MT - Exercise 3: RNNs and Language Modelling

Present your chosen data, does it have any special attributes that you expect to have an influence on the text generation of the model?

The data set we choose for this exercise was the Declaration of Independence of the USA, downloaded from gutenberg.org. We chose this data set due to the following special attributes that could potentially have a significant impact on the text generation of the model:

1. For the declaration was written using an English vocabulary that is not used to communicate in this day and age anymore.
2. The document was phrased in a clear and direct way that could be understand by the general population. Additionally, it's quite short and is straight to the point, so we expect that the generated text will likely has less adjectives and flourishes than other data sets.
3. The declaration uses metaphors like 'the tree of liberty' to help visualise the concepts and ideas written.
4. We noticed that the text is filled with commas and conjunction words which will probably effect the output sample text.

Take a look at the sample generation, what are your impressions?

Looking at the output sample text, we've realised that the English isn't quite up-to-date as seen in the following phrase "the quest to which revere hands into attention shall".
Furthermore, just like we expected, the sample text contains many commas and conjunctions words which ultimately results into longer sentences.
Overall, the generated text clearly has a political tone to its writing which can be expected looking at the topic of our data set.

Can you see a connection between the training, validation and test perplexity? Based on your results, which dropout setting do you think is the best and why?

The following connection can be established between the validation and test perplexity: With every epoch the validation perplexity is reduced. In addition, the test perplexity is dependent on the dropout setting: The larger the dropout, the more '<eos>'-placeholders appear in the sample text and thus, leads to more gaps in the sentences which make the context harder to interpret. This leads to a higher perplexity of the sample text.
Based on our results, we believe that the dropout setting of around 0.3 is best suited because a bigger dropout will result in larger text perplexity. As you can see in the line chart, a dropout of 0.7 will exponentially increase the test perplexity.

Sample some text from the model that obtains the lowest test perplexity, for instance by changing the script scripts/generate.sh. What do you think of its quality? Does it resemble the original training data?

Looking at the sample text with the dropout setting of 0.0, the text has a high quality: Even though some ‘<eos>’-placeholders are embedding in the text, they don’t disrupt the reading-flow and don’t have a negative impact on the understanding of the sample text. Comparing the sample text with the dropout of 0.0 with the dropout of 0.5 (=> the original dropout number) the dropout of 0.0 is significantly clearer to understand than the sample of the original training data. The sample of the original training data consists of incomplete sentences as well as grammar mistakes, inconsistent setting of punctuations and incoherent context.

Sample some text with the highest test perplexity. Can you see a difference to the lowest scoring one?

The differences are similar to the sample of the original training data: In this case, the grammar, the punctuation as well as the context are even worse and less understandable. Additionally, more ‘<eos>’-placeholders appear in the sample text.