CSCI 315: Artificial Intelligence

Assignment 4: Deep Recurrent Neural Network - Data

Due Date: Saturday March 20th @ 11:59PM

Information

For this assignment you will be implementing a recurrent neural network via Keras. Your first step is to manipulate the data associated with this assignment. You will also write a partial analysis about what you did to manipulate this data as well as why it is necessary for learning to occur.

For this assignment you have been given access to consumer\_complaints.csv, which is a csv file of companies’ responses regarding the financial services provided by the United States. This is textual information, which RNNs are exceptional at understanding and analyzing. However, they cannot do this under normal circumstances. RNNs require a fair amount of preprocessing the data in order for the model to function appropriately. This assignment will be accomplishing that goal.

Source Code

There are several steps that need to be accomplished for this preprocessing step to work correctly. Those steps are: 1) removing null or meaningless data points, 2) making the data all lower case, 3) removing symbols, punctuation, brackets, parenthesis, and special characters, 4) removing repetitive and meaningless characters (x in this case), 5) removing stop words, 6) tokenize the remaining natural language, 7) transform that language into a numerical representation, and 8) pad those embeddings so that all input is the same length.

I have given you a file with comments that should help guide your process. You are working with a partner so you should both try to get a good understanding of what is taking place and why it is taking place. Once the data is in the correct format, you are done with this part of the assignment. You can see the fruits of your labor by printing out the tokens for the first x\_train example and seeing if you accomplished all eight of your goals.

The data you are working with is stored in the data column product, which will be our target classification, and the consumer complaint narrative, which is the review the user gave about the product. The narrative text needs to be cleaned and prepared. You also need to preprocess the product such that it is represented numerically. This is best done with the pandas API call get\_dummies.

Turn In

The README is about 30% of your grade, which will solely be focused on the purpose and reasoning behind the eight steps you are to complete. All I want the readme to do is tell me why you need to do these eight steps and why RNNs need data in this format to function correctly. If you have questions about my expectations, please ask and reread the assignment. Please zip up your code and the readme.