained an LSTM NN to classify the data. The samples passed through an embedding layer, then a dropout layer, followed by an LSTM layer and a dense layer. I initially trained the model for 5 epochs, but the training accuracy seemed to rise significantly above the accuracy on the validation dataset, indicating overfitting, so I reduced the training amount to one epoch. After training, the model achieved a classification accuracy of 98% on the held-out, untrained-on test dataset.

22.7)

Words with first letter capitalized + Inc, LLC, group, LLP, LLLP, Corporation, incorporated, trust, joint venture, jv, assoc., assn, Ltd., Company, Co, etc. The string of words preceding the LLC, Ltd, etc. will probably not be too long.

Resulting regex:
([A-Z][a-z]+\s){1,7}\b(Company|Co|Limited|Ltd|Inc|LLC|Group|LLP|LLLP|PLC|Corporation||Corp|Incorporated|Trust|Joint Venture|JV|Assoc|Assn)\b
Corpus: https://money.cnn.com/data/markets/

	Classified as non-company name	Classified as company name
String of consecutive words with capitalized first letters that don't describe a company	210 strings of words	0 string of words
Company name	Apple, Citigroup, General Electric, Google, Microsoft, Starbucks, Unilever. 8 in total.	Dollar Tree Inc, Nielsen Holdings Inc, Activision Blizzard Inc, Boeing Co, Quest Diagnostics Inc, Western Digital Corp, Twitter Inc, Wynn Resorts Ltd, Nordstrom Inc. 9 in total

Recall: 9/(9+8) = .53
Precision: 9/(9+0) = 1.0
Accuracy: (9+210)/(9+210+8+0) = .96