I started working on the challenge on Saturday and the steps I undertook are:

Setup local environment for NLTK processing by installing the important packages such as NLTK, SCIPY, PANDAS, SKLEARN, NUMPY and KERAS. Normally all of the packages aren’t available in a single Python software or IDE; and either they have to be individually installed into python or you get them assembled as a group in anaconda excluding Keras.

After setting up the environment, I studied and analyzed the preprocessing, evaluation, model, chabot and main codes, which posses defined functions of five, five, eight, three and two respectively, that are present within the python files.

My study of the codes is incomplete as I got stuck in a case in which a little tip from an experienced colleague would have made a big difference to proceed in finding a solution to the problem.

After running the preprocessing, evaluation, model and chabot scripts, I couldn’t observe (notice) any bugs or error messages in the command window. But, after executing the ‘main’ script I found an error pointing to **LINE 59** of the preprocessing script linked to LINE 43 of the ‘main’ script. I would love to carry on working on the bug and resolve the challenge, but I run out of time to do it. Moreover, given some more time along with some communications for help I need for the challenge I would write the requested algorithm and test, which model works better on the supplied data. As I have indicated before, I squeezed this challenge along with the commitments I have from my current work.

Please find enclosed below some errors and messages thrown to the command window during the computation of the challenge.

**As for the question about awareness of having a little training data, here below is a statement about my understanding and response for that:**

The larger the training data set is made, the better the model will be trained. Every data received in data science is normally split in to two data subsets: the training data set and the test data set. The training set is the data in which a model is trained at; the test set on the other hand is the data in which the performance of the model is measured at. Data needs to be randomized before we split it into training and test data sets. This would make sure we don’t use data of different characteristics for training and testing a model.

However, in the case of small data set, a more sophisticated approach like cross validation should be carried out to better train a model and have better confidence on the evaluation metrics.

In this attempt, after downloading the “punkt”, an abstract class, which tokenizes text using, unsupervised algorithms I managed to resolve another issue. Enclosed is a screenshot that shows I resolved and another error and got stuck in the after. Currently I am working on this one as well.

Is there any layer type that you want me to use in task 1 and 2 or shall I use those I am familiar with.

Kind regards

Zenawi

Model training in CNN is an iterative procedure; hence we need to specify the number of iterations that a given model needs to be trained. In the language of machine learning those iterations that elapse to train a model are commonly know as **epochs**. Another useful parameter in neural network is **batch size**. This parameter is responsible for how many samples are used in a single epoch *i.e.* how many samples are used in one forward/backward pass.