Part1

1. Word embedding weight is 250 \* 16 = 4000

Embed to hid weight is 128 \* 16 \* 3 = 6144

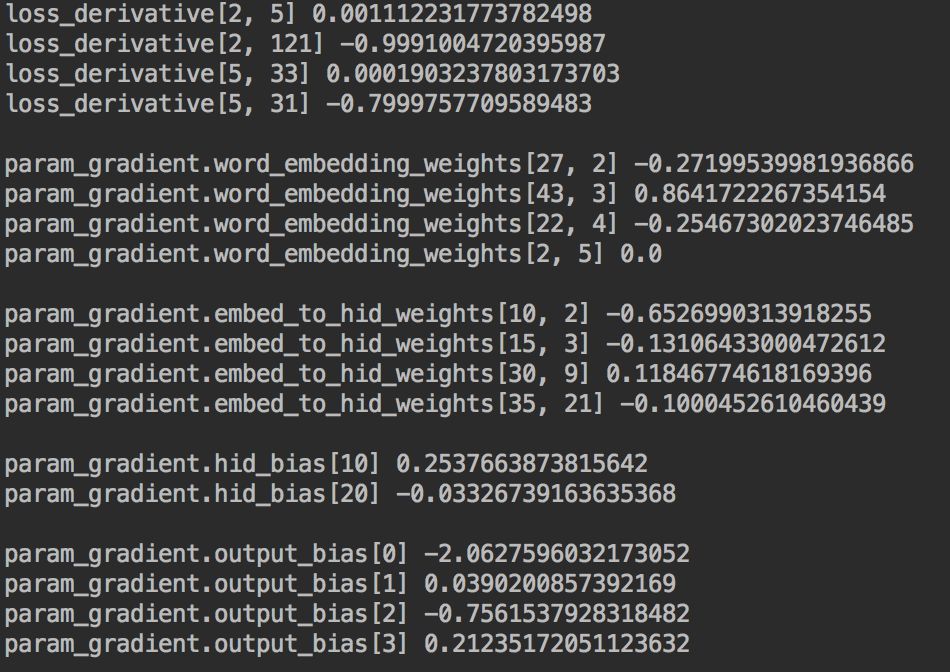
Hid bias is 128 \* 1 = 128

Hid to output weight is 250 \* 128 = 32000

Output bias is 250 \* 1 = 250

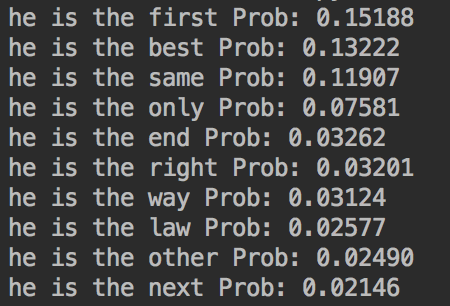
So total number of trainable parameters in the model is 42522, and the part with the largest number of trainable parameters is hidden layer to output layer.

1. The table should have 2504 entries.

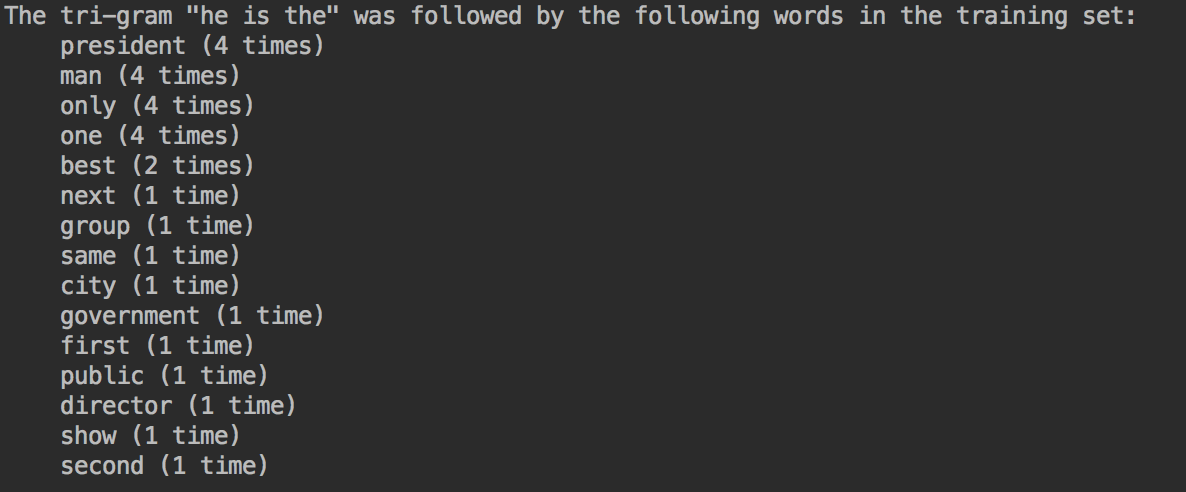
Part2 

Part3

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



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



where doesn’t has other.

1. 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.
2. 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.
3. “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.