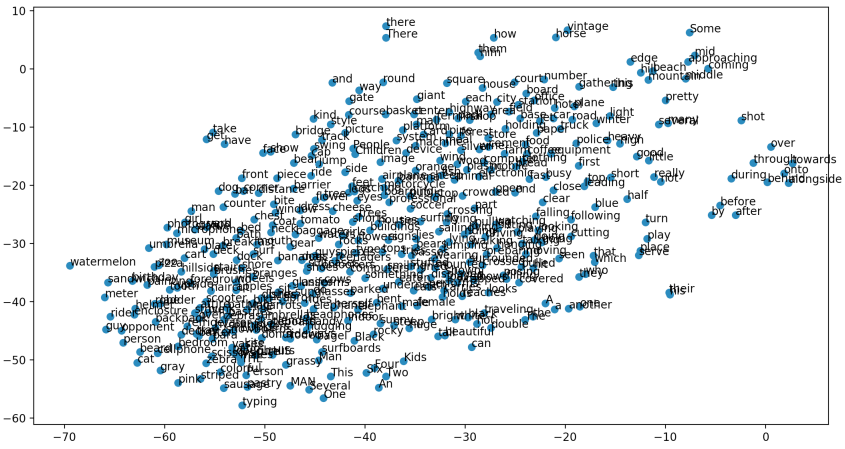
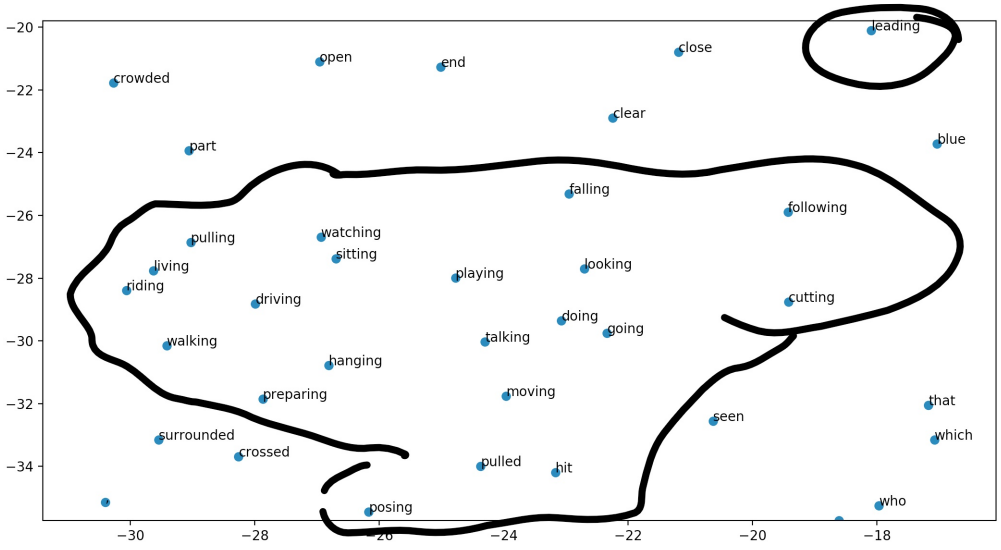
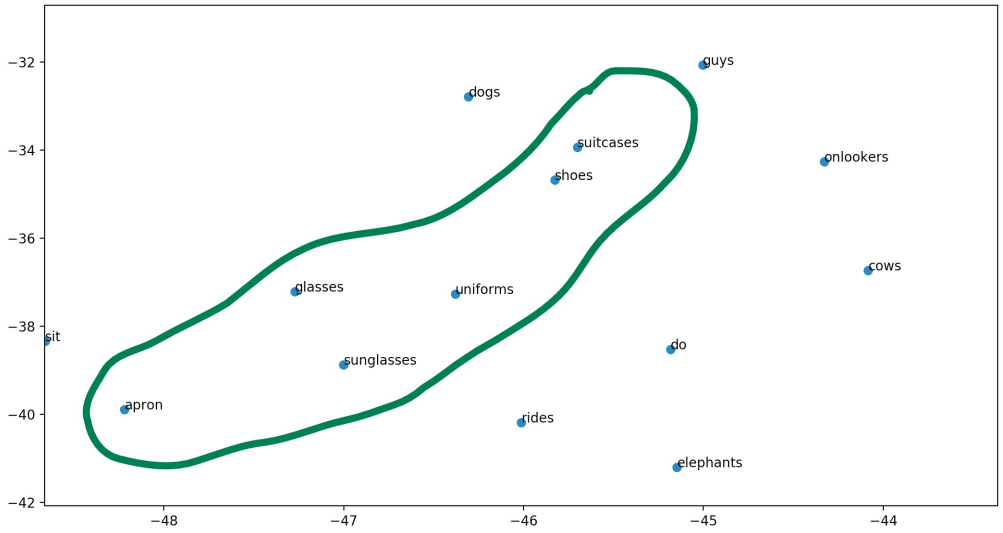
Problem1 Visualize Word Embedding

I increase the amount of words from the original, I choose a subset of [200, 600] from the whole vocabulary.



For this part, I find verb with ing are grouped together as well as all nouns for apparel.

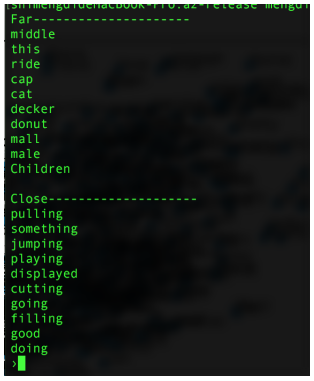




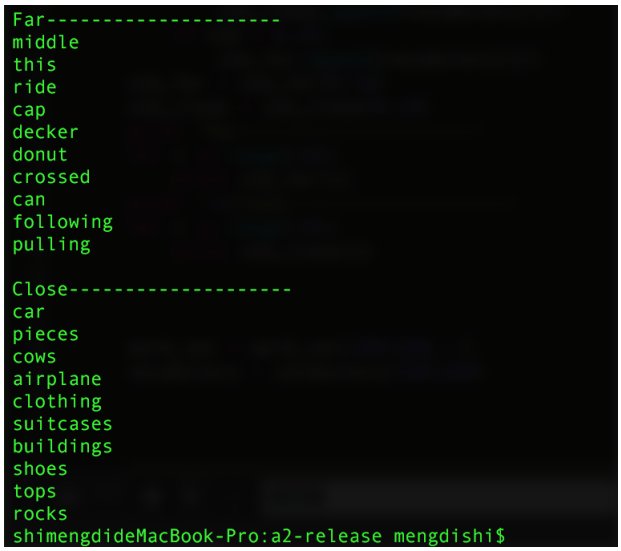
I select some word pairs from two groups, and calculate their cosine similarities in the 2-D diagram:

```
def similarity(x, y):
     return numpy.dot(x, y)/(numpy.linalg.norm(x) * numpy.linalg.norm(y))


#Verbs+ing group
talking = numpy.array([-24.29, -30.09])
doing = numpy.array([-23.07, -29.41])
going = numpy.array([-22.35, -29.85])
looking = numpy.array([-22.71, -22.77])

#apparel group
glasses = numpy.array([-47.24, -37.37])
uniforms = numpy.array([-46.37, -37.42])
shoes = numpy.array([-45.82, -34,72])

similarity(talking, doing)      -- 999902
similarity(doing, going)        -- 0.999747
similarity(doing, looking)      -- 0.992939
similarity(doing, glasses)      -- 0.972198
similarity(glasses, uniforms)   -- 0.999952
similarity(doing, uniforms)     -- 0.974427
```

Besides, I find that they also show cosine similarity in full dimension.
'Far' group means cosine similarity < 0.25, and 'Close' group means large cosine similaritys > 0.75.
As shown in the following diagram, I calculate the cosine similarity to words 'doing'. I find that words has larger cosine similarity to 'doing' has stronger syntactic and meaning relationships(verb + ing form are more likely to be in close group). On the contrary, smaller cosine similarity means less relation close.

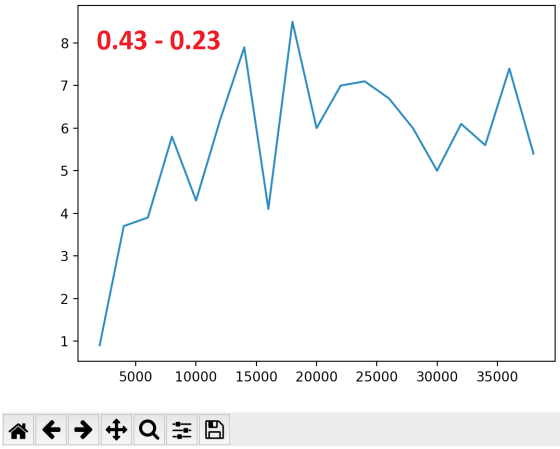


For the noun 'uniforms', the conclusion are the same. As we can see, the noun 'apparel' words are in close groups. In the 'Far' group, many verb words(including verb + ing).

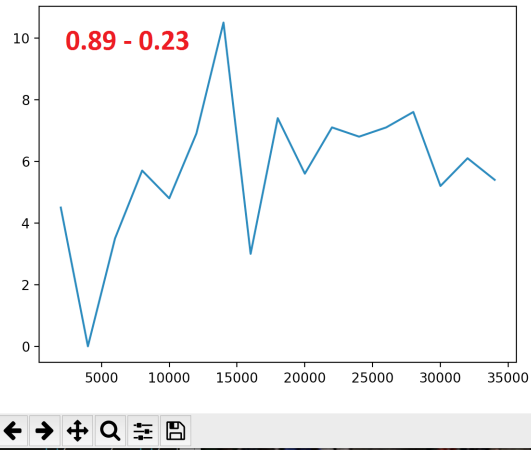


Problem2 Train and Evaluate Model

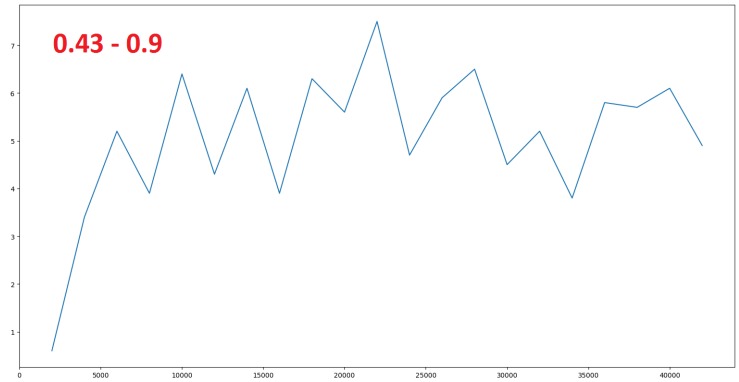
1. My 4 setting of parameters of learning rate and momentum are:
   $[0.43, 0.23] \rightarrow$ bleu score $= 8.5$



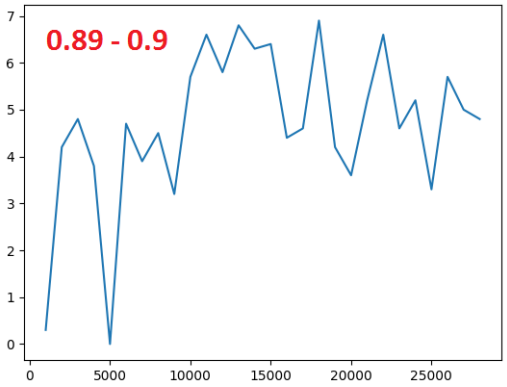
   $[0.89, 0.23] \rightarrow$ bleu score $= 10.5$



$[0.43, 0.9] \rightarrow$ bleu score $= 7.5$

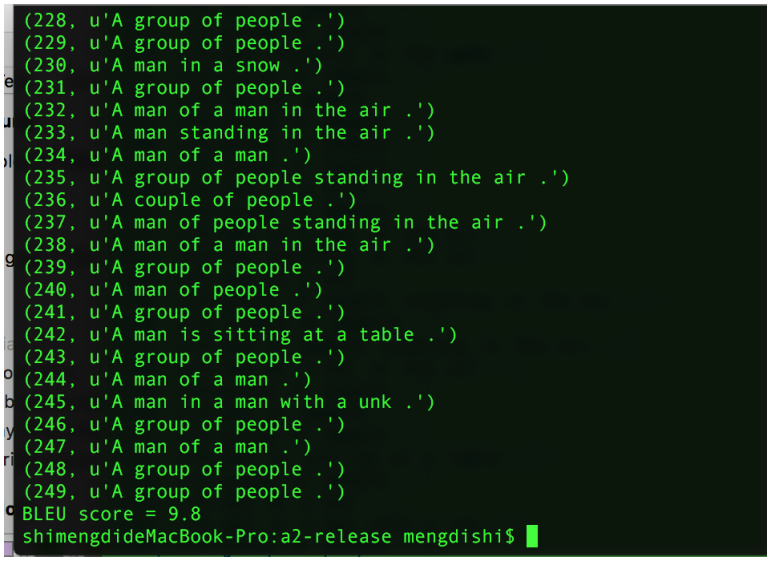


$[0.89, 0.9] \rightarrow$ bleu score $= 6.9$

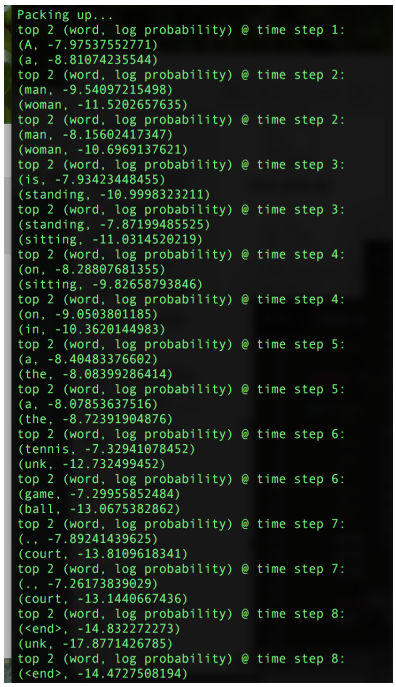


From the experiments, I find that large momentum(0.9) will cause diagram oscillate vertically. Especially when learning rate and momentum are large at the same time, it will cause each time gradient update change a lot, hence more likely to be affected by noice, which could be reflected by the 4th diagram, whose oscilation and amplitude change obviously. To the learning rate, smaller laerning rate smooth curve, but more likely to be trapped in local minimum or maybe saddle point. On the contrary, large learning rate may cause it update more quickly, more likely to avoid local minimum. Therefore, it make sense that the 2nd setting of parameters $[0.89, 0.23]$ get the highest bleu scores.

2. From result above, I set learning rate as 0.89 and momentum as 0.23 as my best model. And get the bleu score of 9.8 in testing:



3. By setting beam width as 2, I get the following output:
It actually keep computing conditional probability given front positions candidates. As we can see can see from the picture, for the 3rd words, although to the 3rd word, 'man' has larger probability follow 'A' than 'woman' follow 'A', so 'A man' has a bigger chance to be the first 2 words. However, for the 3rd word, although 'is' has bigger probability than 'standing', it is actually not the 3rd word in the final interpreter. Since, the probability of a interpreter sentence use a chain rule to calculate, which means it will consider every words probability to get the total probility for the sentence. Hence, althouh some words has tight bound locally, they may not get together at the final.



(4) Just as shown at the above example, some words may have very tight bound, and hence will has very high probability with each other, which will cause bad captions if we just choose sentences based on information locally. But by using beem search, the probability of a word given several front words, which will gain more meanings to who sentences, which will has more chance to make good caption.