Using Deep Learning to Create Fake Yelp Reviews

Elizabeth Willard & Jeffrey Edrington

Background

You can use a neural network for text generation. The model will learn the likelihood of occurrence for a word based off of the previous patterns of words used. Language learning uses long short term memory (LSTM) for pattern prediction.

Database

The database that we used was from Yelp. These are the files are in the database:

yelp\_academic\_dataset\_business

yelp\_academic\_dataset\_checkin

yelp\_academic\_dataset\_review

yelp\_academic\_dataset\_tip

yelp\_academic\_dataset\_user

There is almost 5GB worth of data in this database, but for this project, only the reviews were pulled and used.

Loadthedataandprocess ittoremovepunctuation, and

ObjectivesCreateandtraintheneuralnetworkusingalongshort

tokenize it

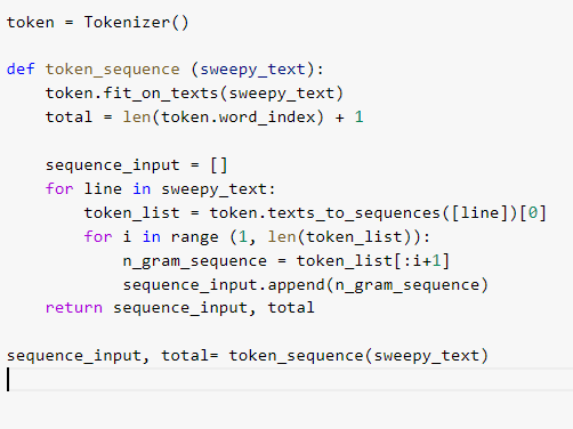
termmemoryarchitecture

Createafunctiontowritefakereviewsusinga"seed"

Tokenization

What exactly is tokenization?

In order for the model to read in the reviews, they all need to be converted to a "sequence" that is legible for the model to read in. Keras has a library function for tokenization that can be called by using 'from keras.preprocessing.text import Tokenizer'

\*sweepy\_text is the cleaned up text for 

processing

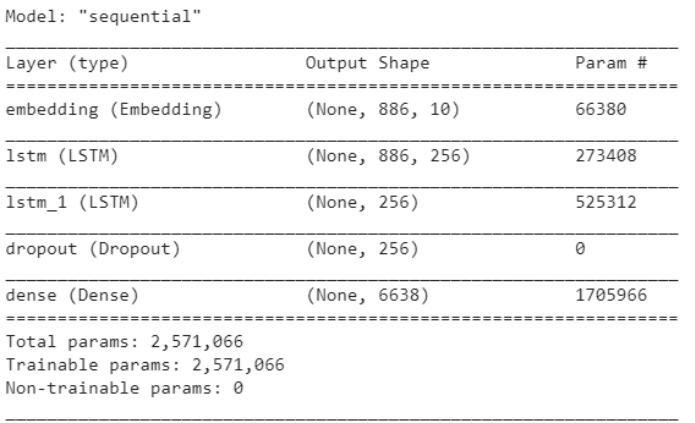
What exactly is LSTM?

Language modelling is used for a variety of purposes, such as speech-to-text. This uses a recurrent neural network (RNN), which propagates from input to output repeatedly and never moves in one direction. By allowing propagation in all directions like this, the neurons can "remember" what has been learnt.

The LSTM affects this by adding an additional state to the RNN that allows the neurons to be more selective about what information it retains.

Coding the Model

There's only three layers in this model, the input, the LSTM, and the dropout layer, which is the bare minimum amount of layers. We used 100 epochs since text generation and RNNs consume a lot of computer power. This model would take about six hours to compile each run.

Model Summary

On the left, as you can see, are the LSTM, which are crucial.

Results Part 03

Resultsfrom theNeural Network

Review#1

Thisplaceisarealgembutiwaslookingforthe foodcomehungryforblackfoodiconcurwiththe borderandcorepriceoverthekidsisreadyto chargenewlockstoharrypotterinotherratethe foodispoliteandcozyismoreabitofacrunchand idecidedtohavewasbetter

It would be very easy for fake reviews to be generated on any websites, so be careful! There are already companies that sell fake reviews, with a better model, you could use this model to detect fake reviews.

**Lessons**

LSTM models require a great deal

of computing power, more than

the average

consumer/programmer may have

on hand.

There is room for improving this model, but if we had more data, we could train this model even more, and adjust the architecture of the model.