Angelina Pustynskaia

MT ÜBUNG 4
RNNs

For the preprocessing, I have chosen the Corpus of Oz (Australian and New Zealand) Early English, which I downoalded from this website https://www.ausnc.org.au/corpora/cooee and used it for my research project in another class. The corpus consists of 1354 files in English language and has 1,545,163 words. I created a script (preprocessing.py) that converted the used corpus to lower case and removed all special characters, such as exclamation marks, periods, dashes, etc. This is important as learning upper/lower cases and special characters would require a larger corpus. Like this, we can learn how words and characters depend on each other solely.

As hyper parameters, I used the following ones:

- Epochs: 10
- Batch Size: 20
- Vocabulary max. Size: 10000

  All by default.

  I had a rather larger corpus, so smaller value did not make sense. On the other hand, bigger values did not improve performance much, so I decided to keep the default ones.

I tried out different parameters for the number of neurons and learning rate, but could not observe any significant effects.

On the dev-set I reached the following *perplexity* (after epoch 10): 116.52
After the *scoring* step, the calculated perplexity was 150.55

I also could not connect to the port to work with tensorboard. That is why I could not finish the exercise.