Neuroinformatics laboratory 3

Name: Radek Bartyzal
Email: rbartyzal1@gmail.com
Code: https://github.com/BartyzalRadek/neuroinformatics-course/blob/master/LSTM.ipynb

LSTM text generation

Dataset = list of names given in US census: http://deron.meranda.us/data/census-derived-all-first.txt

Data preprocessing

Original data:

JAMES

JOHN

ROBERT

MICHAEL

MARY

WILLIAM …

Total: 36122 characters

1. Replace new lines with spaces
2. Turn characters into numbers (27 unique characters)
3. Vectorize characters = turn them into one-hot encoding
4. Split data into sequences of length 20 overlapping each 3 characters = our training data

   The target data for sequence X is the character following sequence X.

Model:

1 layer of 50 LSTM neurons

1 layer of 27 (vocabulary size) softmax neurons

Training:

Training for 20 epochs with batch size 128.

Optimizer = RMSProp with learning rate 0.01

Generating 50 characters after each epoch to see how is the training going. Using seed sequence starting at random point in the data for each epoch.

Results:

```
Epoch 0
12034/12034 [==============================] - 9s - loss: 2.7628
----- Generating with seed: "A DAYSI DARLENA DARC"
H  H  H  H  H  H  H  H  H  H  H  H  H  H  H  H
------------------------------
Epoch 1
12034/12034 [==============================] - 8s - loss: 2.5401
----- Generating with seed: " MAJORIE MAGDA MAC L"
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL
------------------------------
Epoch 2
12034/12034 [==============================] - 8s - loss: 2.4221
----- Generating with seed: " DEBI DARRICK DARLEE"
  II  II  II  II  II  II  II  II  II  II  II  II
------------------------------
Epoch 3
12034/12034 [==============================] - 8s - loss: 2.3185
----- Generating with seed: "H KIMBERLY DEBORAH J"
ARINIA CICININIA CICININIA CICININIA CICININIA CIC
------------------------------
Epoch 4
12034/12034 [==============================] - 9s - loss: 2.2598
----- Generating with seed: "EA DEADRA DAYSI DARL"
ENN JANNE JANNE JANNE JANNE JANNE JANNE JANNE JANN
------------------------------
Epoch 5
12034/12034 [==============================] - 9s - loss: 2.2209
----- Generating with seed: "UBIA NU NORIKO NOHEM"
INNENNENNENNENNENNENNENNENNENNENNENNENNENNEN
------------------------------
Epoch 6
12034/12034 [==============================] - 9s - loss: 2.1863
----- Generating with seed: "A KANDRA KANDIS KAMI"
A ARA ARA ARA ARA ARA ARA ARA ARA ARA ARA ARA ARA
------------------------------
Epoch 7
12034/12034 [==============================] - 9s - loss: 2.1476
----- Generating with seed: "ALA KALYN KALLIE KAL"
LA LANALLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL
------------------------------
Epoch 8
12034/12034 [==============================] - 8s - loss: 2.1007
----- Generating with seed: "E ELANOR EDDA ECHO E"
MANA MARIANA MARIANA MARIANA MARIANA MARIANA MARIA
------------------------------
Epoch 9
12034/12034 [==============================] - 9s - loss: 2.0619
----- Generating with seed: "SA LUCILE LORIE LEAN"
 RISTA RISTA RISTA RISTA RISTA RISTA RISTA RISTA R
------------------------------
Epoch 10
12034/12034 [==============================] - 8s - loss: 2.0248
----- Generating with seed: "ETTA ESTELLA ELVA EF"
RIN RORI RORI RORI RORI RORI RORI RORI RORI RORI R
------------------------------
Epoch 11
12034/12034 [==============================] - 8s - loss: 1.9868
```

```
----- Generating with seed: "IE BARBERA BARBAR BA"
LLE ALLEN GRILL CARLE CLELL GELLE ALLEN GRILL GELL
------------------------------
Epoch 12
12034/12034 [==============================] - 8s - loss: 1.9465
----- Generating with seed: " MCKENZIE MAYE MAYBE"
L MARLI MARLI MARLI MARLIN MARLIN MARLIN MARLIN MA
------------------------------
Epoch 13
12034/12034 [==============================] - 9s - loss: 1.9052
----- Generating with seed: "CKA ELNORA ELLIOTT E"
LENE ELENE ELENE ELENE ELENE ELENE ELENE ELENE ELE
------------------------------
Epoch 14
12034/12034 [==============================] - 9s - loss: 1.8631
----- Generating with seed: "ARIANO MARGOT MA LOU"
DY MARISTY MARISHA MARISTE MARISTY MARISHA MARISTE
------------------------------
Epoch 15
12034/12034 [==============================] - 9s - loss: 1.8075
----- Generating with seed: "COLAS MARISSA LOURDE"
 MARABERTO MARABERTO MARABERTO MARABERTO MARABERTO
------------------------------
Epoch 16
12034/12034 [==============================] - 9s - loss: 1.7676
----- Generating with seed: " MALIA MAIRA MAEGAN "
MICIA MICIA MICIA MICIA MICIA MICIA MICIA MICIA MI
------------------------------
Epoch 17
12034/12034 [==============================] - 10s - loss: 1.7324
----- Generating with seed: " ANTWAN ANNETTA ANNE"
TTA ARINA ARISA ARINA ARISA ARINA ARISA ARINA ARIS
------------------------------
Epoch 18
12034/12034 [==============================] - 9s - loss: 1.6961
----- Generating with seed: "THA SHAWNA RENA ORA "
ROSENE ROSANE RONETTE RESA RESA RENETA RESA RENETA
------------------------------
Epoch 19
12034/12034 [==============================] - 9s - loss: 1.6668
----- Generating with seed: "EARLE PAULETTA PATRI"
E ROBELBRE GRADA ROBELBERTORDE GENNIE BENNIE BENNA
```

Conclusion:

The network first just generates random letters, repeating itself in a loop.

Around epoch 10 it starts to generate names copied from the dataset but still falls into repeating cycles.

By the end it successfully generates unique names without repeating itself in the generated sample, but the repetitions are still there just at a larger scale.

With further training (60 epochs+) the repetition patterns stopped appearing even if I generated 1000 characters.