CSU44061 Machine Learning | Weekly Assignment 9
Brendan McCann | 20332615

Part (i)

(a)

Briefly describing the following datasets:

- `input_childSpeech_trainingSet.txt`

  This dataset contains a series of short sentences with poor grammar and vocabulary - in other words, by a child. The vocabulary size is 40. The sentence length ranges from 13 to 85 characters long.

- `input_childSpeech_testSet.txt`

  This dataset contains a similar style as the previous one, except with different and unseen sentences. The vocabulary size is also 40, with a sentence length ranging between 13 and 83 characters long: very similar to the training set.

- `input_shakespeare.txt`

  The style is much different. This text is from a play, so we have character names preceding their lines - for example: `MARCIUS:...`. The language is richer and more advanced. The vocabulary size is 25,671 and sentences range from lengths of 2 to 63 characters.

(b)

Seeing as the vocabulary of `input_childSpeech_trainingSet.txt` is very small and limited, I chose to reduce the number of word embeddings in the model from the default 384 to 40, which is the vocabulary size (unique characters). This reduced the model parameters to 127,320. My intuition was that it is unnecessary to have more word embeddings than the vocabulary size.

I kept the other hyper-parameters at the same values because the model parameter size dropped drastically after reducing the number of word embeddings.

(c)

(d)

(e)

(Part (ii))

(a)

(b)