Text generator on Kazakh language (Report)

Link to github: https://github.com/abensad/Text-Generator-on-Kazakh-Language

(I add to github dataset and trained model)

Link to youtube: https://youtu.be/hgEzVzJVImI

(Sorry, I could not keep it in 1 minute)

Introduction

Problem

Problem is creating text generator on Kazakh language.

Literature review with links

There are two ways of generation text: by transformers and by RNN. I choose character based RNN. I analyzed several character based RNN models and noticed that they use LSTM and Dropout layers.

Current work

I just read some tutorials and try to understand how to create generator.

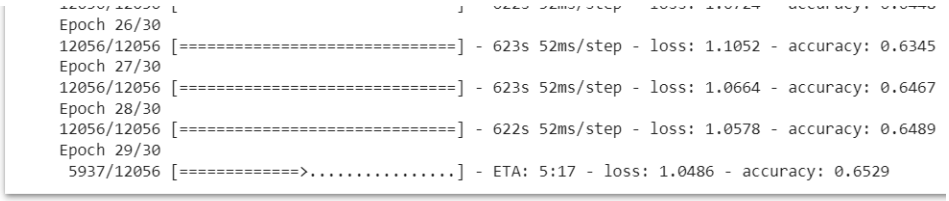
Data and Methods

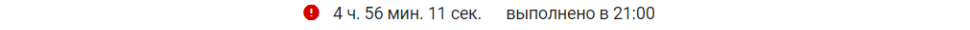
Information about the data

Firstly, I want to use bigger dataset to better accuracy. But I realized that there are limits in use GPU on Google Colab. After I decide to use not 2 toms of «Абай жолы», but only one tom. It means that number of symbols decreased from 1.6 million to 0.7 million.

Description of ML models with some theory

Initially, I want to use bidirectional layers. But when I use it, model has more parameters. And the more parameters they have, the more time it takes to train. But there are limits in use GPU on Google Colab. Because of this sometimes training stopped. There example:





After I decide to use only LSTM and Dropout layers.

Results

As you see before optimizing model, it trained about five hours. After optimizing it trained 1 hour 40 min. Its accuracy 68%:

```
Epoch 30/30
5605/5605 [==============================] - 184s 33ms/step - loss: 0.9423 - accuracy: 0.6851
```

It is not high accuracy. But it is necessary for generation grammatically right texts on Kazakh Language.

Discussion

Critical review of results

Text generator generates only text with lowercase and without punctuation. Consequently, there no sentences and it is almost meaningless text. If you input text which is from “modern” speech, text will output just random words with no sense with input:

Seed: ягами лайт

Generated text:

а жөнелді

абай жаңа жанылдырмай қарап алып кейде жатып қалған сонысы барлас жақындап қалды абай барлады

абай осы кезде қалған бір топ жандар да жаңа келеді

байдалы бар

сонысы екеуі де қарабастың балаларының алдынан танып келе жатқан топ жан жаққа қарай баса бергеннен байқап болған күн ішіндегі барлық жайын абай ең алдымен түсіп жақындап қалды абай барлады

абай осы кезде қалған бір топ жандар да жа

Next steps

To generate better text model should train on bigger dataset. Moreover, dataset should be complex. Dataset should consist of not only from one book, but also from different magazines, articles and modern classics.

Another possible improvement is better architecture. Model should have Embedding, Bidirectional layers and more LSTM and Dropout layers.

Sources

https://www.tensorflow.org/text/tutorials/text_generation

https://towardsdatascience.com/generating-text-with-tensorflow-2-0-6a65c7bdc568

https://www.thepythoncode.com/article/text-generation-keras-python

https://blog.paperspace.com/bidirectional-rnn-keras/

https://coderzcolumn.com/tutorials/artificial-intelligence/keras-text-generation-using-rnn-and-character-embeddings