Part1

1) Word embedding weight is 250 * 16 = 4000
   Embed to hid weight is 128 * 16 * 3 = 6144
   Hid bias is 128 * 1 = 128
   Hid to output weight is 250 * 128 = 32000
   Output bias is 250 * 1 = 250
   So total number of trainable parameters in the model is 42522, and the part with the largest number of trainable parameters is hidden layer to output layer.
2) The table should have 250^4 entries.

Part2

```
loss_derivative[2, 5] 0.001112231773782498
loss_derivative[2, 121] -0.9991004720395987
loss_derivative[5, 33] 0.0001903237803173703
loss_derivative[5, 31] -0.7999757709589483

param_gradient.word_embedding_weights[27, 2] -0.27199539981936866
param_gradient.word_embedding_weights[43, 3] 0.8641722267354154
param_gradient.word_embedding_weights[22, 4] -0.25467302023746485
param_gradient.word_embedding_weights[2, 5] 0.0

param_gradient.embed_to_hid_weights[10, 2] -0.6526990313918255
param_gradient.embed_to_hid_weights[15, 3] -0.13106433000472612
param_gradient.embed_to_hid_weights[30, 9] 0.11846774618169396
param_gradient.embed_to_hid_weights[35, 21] -0.1000452610460439

param_gradient.hid_bias[10] 0.2537663873815642
param_gradient.hid_bias[20] -0.03326739163635368

param_gradient.output_bias[0] -2.0627596032173052
param_gradient.output_bias[1] 0.0390200857392169
param_gradient.output_bias[2] -0.7561537928318482
param_gradient.output_bias[3] 0.21235172051123632
```

Part3

1) Yes, it gives sensible predictions. I used the words “he”, “is”, “the”, and the result shows me this:

```
he is the first Prob: 0.15188
he is the best Prob: 0.13222
he is the same Prob: 0.11907
he is the only Prob: 0.07581
he is the end Prob: 0.03262
he is the right Prob: 0.03201
he is the way Prob: 0.03124
he is the law Prob: 0.02577
he is the other Prob: 0.02490
he is the next Prob: 0.02146
```

Amount this, for example, “he is the other” is a sensible prediction, however, the 4-gram in the dataset is

```
The tri-gram "he is the" was followed by the following words in the training set:
    president (4 times)
    man (4 times)
    only (4 times)
    one (4 times)
    best (2 times)
    next (1 time)
    group (1 time)
    same (1 time)
    city (1 time)
    government (1 time)
    first (1 time)
    public (1 time)
    director (1 time)
    show (1 time)
    second (1 time)
```

where doesn’t has other.

2) Through the graph, I can conclude that words in the same cluster can replace each other grammatically or they have the same part of speech.

3) No, they are not close to each other. This is because as I conclude in question 2, “new” is adj and “york” is noun, they are not the same part of speech.
4) “government” and “political” are closer. This is because they are both noun and they might be described with similar adj word. My result shows that the distance between them is 1.0671188586251534 and the other pair is 1.6132898562556637.