Learnings:

Problems of words appearing in high frequency:

There is a problem with words like: the, and etc, due to their higher frequency and hence they have bigger word vectors, resulting in bigger probabilities. This is true and correct but if we want to show their semantic properties better we have to take care of this high frequency effect. One of the crude ways of solving this problem is by locking the frequency's effect.

Stochastic gradient descent:

Usually our corpus has billions of words we cannot evaluate for the whole corpus in every step of the gradient descent. So what we take is a window (importantly a different window in every step) and calculate the gradient within it (remember: for better results do this in mini batches of 32 or 64).

Problem with stochastic gradient descent:

We might not even encounter many of the words when using windows. The windows might not hold enough variety of words and we might end up with word vector matrices with most of the rows zeroes (initial values). We could solve this by using sparse matrix matrix functions to selectively update certain rows of u and v, or we can use hashing.

WordToVec models:

Skip-Grams model:

In this model we predict the context word given the center word.

Continuous bag of words model:

In this model a bag of context words is used to predict the center word.

Negative sampling optimization:

While calculating the probability of the probability function ,

- $P(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$