Character prediction

dataset:

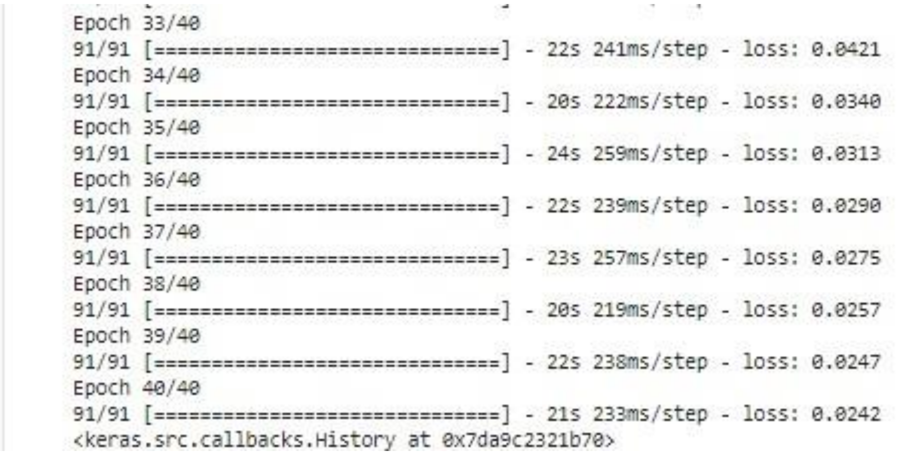
The character-based prediction model is designed to generate accurate predictions based on a dataset consisting of 4 pages from Wikipedia. Before applying the model, the dataset is cleaned by removing HTML tags, URLs, punctuation, white spaces, and digits. The text is then tokenized at the character level.

We use two model

Model1(LSTM):

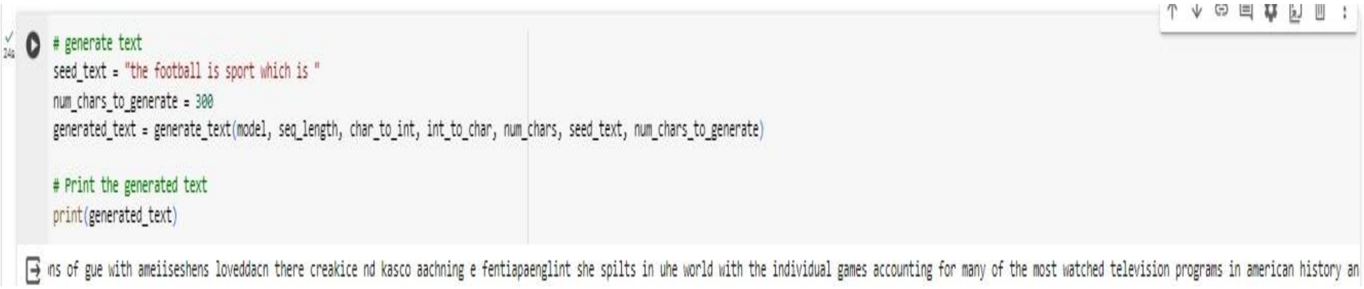
The model architecture starts with 2LSTM layers, one containing 256 And the another contains 128 neurons. The purpose of these layers is to capture the sequential patterns and dependencies present in the input sequence of characters. Finally, the model ends with a softmax layer with a number of units equal to the total number of unique characters in the dataset.

Accuracy: the model loss is decreasing



```
Epoch 33/40
91/91 [==============================] - 22s 241ms/step - loss: 0.0421
Epoch 34/40
91/91 [==============================] - 20s 222ms/step - loss: 0.0340
Epoch 35/40
91/91 [==============================] - 24s 259ms/step - loss: 0.0313
Epoch 36/40
91/91 [==============================] - 22s 239ms/step - loss: 0.0290
Epoch 37/40
91/91 [==============================] - 23s 257ms/step - loss: 0.0275
Epoch 38/40
91/91 [==============================] - 20s 219ms/step - loss: 0.0257
Epoch 39/40
91/91 [==============================] - 22s 238ms/step - loss: 0.0247
Epoch 40/40
91/91 [==============================] - 21s 233ms/step - loss: 0.0242
<keras.src.callbacks.History at 0x7da9c2321b70>
```

Result:



```
# generate text
seed_text = "the football is sport which is "
num_chars_to_generate = 300
generated_text = generate_text(model, seq_length, char_to_int, int_to_char, num_chars, seed_text, num_chars_to_generate)

# Print the generated text
print(generated_text)
```

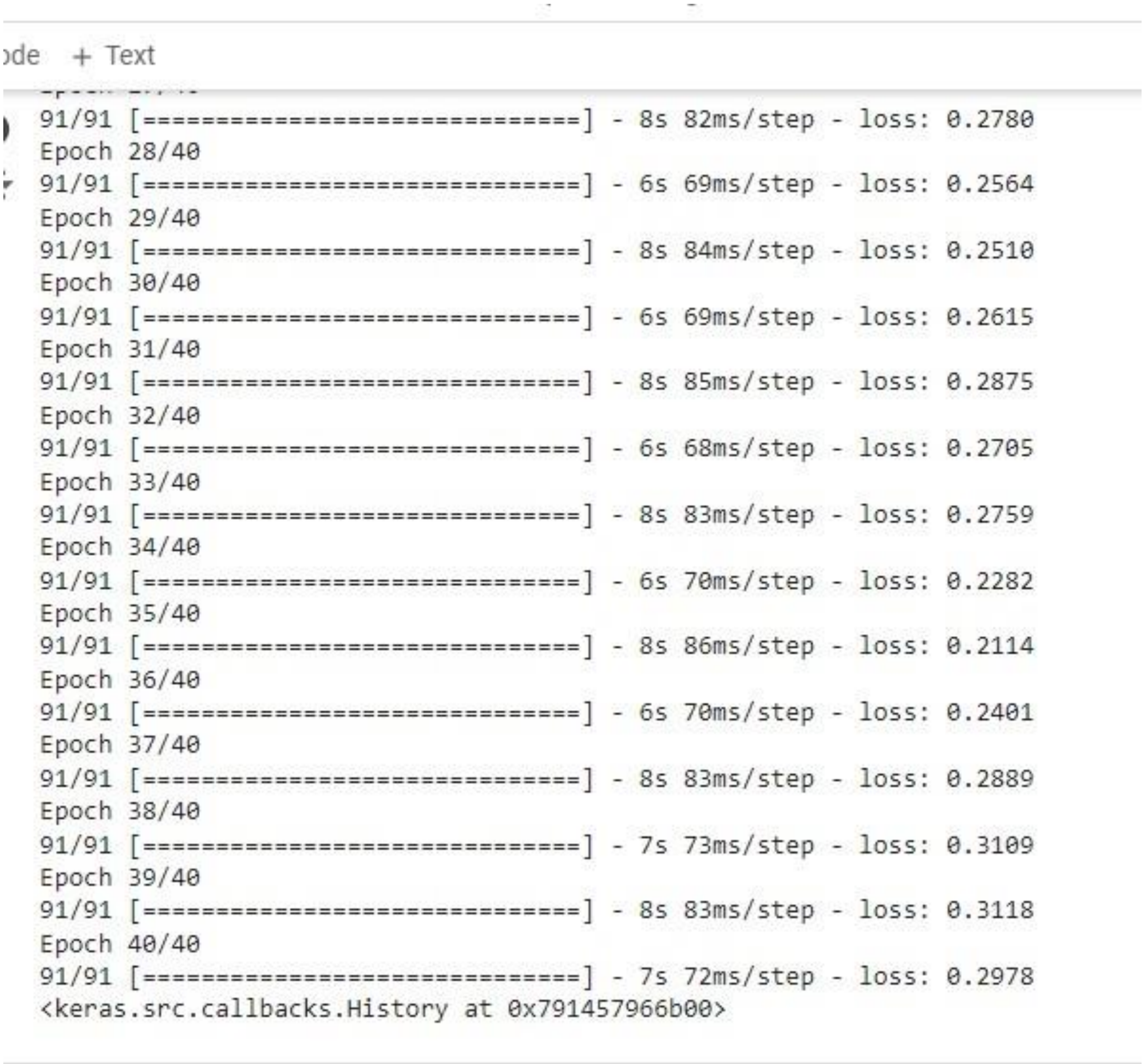
```
ns of gue with ameiiseshens loveddacn there creakice nd kasco aachning e fentiapaenglint she spilts in uhe world with the individual games accounting for many of the most watched television programs in american history an
```

The result is very good the model can print (‘she spilts in uhe world with the individual games accounting for many of the most watched television programs in american history’)it seems model predict is very good according to given data .

Model2(simple RNN):

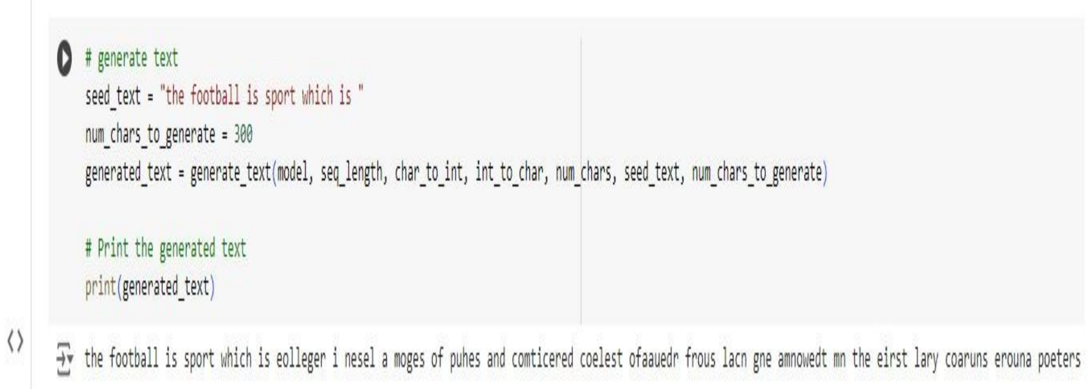
The model architecture starts with 3simple RNN layers, one containing 256 And the another contains 128 neurons and last one contains 64neurons . The purpose of these layers is to capture the sequential patterns and dependencies present in the input sequence of characters. Finally, the model ends with a softmax layer with a number of units equal to the total number of unique characters in the dataset.

Accuracy: the model loss is decreasing



```
91/91 [==============================] - 8s 82ms/step - loss: 0.2780
Epoch 28/40
91/91 [==============================] - 6s 69ms/step - loss: 0.2564
Epoch 29/40
91/91 [==============================] - 8s 84ms/step - loss: 0.2510
Epoch 30/40
91/91 [==============================] - 6s 69ms/step - loss: 0.2615
Epoch 31/40
91/91 [==============================] - 8s 85ms/step - loss: 0.2875
Epoch 32/40
91/91 [==============================] - 6s 68ms/step - loss: 0.2705
Epoch 33/40
91/91 [==============================] - 8s 83ms/step - loss: 0.2759
Epoch 34/40
91/91 [==============================] - 6s 70ms/step - loss: 0.2282
Epoch 35/40
91/91 [==============================] - 8s 86ms/step - loss: 0.2114
Epoch 36/40
91/91 [==============================] - 6s 70ms/step - loss: 0.2401
Epoch 37/40
91/91 [==============================] - 8s 83ms/step - loss: 0.2889
Epoch 38/40
91/91 [==============================] - 7s 73ms/step - loss: 0.3109
Epoch 39/40
91/91 [==============================] - 8s 83ms/step - loss: 0.3118
Epoch 40/40
91/91 [==============================] - 7s 72ms/step - loss: 0.2978
<keras.src.callbacks.History at 0x791457966b00>
```

Result



```
# generate text
seed_text = "the football is sport which is "
num_chars_to_generate = 300
generated_text = generate_text(model, seq_length, char_to_int, int_to_char, num_chars, seed_text, num_chars_to_generate)

# Print the generated text
print(generated_text)
```

```
the football is sport which is eolleger i nesel a moges of puhes and comticered coelest ofaauedr frous lacn gne amnowedt mn the eirst lary coaruns erouna poeters
```

Conclusion:

Character based prediction model need more data to achieve good performance in generating accurate predictions for the next character in a given sequence and also powerful architecture for train the LSTM architecture leads to good results more than simple RNN.