בס"ד

STEVE GUTFREUND – 342873791
YOEL JASNER – 204380992

- TRAIN set size: 1000
- DEV set size: 500
- TEST set size: 200

How?

We created an embedding matrix; each line represents an embedding vector for the digits [1-9] and [a-d].
Every input-sequence is split into a list of its characters. Every character is being translated into its vector representation (embedding vector). Next, they are fed one by one to the LSTM layer. We are interested only in the last output of the LSTM layer, which is then forwarded into an MLP with one hidden layer.
We end up with a 2-dim vector, containing probabilities for each class ("good" or "bad").

Results

The model indeed successfully distinguishes between the languages.
It takes on average between 3 to 5 epochs to get there. After 5 epochs, with high probability, the loss value is less than 0.0001.
The running time is between 24 and 25 seconds (about 4-5 seconds every epoch).

The results are for the train, dev and test sets.
In order to test the accuracy on the test set, which is a unlabeled set, we created test_labeled file, containing the exact same data as 'test' but labeled. With a simple script we can check that out test.pred is exactly the same as test_labeled.

At the beginning the model for some reason got always stuck at 99%, no matter how many epochs. The solution was to change from SimpleSGDTrainer to AdamTrainer, for this task it's a better optimizer.