NLP - HW 2

Guy Tevet (305257206), Gavriel Habib (304946445)

Question 1 – N-gram language model

(b) We calculated the perplexity of our model, using different values of λ_i. We checked all the combinations of $\lambda_{1,2} = 0,0.1, \dots, 1$, and $\lambda_3 = 1 - \lambda_1 - \lambda_2$.
The minimum of the perplexity is when $\lambda_1 = 0.4; \lambda_2 = 0.5; \lambda_3 = 0.1$. It means that bigram and trigram works well for this problem.
Be aware that for some combinations we got perplexity = ∞, but it happened only when $\lambda_3 = 1$. It means that in the test set there were bigrams or trigrams that have not been in the train set.

Full results:

[l1 = 0.0 , l2 = 0.0 , l3 = 1.0] #perplexity: 158.9
[l1 = 0.0 , l2 = 0.1 , l3 = 0.9] #perplexity: 94.5
[l1 = 0.0 , l2 = 0.2 , l3 = 0.8] #perplexity: 75.9
[l1 = 0.0 , l2 = 0.3 , l3 = 0.7] #perplexity: 65.3
[l1 = 0.0 , l2 = 0.4 , l3 = 0.6] #perplexity: 58.2
[l1 = 0.0 , l2 = 0.5 , l3 = 0.5] #perplexity: 53.2
[l1 = 0.0 , l2 = 0.6 , l3 = 0.4] #perplexity: 49.4
[l1 = 0.0 , l2 = 0.7 , l3 = 0.3] #perplexity: 46.6
[l1 = 0.0 , l2 = 0.8 , l3 = 0.2] #perplexity: 44.7
[l1 = 0.0 , l2 = 0.9 , l3 = 0.1] #perplexity: 44.0
[l1 = 0.0 , l2 = 1.0 , l3 = 0.0] #perplexity: inf
[l1 = 0.1 , l2 = 0.0 , l3 = 0.9] #perplexity: 84.2
[l1 = 0.1 , l2 = 0.1 , l3 = 0.8] #perplexity: 65.8
[l1 = 0.1 , l2 = 0.2 , l3 = 0.7] #perplexity: 57.1
[l1 = 0.1 , l2 = 0.3 , l3 = 0.6] #perplexity: 51.3
[l1 = 0.1 , l2 = 0.4 , l3 = 0.5] #perplexity: 47.2
[l1 = 0.1 , l2 = 0.5 , l3 = 0.4] #perplexity: 44.2
[l1 = 0.1 , l2 = 0.6 , l3 = 0.3] #perplexity: 41.9
[l1 = 0.1 , l2 = 0.7 , l3 = 0.2] #perplexity: 40.4
[l1 = 0.1 , l2 = 0.8 , l3 = 0.1] #perplexity: 39.9
[l1 = 0.1 , l2 = 0.9 , l3 = 0.0] #perplexity: inf
[l1 = 0.2 , l2 = 0.0 , l3 = 0.8] #perplexity: 68.7
[l1 = 0.2 , l2 = 0.1 , l3 = 0.7] #perplexity: 55.7
[l1 = 0.2 , l2 = 0.2 , l3 = 0.6] #perplexity: 49.6
[l1 = 0.2 , l2 = 0.3 , l3 = 0.5] #perplexity: 45.4
[l1 = 0.2 , l2 = 0.4 , l3 = 0.4] #perplexity: 42.4
[l1 = 0.2 , l2 = 0.5 , l3 = 0.3] #perplexity: 40.2
[l1 = 0.2 , l2 = 0.6 , l3 = 0.2] #perplexity: 38.8
[l1 = 0.2 , l2 = 0.7 , l3 = 0.1] #perplexity: 38.3
[l1 = 0.2 , l2 = 0.8 , l3 = 0.0] #perplexity: inf
[l1 = 0.3 , l2 = 0.0 , l3 = 0.7] #perplexity: 60.7
[l1 = 0.3 , l2 = 0.1 , l3 = 0.6] #perplexity: 49.9
[l1 = 0.3 , l2 = 0.2 , l3 = 0.5] #perplexity: 45.1
[l1 = 0.3 , l2 = 0.3 , l3 = 0.4] #perplexity: 41.8
[l1 = 0.3 , l2 = 0.4 , l3 = 0.3] #perplexity: 39.5
[l1 = 0.3 , l2 = 0.5 , l3 = 0.2] #perplexity: 38.0
[l1 = 0.3 , l2 = 0.6 , l3 = 0.1] #perplexity: 37.5
[l1 = 0.3 , l2 = 0.7 , l3 = -0.0] #perplexity: inf
[l1 = 0.4 , l2 = 0.0 , l3 = 0.6] #perplexity: 55.9
[l1 = 0.4 , l2 = 0.1 , l3 = 0.5] #perplexity: 46.2
[l1 = 0.4 , l2 = 0.2 , l3 = 0.4] #perplexity: 42.2
[l1 = 0.4 , l2 = 0.3 , l3 = 0.3] #perplexity: 39.5
[l1 = 0.4 , l2 = 0.4 , l3 = 0.2] #perplexity: 37.8
[l1 = 0.4 , l2 = 0.5 , l3 = 0.1] #perplexity: 37.2
[l1 = 0.4 , l2 = 0.6 , l3 = -0.0] #perplexity: inf
[l1 = 0.5 , l2 = 0.0 , l3 = 0.5] #perplexity: 53.1
[l1 = 0.5 , l2 = 0.1 , l3 = 0.4] #perplexity: 43.9
[l1 = 0.5 , l2 = 0.2 , l3 = 0.3] #perplexity: 40.3
[l1 = 0.5 , l2 = 0.3 , l3 = 0.2] #perplexity: 38.3
[l1 = 0.5 , l2 = 0.4 , l3 = 0.1] #perplexity: 37.5
[l1 = 0.5 , l2 = 0.5 , l3 = 0.0] #perplexity: inf
[l1 = 0.6 , l2 = 0.0 , l3 = 0.4] #perplexity: 51.8
[l1 = 0.6 , l2 = 0.1 , l3 = 0.3] #perplexity: 42.6
[l1 = 0.6 , l2 = 0.2 , l3 = 0.2] #perplexity: 39.5
[l1 = 0.6 , l2 = 0.3 , l3 = 0.1] #perplexity: 38.3
[l1 = 0.6 , l2 = 0.4 , l3 = -0.0] #perplexity: inf
[l1 = 0.7 , l2 = 0.0 , l3 = 0.3] #perplexity: 52.0
[l1 = 0.7 , l2 = 0.1 , l3 = 0.2] #perplexity: 42.3

[l1 = 0.7 , l2 = 0.2 , l3 = 0.1] #perplexity: 40.0
[l1 = 0.7 , l2 = 0.3 , l3 = -0.0] #perplexity: inf
[l1 = 0.8 , l2 = 0.0 , l3 = 0.2] #perplexity: 54.4
[l1 = 0.8 , l2 = 0.1 , l3 = 0.1] #perplexity: 43.6
[l1 = 0.8 , l2 = 0.2 , l3 = -0.0] #perplexity: inf
[l1 = 0.9 , l2 = 0.0 , l3 = 0.1] #perplexity: 61.9
[l1 = 0.9 , l2 = 0.1 , l3 = -0.0] #perplexity: inf
[l1 = 1.0 , l2 = 0.0 , l3 = 0.0] #perplexity: inf

MINIMUM: [l1 = 0.4 , l2 = 0.5 , l3 = 0.1] #perplexity: 37.2

Question 2 - Neural language model

(a) $J = CE(y, \hat{y}) = -\sum_i y_i \log(\hat{y}_i) = -\log\left(\frac{e^{\theta_i}}{\sum_j e^{\theta_j}}\right)$

$$\frac{\partial \mathrm{J}}{\partial \theta_i} = -\frac{\frac{\sum_j e^{\theta_j}}{e^{\theta_i}}\left(e^{\theta_i}\sum_j e^{\theta_j} - e^{\theta_i}e^{\theta_i}\right)}{\left(\sum_j e^{\theta_j}\right)^2} = \frac{e^{\theta_i}}{\sum_j e^{\theta_j}} - 1$$

$$\frac{\partial J}{\partial \theta_j} = -\frac{\frac{\sum_j e^{\theta_j}}{e^{\theta_i}}\left(-e^{\theta_j}e^{\theta_i}\right)}{\left(\sum_j e^{\theta_j}\right)^2} = \frac{e^{\theta_j}}{\sum_j e^{\theta_j}}$$

$$\frac{\partial J}{\partial \theta} = \hat{y} - y$$

(b) $J = CE(y, \hat{y}) = -\sum_i y_i \log(\hat{y}_i)$

$$\hat{y} = softmax(z) = softmax(hW_2 + b_2)$$

$$h = \sigma(a) = \sigma(xW_1 + b_1)$$

By the chain rule:

$$\frac{\partial J}{\partial x} = \frac{\partial J}{\partial z} * \frac{\partial z}{\partial h} * \frac{\partial h}{\partial a} * \frac{\partial a}{\partial x}$$

$$\frac{\partial J}{\partial z} = \hat{y} - y$$

$$\frac{\partial z}{\partial h} = W_2^T$$

$$\frac{\partial h}{\partial a} = \sigma(a) \circ \left(1 - \sigma(a)\right) = h \circ (1 - h)$$

$$\frac{\partial a}{\partial x} = W_1^T$$

Where ∘ means multiplication element-wise.

Therefore:

$$\frac{\partial J}{\partial x} = (\hat{y} - y)W_2^T \circ h \circ (1 - h)W_1^T$$

In addition (For back propagation):

$$\frac{\partial J}{\partial W_1} = \left(\left((\hat{y} - y)W_2^T \circ h \circ (1 - h)\right)^T x\right)^T$$

$$\frac{\partial J}{\partial W_2} = ((\hat{y} - y)^T h)^T$$

$$\frac{\partial J}{\partial b_1} = (\hat{y} - y)W_2^T \circ h \circ (1 - h)$$

$$\frac{\partial J}{\partial b_2} = (\hat{y} - y)$$

(c) Python implementation

(d) Final results:

```
#params: 104550
#train examples: 1118296
iter 1000: 7.626785
iter 2000: 7.668758
iter 3000: 7.615230
iter 4000: 7.644987
iter 5000: 7.598096
iter 6000: 7.584239
iter 7000: 7.508846
iter 8000: 7.479880
iter 9000: 7.465934
iter 10000: 7.382509
iter 11000: 7.351573
iter 12000: 7.342520
iter 13000: 7.371616
iter 14000: 7.332590
iter 15000: 7.309462
iter 16000: 7.343333
iter 17000: 7.322285
iter 18000: 7.321722
iter 19000: 7.319961
iter 20000: 7.286555
iter 21000: 7.253136
iter 22000: 7.196400
iter 23000: 7.156063
iter 24000: 7.129687
iter 25000: 7.140243
iter 26000: 7.050707
iter 27000: 7.049264
iter 28000: 7.000309
iter 29000: 7.038164
iter 30000: 6.949937
iter 31000: 6.970294
iter 32000: 6.962567
iter 33000: 6.943787
iter 34000: 6.929166
iter 35000: 6.935447
iter 36000: 6.885141
iter 37000: 6.921364
iter 38000: 6.956234
iter 39000: 6.929568
iter 40000: 6.911343
training took 27011 seconds
dev perplexity : 112.967665327
test perplexity will be evaluated only at test time!
```