Previous:

* Followed 2 tutorials on creating tensorflow chatbot
* 1) Simple DNN network on small dataset picking random answer from tag as response
* 2) Large Reddit dataset that uses nmt (neural machine translation) model trained on 100000 pairs + to create responses.
* Downloaded amazon QA dataset –

6/03

* Follow tensorflow tutorial on text classification

From the tutorial: Build a tf.keras.Sequential model and start with an embedding layer. An embedding layer stores one vector per word. When called, it converts the sequences of word indices to sequences of vectors. These vectors are trainable. After training (on enough data), words with similar meanings often have similar vectors.

This index-lookup is much more efficient than the equivalent operation of passing a one-hot encoded vector through a tf.keras.layers.Dense layer.

A recurrent neural network (RNN) processes sequence input by iterating through the elements. RNNs pass the outputs from one timestep to their input—and then to the next.

The tf.keras.layers.Bidirectional wrapper can also be used with an RNN layer. This propagates the input forward and backwards through the RNN layer and then concatenates the output. This helps the RNN to learn long range dependencies.

7/03

* Look into reddit data to create conversational dataset with multiple threads from a conversation
* Installed bigquery command line to read the data
* Costs reached free tier limit
* Running dataflow command caused recursive error…
* Fixed by running outside virtualenv, installing pip
* curl [https://bootstrap.pypa.io/get-pip.py -o get-pip.py](https://bootstrap.pypa.io/get-pip.py%20-o%20get-pip.py)
* sudo python get-pip.py
* from here ran pip install apache-beam[gcp]
* ran the dataflow script outside virtualenv – now up and running, constantly checking quota limits to keep an eye on costs not
* wanted to extract score for each comment in reddit data. added score to createdata.py
* changed data structure of JSON to be nested and clear with scores
* made reddit dataset really small on big query for testing for structural purposes

8/03

* follow reddit tutorial
* cloned github repo
* used their data (as it was more than I had…. And claims to be ‘small’ amount of data)
* upgraded tensorflow, uninstalled tensorflow, installed tensorflow 1.14 - worked