Olga Klimashevska, 12745246, MT19

Exercise 4

As the dataset I took the novel “Anna Karenina” by Leo Tolstoi from https://www.gutenberg.org/files/1399/1399-0.txt as it matched the size: it is roughly 2 Mb. In order to speed up the process and due to redundancy of the first lines I used the following command to create a shorter dataset:

```
 head -n 25057 1399-0.txt | tail -n 25000 > text.en
```

The problem was, however, that the text was split by the width, and not sentence-wise. In order to return sentence-wise text I applied the script split-lines.py to “text.en” which create “text_lines.en”. The last line was not a complete sentence so I dropped it and the final name is “text_lines_final.en”. It has 10361 rows. Then I shuffled it to “text.shuf” and created dev data text “text.dev” with 1036 rows, since it is 10% of the content and less than 2000 rows. The rest was put to training data “text.train”. I then normalized the data and afterwards tokenized it so that “.” is a separate token via moses-decoder software from Exercise 2. I used the last outputs: “text.train.tokenized.en” and “text.dev.tokenized.en” for training and scoring correspondingly for romanesco.

Model1 was default model. The perplexities are:

```
2019-04-30 06:25:08,960 - INFO - Perplexity on training data after epoch 8:
 179.87

2019-04-30 06:25:37,974 - INFO - Perplexity on training data after  epoch 9: 148.38

2019-04-30 06:26:06,934 - INFO - Perplexity on training data
 after epoch 10: 127.02
```

For Scoring: Perplexity: 397.48

Model 2 had batch size of 200, since it’s bigger than the default value. The perplexities are

```
2019-04-30 11:16:09,732 - INFO - Perplexity on training data after epoch 8: 459.42

2019-04-30 11:16:33,794 - INFO - Perplexity on training data after epoch 9: 446.24

2019-04-30 11:16:57,678 - INFO - Perplexity on training data after epoch 10: 415.46
```

For Scoring: Perplexity: 124.31