1. Hyper-parameters explored w/ a description of what each parameter controls and how varying each is likely to affect the training.

2.

a) Experiment with different batch size

###
###########

skip_window - How many words to consider left and right.

1. num_skips - How many times to reuse an input to generate a label.
2. max_num_steps - maximum training step
3. batch_size - small subset of training data

Configuration 1:
batch_size = 16, skip_window = 4, num_skips = 8, max_num_steps = 200001
cross entropy - Average loss at step 200000 : 2.7589261856555938
nce - Average loss at step 200000 : 744.2880624351501

Configuration 2:
batch_size = 32, skip_window = 4, num_skips = 8, max_num_steps = 200001
cross entropy - Average loss at step 200000 : 3.432030707454681
nce - Average loss at step 200000 : 16.635339168804208

Configuration 3:
batch_size = 64, skip_window = 4, num_skips = 8, max_num_steps = 200001
cross entropy - Average loss at step 200000 : 4.126185366058349
nce - Average loss at step 200000 : 1.3590045896619558

Configuration 4:
batch_size = 128, skip_window = 4, num_skips = 8, max_num_steps = 200001
cross entropy - Average loss at step 200000 : 4.822538606834412
nce - Average loss at step 200000 : 1.388805592134595

Configuration 5:
batch_size = 256, skip_window = 4, num_skips = 8, max_num_steps = 200001
cross entropy - Average loss at step 200000 : 5.509526480007172
nce - Average loss at step 200000 : 1.4231902356922626

Configuration 6:
batch_size = 512, embedding_size = 128, skip_window = 4, num_skips = 8, max_num_steps = 200001
cross entropy - Average loss at step 200000 : 6.176199967956543
nce - Average loss at step 200000 : 1.5417432206258177

Observation:
Ideal batch_size = 64, As batch size increases time taken increases and value of average loss increases which is not preferred (but after going below 64 for batch size overfitting happens as observed)

b) Experiment with different skip window and num skips

###
###########

Configuration 1:
batch_size = 128, skip_window = 2, num_skips = 4, max_num_steps = 200001

cross entropy - Average loss at step   200000 :   4.6985910774707795
nce - Average loss at step   200000 :   1.3264403286993502

Configuration 2:
batch_size = 128, skip_window = 4, num_skips = 8, max_num_steps   = 200001
cross entropy - Average loss at step   200000 :   4.822538606834412
nce - Average loss at step   200000 :   1.388805592134595

Configuration 3:
batch_size = 128, skip_window = 8, num_skips = 16, max_num_steps   = 200001
cross_entropy - Average loss at step   200000 :   4.83617493801117
nce - Average loss at step   200000 :   1.5599304841637611

Observation:
Ideal skip_window = 2, num_skips = 4, As skip window and num skips increase, value of average loss increases which is not preferred and also the time taken increases

c) Experiment with different max num skips
###
###########

Configuration 1:
batch_size = 128, skip_window = 4, num_skips = 8, max_num_steps   = 50001
cross entropy - Average loss at step   50000 :   4.831868802452087
nce - Average loss at step   50000 :   1.25496210757792

Configuration 2:
batch_size = 128, skip_window = 4, num_skips = 8, max_num_steps   = 100001
cross entropy - Average loss at step   100000 :   4.822567562198639
nce - Average loss at step   100000 :   1.29812714253664

Configuration 3:
batch_size = 128, skip_window = 4, num_skips = 8, max_num_steps   = 200001
cross entropy - Average loss at step   200000 :   4.822538606834412
nce - Average loss at step   200000 :   1.388805592134595

Configuration 4:
batch_size = 128, skip_window = 4, num_skips = 8, max_num_steps   = 400001
cross entropy - Average loss at step   400000 :   4.822009043216705
nce - Average loss at step   400000 :   1.1749355765908958

Observation:
Ideal max_num_steps = 400001, As max_num_steps increase, value of average loss decreases but also the time taken increases

Ideal Configuration:
batch_size = 64, skip_window = 2, num_skips = 4, max_num_steps   = 400001

->cross entropy - Average loss at step   400000 :   4.031693156671524

Word analogy output on dev.txt
Generated by: score_maxdiff.pl
Mechanical Turk File: word_analogy_dev_mturk_answers.txt

```
Test File:
word_analogy_test_predictions_cross_entropy.txt
Number of MaxDiff Questions:                       914
Number of Least Illustrative Guessed Correctly:    264
Number of Least Illustrative Guessed Incorrectly: 650
Accuracy of Least Illustrative Guesses:            28.9%
Number of Most Illustrative Guessed Correctly:     333
Number of Most Illustrative Guessed Incorrectly:   581
Accuracy of Most Illustrative Guesses:             36.4%
Overall Accuracy:                                  32.7%
```

->nce - Average loss at step   400000 :   1.1813166753590107

Word analogy output on dev.txt

```
Generated by:                                      score_maxdiff.pl
Mechanical Turk File:                             word_analogy_dev_mturk_answers.txt
Test File:                                       word_analogy_test_predictions_nce.txt
Number of MaxDiff Questions:                       914
Number of Least Illustrative Guessed Correctly:    280
Number of Least Illustrative Guessed Incorrectly: 634
Accuracy of Least Illustrative Guesses:            30.6%
Number of Most Illustrative Guessed Correctly:     330
Number of Most Illustrative Guessed Incorrectly:   584
Accuracy of Most Illustrative Guesses:             36.1%
Overall Accuracy:                                  33.4%
```

2) Results on the analogy task for five different configurations of hyper-parameters, along with the best configuration

-> Ideal vs default configuration

Default Configuration:

batch_size = 128, skip_window = 4, num_skips = 8, max_num_steps   = 200001
cross entropy - Overall Accuracy: 32.9%
nce - Overall Accuracy: 33.2%

Configuration 1: Ideal

batch_size = 64, skip_window = 2, num_skips = 4, max_num_steps   = 400001
cross entropy - Overall Accuracy:32.7%
nce - Overall Accuracy : 33.4%

Observation - Approximately the same result

-> Experimentation with different batch size

Default Configuration:

batch_size = 128, skip_window = 4, num_skips = 8, max_num_steps   = 200001
cross entropy - Overall Accuracy: 32.9%
nce - Overall Accuracy: 33.2%

Configuration 2:

batch_size = 64, skip_window = 4, num_skips = 8, max_num_steps = 200001
cross entropy - Overall Accuracy: 33.2%
nce - Overall Accuracy: 33.1%

Observation - Noise decreases as batch size decreases

-> Experimentation with different skip window and num skips

Default Configuration:
batch_size = 128, skip_window = 4, num_skips = 8, max_num_steps = 200001
cross entropy - Overall Accuracy: 32.9%
nce - Overall Accuracy: 33.2%

Configuration 3:
batch_size = 128, skip_window = 2, num_skips = 4, max_num_steps = 200001
cross entropy - Overall Accuracy: 32.9%
nce - Overall Accuracy: 33.0%

Configuration 4:
batch_size = 128, skip_window = 1, num_skips = 2, max_num_steps = 200001
cross entropy - Overall Accuracy: 32.4%
nce - Overall Accuracy: 33.1%

Observation - Noise of nce can slightly increase as skip window and num skips decreases, where as noise of cross entropy can increase by a higher margin as skip window and num skips decreases.

-> Experimentation with different max num steps

Default Configuration:
batch_size = 128, skip_window = 4, num_skips = 8, max_num_steps = 200001
cross entropy - Overall Accuracy: 32.9%
nce - Overall Accuracy: 33.2%

Configuration 5:
batch_size = 128, skip_window = 4, num_skips = 8, max_num_steps = 50001
cross entropy - Overall Accuracy: 32.9%
nce - Overall Accuracy: 33.3%

Configuration 6:
batch_size = 128, skip_window = 4, num_skips = 8, max_num_steps = 20001
cross entropy - Overall Accuracy: 32.9%
nce - Overall Accuracy: 33.3%

Observation - Noise almost remains the same as max num steps decreases

3) Top 20 similar words according to NCE

Nearest to first: their, some, her, no, others, western, human, what, so, american, anti, how, all, a, other, william, political, abuse, economic, they

Nearest to american: could, can, would, march, november, august, december, have, june, might, october, will, must, january, has, february, july, until, traits, september

Nearest to would: he, this, there, often, now, not, also, still, sometimes, generally, usually, that, very, she, what, similar, but, to, widely, difficult

Top 20 similar words according to cross entropy

Nearest to first: through, into, between, within, since, during, against, in, with, until, at, before, after, on, including, under, when, across, began, due

Nearest to american: many, both, some, several, various, any, each, most, every, those, only, among, especially, just, include, certain, making, viking, following, the

Nearest to would: beginning, end, america, soviet, u, bottom, addition, harriman, case, west, netherlands, rest, atlantic, southern, gooding, middle, eastern, east, psychiatric, popularity

4) Summary of the justification behind the NCE method loss.

NCE reduces density estimation to the probabilistic binary classification, hence this is used on un-normalized models. NCE trains a logistic regression to distinguish data from positive (actual data) and noise distribuition (negative) samples. If the model contains data distribution with an approximate normalized form, then it returns the perfect normalized form. Key advantage is that the training time is independent of the vocabulary size.

An auxiliary binary classification is created from the training data and noise distribution to give positive and negative samples respectively. Unigram distribution of input data is used to avoid zero probabilities for noise distirbution.
We then fit the model by maximizing the log posterior probability over positive samples and averaged over both positive and negative samples. As sum over the k negative samples is used, training time becomes linear, as we won’t use the entire vocabulary.