Report

Tao Ma, ZeHao Chen

Task I:

LSTM network: The first layer is embedding(total_words, 10, input_length = max_sequence_len-1). The embedding layer's role is to train the matrix W and embed the words and assign each word its corresponding word vector. The total_words means the size of the vocabulary. 10 means the dimension of the dense embedding. The input_length is the length of input sequence, which is 100 in here.

The second layer is LSTM(150, return_sequence=True). Long Short-Term Memory layer. 150 means the output dimension. Return_sequence means returning the entire sequence it true, otherwise, only the last output of the output sequence is returned.

The third layer is LSTM(100). The output dimension is 100 and the return sequence is last output of the output.

The forth layer is Dense(total_words, activation='softmax'). The operation implemented is output = activation(dot(input, kernel)+bias). Where activation is the activation function calculated by element, kernel is the weight matrix of this layer, and bias is the bias vector, which is added only when use_bias=True. The output dimension is the size of total_words. And use the softmax to classify.

Next is compile and fit. Use categorical_crossentropy as loss function that can deal with the multi-class question, and use adam as optimizer, and set the early stopping to avoid overfitting.

LSTM can avoid the gradient vanish of regular RNN using state cell, forget gate, which determine what information we are going to throw out of the cellular state, input gate, which determine which new information we will store in the cellular state, and output gate that determine the final output based on the current cellular state. And the GRU only has two gates. I think the LSTM is better.

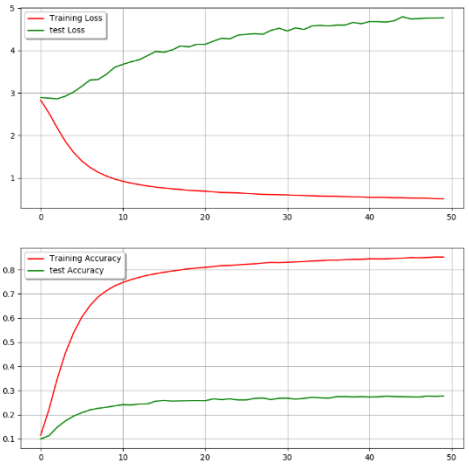
Task II:

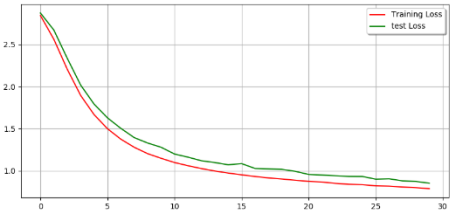
I made the Weight Initialization: keras.initializers.he_uniform(seed=None).

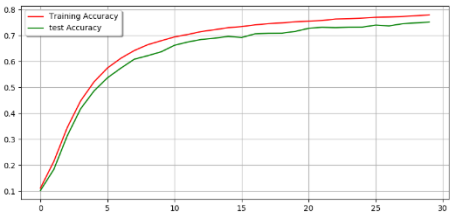
It extracts samples from the uniform distribution in [-limit, limit], where limit is SQRT (6 / fan_in), where fan_in is the number of input units in the weight tensor.

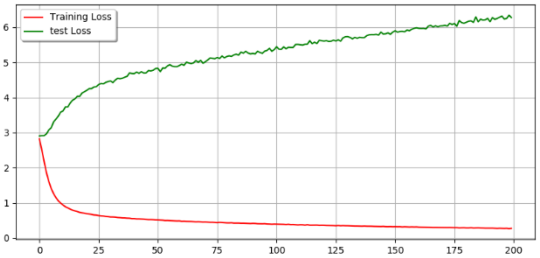
The first Hyperparameter i adjusted was epochs. The first time, I choose epochs as 20, but it is not enough because the accuracy is still growth. The second time, I choose epochs as 40, it is keep the accuracy as 80%. I hope it can be higher accuracy. So, the last time, I choose epochs as 60. It is enough.

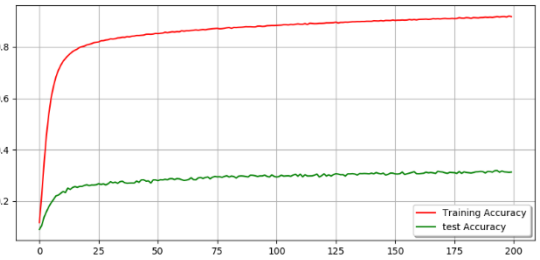
The second hyperparameter I adjusted was earlystopping. At the beginning, because I choose some samples to train, I choose different samples to test. I got the test accuracy that is very low. So, it always was halted due to low accuracy. After, I choose the first 1000 samples to train, and pick some to test. It works.



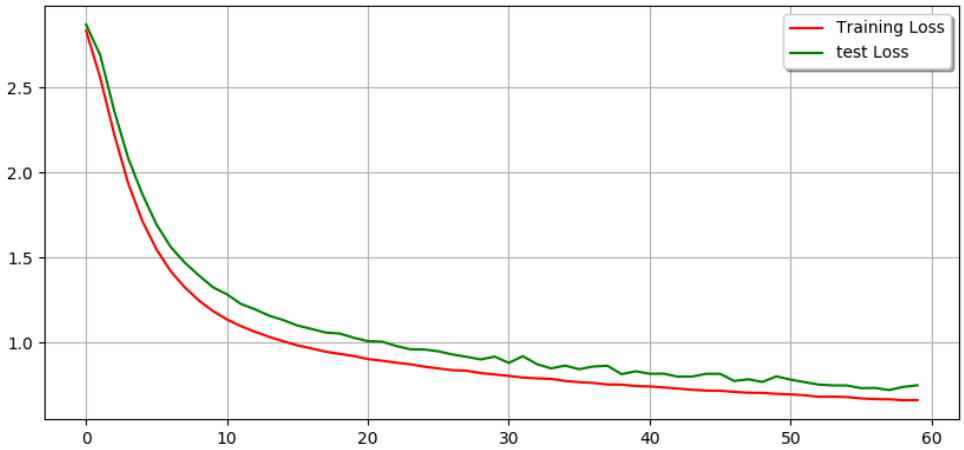


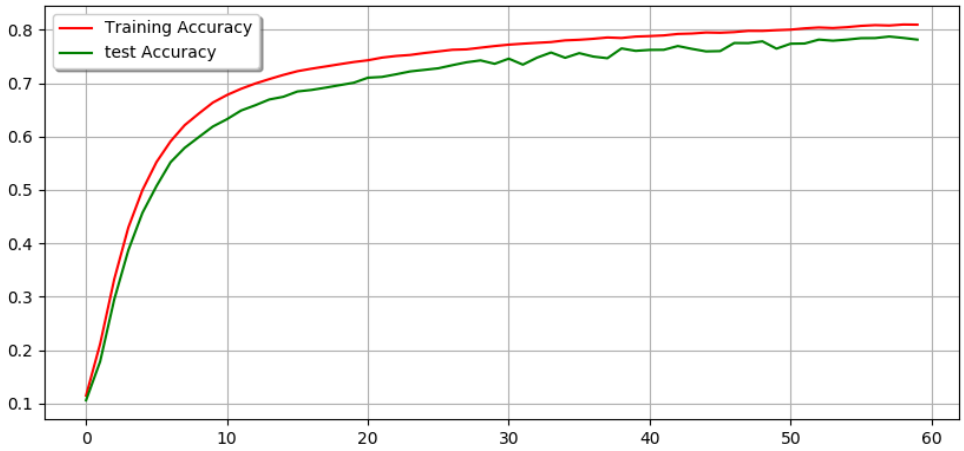






These are some wrong results.





Because my computer cannot deal with all samples, I use the 1000 samples in pdb_seqres.txt. And I set the max_len is 25. 800 samples are used to train and 200 samples are used to test. So, I get the result. The train accuracy is more than 80% and the test accuracy is close to 80%.

I changed the letters one at a time, and made predictions, and then compared them to see if they made a difference. For example, I use “m n i f e m l r i d e g l r l k i y k d t e g y y” as sequence, the first time, I changed the first letter “m” to see if they made a difference by predicting. And then I kept the first no change, and changed the second letter to see if they made a difference. Until I changed the last one. I test the 500 samples and the length of every sample is 25. I counted the times that we can get the correct result.

```
[[24], [], [], [24], [24], [24], [], [24], [], [], [], [], [24], [24], [], [], [], [24], [], [24], [23], [24], [24], [], [],
Counter({24: 187, 23: 54, 22: 31, 21: 19, 19: 10, 20: 10, 16: 6, 15: 4, 18: 4, 14: 2, 7: 2, 3: 2, 10: 2, 17: 2, 12: 1})
```

This is the result, when i can predict correctly, changing the 24th letters get more chances. But sometimes, only changing the 16th, 15th, 18th can get the correct prediction.

Task III:

I generated these five samples:

MQNGYTYEDYQDTAKWLLSHTEQRP

RSRLTADEYLKIYQAAESSPCWLRL

MQNGYTYEDYQDTAKWLLSHTEQRP

MVLSEGEWQLVLHVWAKVEADVAGH

STAGKVIKCKAAVLWEEKKPFSIEE

Sample_800th

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19+
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

xla MQNGYTYEDYQDTAKWLLSHTEQRP

fix k sequence ['M']

xpre ['M', 'N', 'I', 'F', 'E', 'M', 'L', 'R', 'I', 'D', 'E', 'G', 'L', 'R', 'L', 'K', 'I', 'Y', 'K', 'D', 'T', 'E', 'G', 'Y', 'Y']

fix k: 1

fix k sequence ['M', 'Q']

xpre ['M', 'Q', 'T', 'I', 'K', 'C', 'V', 'V', 'V', 'G', 'G', 'D', 'D', 'D', 'V', 'H', 'F', 'F', 'I', 'L', 'T', 'N', 'N', 'F', 'P']

fix k: 2

fix k sequence ['M', 'Q', 'N']

xpre ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y', 'E', 'D', 'D', 'R', 'S', 'T', 'I', 'G', 'A', 'S', 'L', 'L', 'N', 'E', 'K', 'F', 'T', 'Q']

fix k: 3

fix k sequence ['M', 'Q', 'N', 'G']

xpre ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y', 'E', 'D', 'D', 'R', 'S', 'T', 'I', 'G', 'A', 'S', 'L', 'L', 'N', 'E', 'K', 'F', 'T', 'Q']

fix k: 4

fix k sequence ['M', 'Q', 'N', 'G', 'Y']

xpre ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y', 'E', 'D', 'D', 'R', 'S', 'T', 'I', 'G', 'A', 'S', 'L', 'L', 'N', 'E', 'K', 'F', 'T', 'Q']

fix k: 5

fix k sequence ['M', 'Q', 'N', 'G', 'Y', 'T']

xpre ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y', 'E', 'D', 'D', 'R', 'S', 'T', 'I', 'G', 'A', 'S', 'L', 'L', 'N', 'E', 'K', 'F', 'T', 'Q']

fix k: 6

fix k sequence ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y']

xpre ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y', 'E', 'D', 'D', 'R', 'S', 'T', 'I', 'G', 'A', 'S', 'L', 'L', 'N', 'E', 'K', 'F', 'T', 'Q']

fix k: 7

fix k sequence ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y', 'E']

xpre ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y', 'E', 'D', 'D', 'R', 'S', 'T', 'I', 'G', 'A', 'S', 'L', 'L', 'N', 'E', 'K', 'F', 'T', 'Q']

fix k: 8

fix k sequence ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y', 'E', 'D']

xpre ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y', 'E', 'D', 'D', 'R', 'S', 'T', 'I', 'G', 'A', 'S', 'L', 'L', 'N', 'E', 'K', 'F', 'T', 'Q']

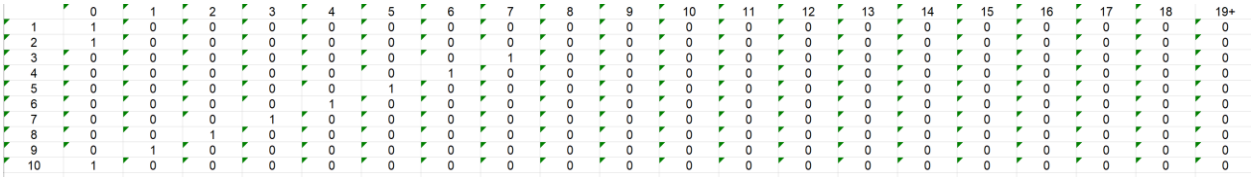
fix k: 9

fix k sequence ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y', 'E', 'D', 'Y']

xpre ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y', 'E', 'D', 'Y', 'Q', 'D', 'T', 'A', 'K', 'W', 'L', 'I', 'E', 'K', 'K', 'W', 'D', 'G', 'V']

fix k: 10

Sample_999th



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19+
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

RSRLTADEYLKIYQAAESSPCWLRL

xla RSRLTADEYLKIYQAAESSPCWLRL

fix k sequence ['R']

xpre ['R', 'S', 'A', 'A', 'T', 'H', 'A', 'A', 'A', 'E', 'W', 'I', 'Q', 'Q', 'K', 'K', 'I', 'I', 'C', 'I', 'E', 'D', 'D', 'S', 'R']

fix k: 1

fix k sequence ['R', 'S']

xpre ['R', 'S', 'A', 'A', 'T', 'H', 'A', 'A', 'A', 'E', 'W', 'I', 'Q', 'Q', 'K', 'K', 'I', 'I', 'C', 'I', 'E', 'D', 'D', 'S', 'R']

fix k: 2

fix k sequence ['R', 'S', 'R']

xpre ['R', 'S', 'R', 'L', 'T', 'A', 'D', 'E', 'Y', 'L', 'K', 'E', 'K', 'I', 'T', 'E', 'E', 'K', 'I', 'K', 'E', 'F', 'T', 'E', 'F']

fix k: 3

fix k sequence ['R', 'S', 'R', 'L']

xpre ['R', 'S', 'R', 'L', 'T', 'A', 'D', 'E', 'Y', 'L', 'K', 'E', 'K', 'I', 'T', 'E', 'E', 'K', 'I', 'K', 'E', 'F', 'T', 'E', 'F']

fix k: 4

fix k sequence ['R', 'S', 'R', 'L', 'T']

xpre ['R', 'S', 'R', 'L', 'T', 'A', 'D', 'E', 'Y', 'L', 'K', 'E', 'K', 'I', 'T', 'E', 'E', 'K', 'I', 'K', 'E', 'F', 'T', 'E', 'F']

fix k: 5

fix k sequence ['R', 'S', 'R', 'L', 'T', 'A']

xpre ['R', 'S', 'R', 'L', 'T', 'A', 'D', 'E', 'Y', 'L', 'K', 'E', 'K', 'I', 'T', 'E', 'E', 'K', 'I', 'K', 'E', 'F', 'T', 'E', 'F']

fix k: 6

fix k sequence ['R', 'S', 'R', 'L', 'T', 'A', 'D']

xpre ['R', 'S', 'R', 'L', 'T', 'A', 'D', 'E', 'Y', 'L', 'K', 'E', 'K', 'I', 'T', 'E', 'E', 'K', 'I', 'K', 'E', 'F', 'T', 'E', 'F']

fix k: 7

fix k sequence ['R', 'S', 'R', 'L', 'T', 'A', 'D', 'E']

xpre ['R', 'S', 'R', 'L', 'T', 'A', 'D', 'E', 'Y', 'L', 'K', 'E', 'K', 'I', 'T', 'E', 'E', 'K', 'I', 'K', 'E', 'F', 'T', 'E', 'F']

fix k: 8

fix k sequence ['R', 'S', 'R', 'L', 'T', 'A', 'D', 'E', 'Y']

xpre ['R', 'S', 'R', 'L', 'T', 'A', 'D', 'E', 'Y', 'L', 'K', 'E', 'K', 'I', 'T', 'E', 'E', 'K', 'I', 'K', 'E', 'F', 'T', 'E', 'F']

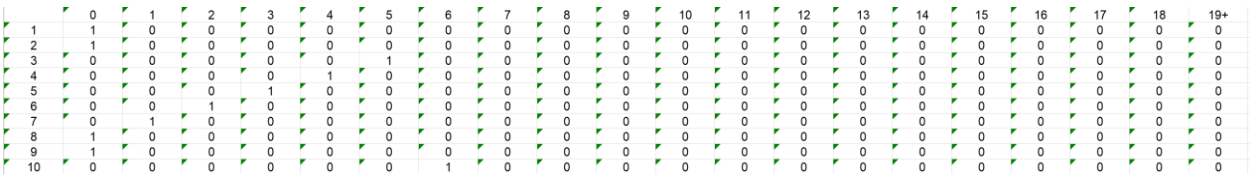
fix k: 9

fix k sequence ['R', 'S', 'R', 'L', 'T', 'A', 'D', 'E', 'Y', 'L']

xpre ['R', 'S', 'R', 'L', 'T', 'A', 'D', 'E', 'Y', 'L', 'K', 'E', 'K', 'I', 'T', 'E', 'E', 'K', 'I', 'K', 'E', 'F', 'T', 'E', 'F']

fix k: 10

Sample_801th



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19+
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

xla MQNGYTYEDYQDTAKWLLSHTEQRP

fix k sequence ['M']

xpre ['M', 'N', 'I', 'F', 'E', 'M', 'L', 'R', 'I', 'D', 'E', 'G', 'L', 'R', 'L', 'K', 'I', 'Y', 'K', 'D', 'T', 'E', 'G', 'Y', 'Y']

fix k: 1

fix k sequence ['M', 'Q']

xpre ['M', 'Q', 'T', 'I', 'K', 'C', 'V', 'V', 'V', 'G', 'G', 'D', 'D', 'D', 'V', 'H', 'F', 'F', 'I', 'L', 'T', 'N', 'N', 'F', 'P']

fix k: 2

fix k sequence ['M', 'Q', 'N']

xpre ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y', 'E', 'D', 'D', 'R', 'S', 'T', 'I', 'G', 'A', 'S', 'L', 'L', 'N', 'E', 'K', 'F', 'T', 'Q']

fix k: 3

fix k sequence ['M', 'Q', 'N', 'G']

xpre ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y', 'E', 'D', 'D', 'R', 'S', 'T', 'I', 'G', 'A', 'S', 'L', 'L', 'N', 'E', 'K', 'F', 'T', 'Q']

fix k: 4

fix k sequence ['M', 'Q', 'N', 'G', 'Y']

xpre ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y', 'E', 'D', 'D', 'R', 'S', 'T', 'I', 'G', 'A', 'S', 'L', 'L', 'N', 'E', 'K', 'F', 'T', 'Q']

fix k: 5

fix k sequence ['M', 'Q', 'N', 'G', 'Y', 'T']

xpre ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y', 'E', 'D', 'D', 'R', 'S', 'T', 'I', 'G', 'A', 'S', 'L', 'L', 'N', 'E', 'K', 'F', 'T', 'Q']

fix k: 6

fix k sequence ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y']

xpre ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y', 'E', 'D', 'D', 'R', 'S', 'T', 'I', 'G', 'A', 'S', 'L', 'L', 'N', 'E', 'K', 'F', 'T', 'Q']

fix k: 7

fix k sequence ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y', 'E']

xpre ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y', 'E', 'D', 'D', 'R', 'S', 'T', 'I', 'G', 'A', 'S', 'L', 'L', 'N', 'E', 'K', 'F', 'T', 'Q']

fix k: 8

fix k sequence ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y', 'E', 'D']

xpre ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y', 'E', 'D', 'D', 'R', 'S', 'T', 'I', 'G', 'A', 'S', 'L', 'L', 'N', 'E', 'K', 'F', 'T', 'Q']

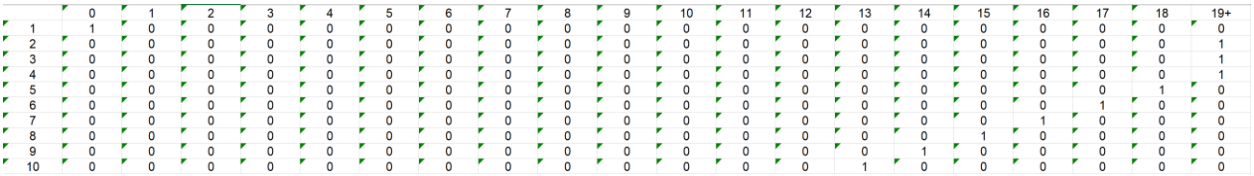
fix k: 9

fix k sequence ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y', 'E', 'D', 'Y']

xpre ['M', 'Q', 'N', 'G', 'Y', 'T', 'Y', 'E', 'D', 'Y', 'Q', 'D', 'T', 'A', 'K', 'W', 'L', 'I', 'E', 'K', 'K', 'W', 'D', 'G', 'V']

fix k: 10

Sample_1th



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19+
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0

xla MVLSEGEWQLVLHVWAKVEADVAGH

fix k sequence ['M']

xpre ['M', 'N', 'I', 'F', 'E', 'M', 'L', 'R', 'I', 'D', 'E', 'G', 'L', 'R', 'L', 'K', 'I', 'Y', 'K', 'D', 'T', 'E', 'G', 'Y', 'Y']

fix k: 1

fix k sequence ['M', 'V']

xpre ['M', 'V', 'L', 'S', 'E', 'G', 'E', 'W', 'Q', 'L', 'V', 'L', 'H', 'V', 'W', 'A', 'K', 'V', 'E', 'A', 'D', 'V', 'A', 'G', 'H']

fix k: 2

fix k sequence ['M', 'V', 'L']

xpre ['M', 'V', 'L', 'S', 'E', 'G', 'E', 'W', 'Q', 'L', 'V', 'L', 'H', 'V', 'W', 'A', 'K', 'V', 'E', 'A', 'D', 'V', 'A', 'G', 'H']

fix k: 3

fix k sequence ['M', 'V', 'L', 'S']

xpre ['M', 'V', 'L', 'S', 'E', 'G', 'E', 'W', 'Q', 'L', 'V', 'L', 'H', 'V', 'W', 'A', 'K', 'V', 'E', 'A', 'D', 'V', 'A', 'G', 'H']

fix k: 4

fix k sequence ['M', 'V', 'L', 'S', 'E']

xpre ['M', 'V', 'L', 'S', 'E', 'G', 'E', 'W', 'Q', 'L', 'V', 'L', 'H', 'V', 'W', 'A', 'K', 'V', 'E', 'A', 'D', 'V', 'A', 'G', 'H']

fix k: 5

fix k sequence ['M', 'V', 'L', 'S', 'E', 'G']

xpre ['M', 'V', 'L', 'S', 'E', 'G', 'E', 'W', 'Q', 'L', 'V', 'L', 'H', 'V', 'W', 'A', 'K', 'V', 'E', 'A', 'D', 'V', 'A', 'G', 'H']

fix k: 6

fix k sequence ['M', 'V', 'L', 'S', 'E', 'G', 'E']

xpre ['M', 'V', 'L', 'S', 'E', 'G', 'E', 'W', 'Q', 'L', 'V', 'L', 'H', 'V', 'W', 'A', 'K', 'V', 'E', 'A', 'D', 'V', 'A', 'G', 'H']

fix k: 7

fix k sequence ['M', 'V', 'L', 'S', 'E', 'G', 'E', 'W']

xpre ['M', 'V', 'L', 'S', 'E', 'G', 'E', 'W', 'Q', 'L', 'V', 'L', 'H', 'V', 'W', 'A', 'K', 'V', 'E', 'A', 'D', 'V', 'A', 'G', 'H']

fix k: 8

fix k sequence ['M', 'V', 'L', 'S', 'E', 'G', 'E', 'W', 'Q']

xpre ['M', 'V', 'L', 'S', 'E', 'G', 'E', 'W', 'Q', 'L', 'V', 'L', 'H', 'V', 'W', 'A', 'K', 'V', 'E', 'A', 'D', 'V', 'A', 'G', 'H']

fix k: 9

fix k sequence ['M', 'V', 'L', 'S', 'E', 'G', 'E', 'W', 'Q', 'L']

xpre ['M', 'V', 'L', 'S', 'E', 'G', 'E', 'W', 'Q', 'L', 'V', 'L', 'H', 'V', 'W', 'A', 'K', 'V', 'E', 'A', 'D', 'V', 'A', 'G', 'H']

fix k: 10

sample 950th

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19+
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0

xla STAGKVIKCKAAVLWEEKKPFSIEE

fix k sequence ['S']

xpre ['S', 'T', 'A', 'G', 'K', 'V', 'I', 'K', 'C', 'K', 'A', 'A', 'V', 'L', 'W', 'E', 'E', 'K', 'K', 'P', 'F', 'S', 'I', 'E', 'E']

fix k: 1

fix k sequence ['S', 'T']

xpre ['S', 'T', 'A', 'G', 'K', 'V', 'I', 'K', 'C', 'K', 'A', 'A', 'V', 'L', 'W', 'E', 'E', 'K', 'K', 'P', 'F', 'S', 'I', 'E', 'E']

fix k: 2

fix k sequence ['S', 'T', 'A']

xpre ['S', 'T', 'A', 'G', 'K', 'V', 'I', 'K', 'C', 'K', 'A', 'A', 'V', 'L', 'W', 'E', 'E', 'K', 'K', 'P', 'F', 'S', 'I', 'E', 'E']

fix k: 3

fix k sequence ['S', 'T', 'A', 'G']

xpre ['S', 'T', 'A', 'G', 'K', 'V', 'I', 'K', 'C', 'K', 'A', 'A', 'V', 'L', 'W', 'E', 'E', 'K', 'K', 'P', 'F', 'S', 'I', 'E', 'E']

fix k: 4

fix k sequence ['S', 'T', 'A', 'G', 'K']

xpre ['S', 'T', 'A', 'G', 'K', 'V', 'I', 'K', 'C', 'K', 'A', 'A', 'V', 'L', 'W', 'E', 'E', 'K', 'K', 'P', 'F', 'S', 'I', 'E', 'E']

fix k: 5

fix k sequence ['S', 'T', 'A', 'G', 'K', 'V']

xpre ['S', 'T', 'A', 'G', 'K', 'V', 'I', 'K', 'C', 'K', 'A', 'A', 'V', 'L', 'W', 'E', 'E', 'K', 'K', 'P', 'F', 'S', 'I', 'E', 'E']

fix k: 6

fix k sequence ['S', 'T', 'A', 'G', 'K', 'V', 'I']

xpre ['S', 'T', 'A', 'G', 'K', 'V', 'I', 'K', 'C', 'K', 'A', 'A', 'V', 'L', 'W', 'E', 'E', 'K', 'K', 'P', 'F', 'S', 'I', 'E', 'E']

fix k: 7

fix k sequence ['S', 'T', 'A', 'G', 'K', 'V', 'I', 'K']

xpre ['S', 'T', 'A', 'G', 'K', 'V', 'I', 'K', 'C', 'K', 'A', 'A', 'V', 'L', 'W', 'E', 'E', 'K', 'K', 'P', 'F', 'S', 'I', 'E', 'E']

fix k: 8

fix k sequence ['S', 'T', 'A', 'G', 'K', 'V', 'I', 'K', 'C']

xpre ['S', 'T', 'A', 'G', 'K', 'V', 'I', 'K', 'C', 'K', 'A', 'A', 'V', 'L', 'W', 'E', 'E', 'K', 'K', 'P', 'F', 'S', 'I', 'E', 'E']

fix k: 9

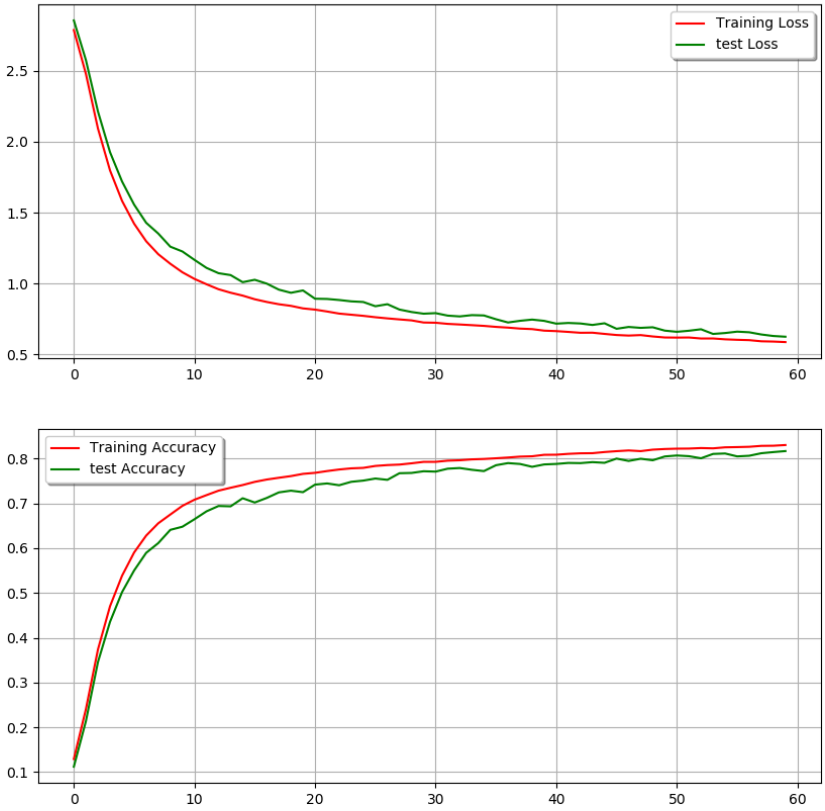
fix k sequence ['S', 'T', 'A', 'G', 'K', 'V', 'I', 'K', 'C', 'K']

xpre ['S', 'T', 'A', 'G', 'K', 'V', 'I', 'K', 'C', 'K', 'A', 'A', 'V', 'L', 'W', 'E', 'E', 'K', 'K', 'P', 'F', 'S', 'I', 'E', 'E']

fix k: 10

I only compute the several samples and get the table. If you want to get more table, you can run my get_table.py. it will produce the table automatically.

(2) I trained a new model (4 gram model).



This is a little better than last model. The accuracy of training is about 85% and the accuracy of test is about 0.82.

And then I compute the probability of MQNG in MQNGYTYEDYQDTAKWLLSHTEQRP.

In the first model, P(M)= 0.014747879, P(Q|M)= 0.008295132, P(N|QM)= 0.035728157, P(G|MQN)=0.0030834817. So, P(MQNG)= 1.3477359e-08.

In the 4-gram model, P(M)= 0.00069047284, P(Q|M)= 0.0008694072, P(N|QM)= 1.9276287e-05, P(G|MQN)= 2.190843e-05. So, P(MQNG)= 2.5351548e-16.

And then I compute the probability of QNGY in MQNGYTYEDYQDTAKWLLSHTEQRP.

In the first model, P(Q)= 0.018408585, P(N|Q)= 0.0016411562, P(G|QN)= 0.003831849, P(Y|QNG)= 0.008317778, P(MQNG)= 9.629106e-10.

In the 4-gram model, P(Q)= 0.0006098945, P(N|Q)= 0.0017425038, P(G|QN 0.0022819682, P(Y|QNG)= 0.0002524591, P(MQNG)= 6.1225043e-13.

And then I compute the probability of NGYT in MQNGYTYEDYQDTAKWLLSHTEQRP.

In the first model, P(N)=0.0810478, P(G|N)= 0.018971322, P(Y|GN)= 0.0032301378, P(T|NGY)= 0.008137504, P(NGYT)= 4.041579e-08.

In the 4-gram model, P(N)= 0.0008805519, P(G|N)= 0.00966275, P(Y|GN)= 0.00047660992, P(T|NGY)= 0.016990261, P(NGYT)= 6.8899934e-11.

Like this order. This is the all result.

The last model:

number: 0

P(first) 0.014747879

P(second|first) 0.008295132

P(third|first two) 0.035728157

P(forth|first three) 0.0030834817

P(all four) 1.3477359e-08

number: 1

P(first) 0.018408585

P(second|first) 0.0016411562

P(third|first two) 0.003831849

P(forth|first three) 0.008317778

P(all four) 9.629106e-10

number: 2

P(first) 0.0810478

P(second|first) 0.018971322

P(third|first two) 0.0032301378

P(forth|first three) 0.008137504

P(all four) 4.041579e-08

number: 3

P(first) 0.017519737

P(second|first) 0.073552534

P(third|first two) 0.088673174

P(forth|first three) 0.000127867

P(all four) 1.46108645e-08

number: 4

P(first) 0.008243395

P(second|first) 0.11343215

P(third|first two) 6.111665e-05

P(forth|first three) 0.00017919696

P(all four) 1.0240765e-11

number: 5

P(first) 0.025790889

P(second|first) 0.009197957

P(third|first two) 0.16028535

P(forth|first three) 0.001318955

P(all four) 5.0151222e-08

number: 6

P(first) 0.008243395

P(second|first) 0.103414655

P(third|first two) 0.0049312273

P(forth|first three) 0.0046853516

P(all four) 1.9696335e-08

number: 7

P(first) 0.07844744

P(second|first) 0.046892595

P(third|first two) 0.022979332

P(forth|first three) 0.010494114

P(all four) 8.87087e-07

number: 8

P(first) 0.074565716

P(second|first) 0.013630279

P(third|first two) 0.20924231

P(forth|first three) 3.5903064e-05

P(all four) 7.63528e-09

number: 9

P(first) 0.008243395

P(second|first) 0.11532303

P(third|first two) 0.003976098

P(forth|first three) 2.483703e-05

P(all four) 9.388126e-11

number: 10

P(first) 0.018408585

P(second|first) 0.04642617

P(third|first two) 0.023205513

P(forth|first three) 0.011307549

P(all four) 2.2425539e-07

number: 11

P(first) 0.074565716

P(second|first) 0.00015660012

P(third|first two) 0.0008514389

P(forth|first three) 0.022794148

P(all four) 2.2662515e-10

number: 12

P(first) 0.025790889

P(second|first) 0.007583688

P(third|first two) 0.0024863482

P(forth|first three) 0.025147868

P(all four) 1.2229533e-08

number: 13

P(first) 0.03318773

P(second|first) 0.024727233

P(third|first two) 0.0026278591

P(forth|first three) 2.5823996e-11

P(all four) 5.569017e-17

number: 14

P(first) 0.008573342

P(second|first) 0.0065386724

P(third|first two) 2.021876e-11

P(forth|first three) 2.215972e-10

P(all four) 2.5116466e-25

number: 15

P(first) 0.0034748893

P(second|first) 4.3097175e-11

P(third|first two) 3.24165e-12

P(forth|first three) 0.0010413206

P(all four) 5.0552243e-28

number: 16

P(first) 2.839347e-11

P(second|first) 3.6253143e-11

P(third|first two) 0.024407143

P(forth|first three) 2.4937628e-06

P(all four) 6.2652186e-29

number: 17

P(first) 2.839347e-11

P(second|first) 0.053428054

P(third|first two) 0.077874124

P(forth|first three) 0.02375308

P(all four) 2.806086e-15

number: 18

P(first) 0.43453106

P(second|first) 0.005971057

P(third|first two) 0.022562245

P(forth|first three) 0.32541588

P(all four) 1.9049918e-05

number: 19

P(first) 0.0038999673

P(second|first) 0.0678688

P(third|first two) 0.025589522

P(forth|first three) 0.0005148288

P(all four) 3.4870336e-09

number: 20

P(first) 0.025790889

P(second|first) 0.08799851

P(third|first two) 0.01339587

P(forth|first three) 0.0015695745

P(all four) 4.7719343e-08

number: 21

P(first) 0.07844744

P(second|first) 0.05410671

P(third|first two) 0.01073887

P(forth|first three) 0.00083687576

P(all four) 3.814604e-08

The 4-gram model.

number: 0

P(first) 0.00069047284

P(second|first) 0.0008694072

P(third|first two) 1.9276287e-05

P(forth|first three) 2.190843e-05

P(all four) 2.5351548e-16

number: 1

P(first) 0.0006098945

P(second|first) 0.0017425038

P(third|first two) 0.0022819682

P(forth|first three) 0.0002524591

P(all four) 6.1225043e-13

number: 2

P(first) 0.0008805519

P(second|first) 0.00966275

P(third|first two) 0.00047660992

P(forth|first three) 0.016990261

P(all four) 6.8899934e-11

number: 3

P(first) 8.6090195e-06

P(second|first) 0.019537915

P(third|first two) 0.38420615

P(forth|first three) 5.0279887e-05

P(all four) 3.2493051e-12

number: 4

P(first) 0.0025733314

P(second|first) 0.17210785

P(third|first two) 0.009881633

P(forth|first three) 5.681735e-05

P(all four) 2.486601e-10

number: 5

P(first) 0.006751065

P(second|first) 0.003597087

P(third|first two) 0.03151011

P(forth|first three) 0.012094953

P(all four) 9.25502e-09

number: 6

P(first) 0.0025733314

P(second|first) 0.00084944017

P(third|first two) 1.4324235e-05

P(forth|first three) 0.000118348464

P(all four) 3.705634e-15

number: 7

P(first) 0.00077700504

P(second|first) 0.00092082936

P(third|first two) 0.0013991538

P(forth|first three) 0.001377505

P(all four) 1.3789918e-12

number: 8

P(first) 0.025610765

P(second|first) 0.00049289

P(third|first two) 0.0010956441

P(forth|first three) 9.941298e-07

P(all four) 1.37494445e-14

number: 9

P(first) 0.0025733314

P(second|first) 0.5639717

P(third|first two) 0.00012661272

P(forth|first three) 0.0026616142

P(all four) 4.89075e-10

number: 10

P(first) 0.0006098945

P(second|first) 0.00061442447

P(third|first two) 0.33469748

P(forth|first three) 0.00900883

P(all four) 1.1299107e-09

number: 11

P(first) 0.025610765

P(second|first) 0.015028269

P(third|first two) 0.3095523

P(forth|first three) 0.4840569

P(all four) 5.7671594e-05

number: 12

P(first) 0.006751065

P(second|first) 0.02532412

P(third|first two) 0.35636517

P(forth|first three) 4.711925e-06

P(all four) 2.8707825e-10

number: 13

P(first) 0.017747395

P(second|first) 0.1316648

P(third|first two) 4.0570632e-07

P(forth|first three) 4.1508372e-14

P(all four) 3.9350638e-23

number: 14

P(first) 0.040142383

P(second|first) 8.610282e-06

P(third|first two) 1.5578927e-12

P(forth|first three) 1.6802632e-12

P(all four) 9.047642e-31

number: 15

P(first) 0.0004793588

P(second|first) 6.1003535e-13

P(third|first two) 1.1993953e-13

P(forth|first three) 0.35609987

P(all four) 1.24896384e-29

number: 16

P(first) 3.154449e-11

P(second|first) 3.2150103e-12

P(third|first two) 0.30677992

P(forth|first three) 5.607376e-07

P(all four) 1.7445863e-29

number: 17

P(first) 3.154449e-11

P(second|first) 0.0002289541

P(third|first two) 0.0004981382

P(forth|first three) 0.054332107

P(all four) 1.9546919e-19

number: 18

P(first) 0.076522686

P(second|first) 0.092554644

P(third|first two) 0.16332652

P(forth|first three) 0.27100247

P(all four) 0.00031348615

number: 19

P(first) 0.037226405

P(second|first) 0.039306037

P(third|first two) 0.09950083

P(forth|first three) 0.0004533991

P(all four) 6.6011204e-08

number: 20

P(first) 0.006751065

P(second|first) 0.011432434

P(third|first two) 0.0012566386

P(forth|first three) 0.60313356

P(all four) 5.8497175e-08

number: 21

P(first) 0.00077700504

P(second|first) 0.015191201

P(third|first two) 0.5139569

P(forth|first three) 0.0013561219

P(all four) 8.226998e-09

And then I calculate the L2 norm:

MQNGYTYEDYQDTAKWLLSHTEQRP.

[0.039207775, 0.04069929, 0.04001626, 0.018126749, 0.019602427, 0.019602427, 0.08058691, 0.081122674, 0.08115459, 0.08123663, 0.08123663, 0.021891603, 0.057534095, 0.30106863, 0.30093846, 0.30093846, 0.30048206, 0.30135483, 0.05976149, 0.059761565, 0.059761565, 0.062464472]

DO the same thing like above:

STAGKVIKCKAAVLWEEKKPFSIEE

[0.38184065, 0.13414468, 0.03431995, 0.027517552, 0.5201075, 0.5201075, 0.5199882, 0.5209995, 0.59083956, 0.28165892, 0.28165892, 0.28178003, 0.28209049, 0.043571703, 0.3949226, 0.3949226, 0.3957952, 0.40030295, 0.40036166, 0.079655625, 0.079655625, 0.08848169]

MVLSEGEWQLVLHVWAKVEADVAGH

[0.019436693, 0.051582653, 0.050369024, 0.35893875, 0.35886016, 0.35886016, 0.35538715, 0.35934687, 0.054046277, 0.06747922, 0.06747922, 0.3643123, 0.37322173, 0.37348044, 0.4040606, 0.4040606, 0.20279959, 0.19148451, 0.571188, 0.5484942, 0.5484942, 0.5432493]

From the result above, we can sort all of them.

[0.018126749, 0.019436693, 0.019602427, 0.019602427, 0.021891603, 0.027517552, 0.03431995, 0.039207775, 0.04001626, 0.04069929, 0.043571703, 0.050369024, 0.051582653, 0.054046277, 0.057534095, 0.05976149, 0.059761565, 0.059761565, 0.062464472, 0.06747922, 0.06747922, 0.079655625, 0.079655625, 0.08058691, 0.081122674, 0.08115459, 0.08123663, 0.08123663, 0.08848169, 0.13414468, 0.19148451, 0.20279959, 0.28165892, 0.28165892, 0.28178003, 0.28209049, 0.30048206, 0.30093846, 0.30093846, 0.30106863, 0.30135483, 0.35538715, 0.35886016, 0.35886016, 0.35893875, 0.35934687, 0.3643123, 0.37322173, 0.37348044, 0.38184065, 0.3949226, 0.3949226, 0.3957952, 0.40030295, 0.40036166, 0.4040606, 0.4040606, 0.5199882, 0.5201075, 0.5201075, 0.5209995, 0.5432493, 0.5484942, 0.5484942, 0.571188, 0.59083956]

The yellow is the 20 most closely matched entries.

The green is the 20 most different matched entries.

I know the L2 norm means The smaller, the more similar the two users are. The bigger, they are not similar.

We can find the first model accuracy is better than 4-gram model. And when this entry has been predicted before, the similarity is high, especially these entries in the same sample. But not all of entries follow this rule. I think it because my data is not enough and sometimes it predict wrongly.