# CSC421: Assignment 1

Due on Thrusday, Jan 31, 2019

**Zhongtian Ouyang**
**1002341012**

# Part 1

1.
word_embedding_weights: $250 \times 16 = 4000$
embed_to_hid_weights: $(16 \times 3) \times 128 = 6144$
hid_bias: 128
hid_to_output_weights: $128 \times 250 = 32000$
output_bias: 250
We can see that we have $4000 + 6144 + 128 + 32000 + 250 = 42522$ trainable parameters
hid_to_output_weights has the largest number of trainable parameters.

2.
In a 4-gram model, each word position has 250 possible choices. We need $250^4 = 3,906,250,000$ entries to store the counts of all possible 4-grams explicitly.

# Part 2

```
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] −0.2546730202374649
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] −0.10004526104604039

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] 0.21235172051123635
```

# Part 3

1.
The three-word phrase I pick is “he is the”. I get the following result from model:
he is the best Prob: 0.23524
he is the same Prob: 0.10197
he is the only Prob: 0.07203
he is the first Prob: 0.05316
he is the right Prob: 0.04734
he is the last Prob: 0.03425
he is the other Prob: 0.03199
he is the people Prob: 0.03023
he is the case Prob: 0.02653
he is the end Prob: 0.02199
All these predictions seems to be sensible. Among these predictions, “he is the right”, “he is the last”, “he is the other”, “he is the people”, “he is the case” and “he is the end” weren’t present in the dataset, and all these 4-grams seems plausible, especially “he is the right”, “he is the last”, “he is the other”.

2.
(should, might, may, would, could, can, will) is a cluster of words. We can see the usage of these words are very similar. They appears in the same position of a sentence with somewhat similar meanings.
(how, when, who, where, what) is a cluster of words. These words are all used as part of a question.
(does, do, did) and (be, been, was, is, were, are) are two clusters of words. Words in the first cluster are different tenses of “do” while words in the second cluster are different tenses of “be”.

3.
The word distance for “new” and “york” is 3.57. And from the tsne graph, we can see that they are not really close to each other. The reason is that even though the phrase “new york” appears a lot, “new” and “york” have different functions. The word “new” is also used a lot as a adjective describing items, while “york” can’t be used like that.

4.
The word distance for “government” and “university” is 1.044 while the word distance for “government” and “political” is 1.376. It is obvious that (government, university) is closer together. The reason could be that even though “government” and “policical” have related meanings, “government” and “university” have closer encodings because they are both nouns and types of organizations. They appear in similar positions of a sentence.

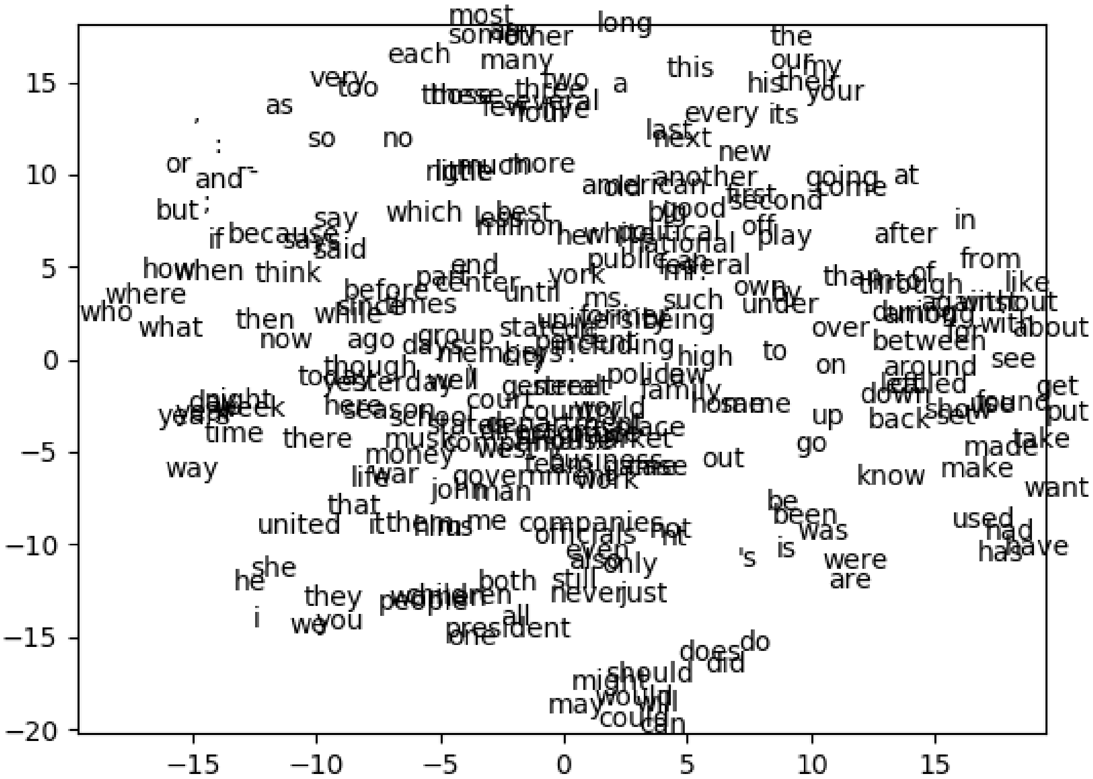

Figure 1: TSNE Graph