(a) It's intuitive that only when w equals to o, y_w wont be 0.

(b) First, we have $\hat{y} = softmax(U^T v_c)$. Then

$$\begin{aligned}\frac{\partial J}{\partial v_c} &= \frac{\partial J}{\partial U^T v_c}\frac{\partial U^T v_c}{\partial v_c} \\ &= U^T(\hat{y} - y)\end{aligned}$$

(c) The first steps are the same as the last problem. We have $\hat{y} = softmax(U^T v_c)$. Then

$$\begin{aligned}\frac{\partial J}{\partial u_w} &= \frac{\partial J}{\partial U^T v_c}\frac{\partial U^T v_c}{\partial u_w} \\ &= v_c(\hat{y} - y)^T\end{aligned}$$

(d) The process is as below,

$$\begin{aligned}\frac{\partial \sigma(x)}{\partial x} &= \frac{e^x(e^x+1) - e^x e^x}{\partial (e^x+1)^2} \\ &= \sigma(x)(1-\sigma(x))\end{aligned}$$

(e) The partial derivative for v_c is

$$(\sigma(u_o^T v_c) - 1)u_o + \sum_{k=1}^{K}(\sigma(-u_k^T v_c) - 1)u_k$$

.

The partial derivative for u_o is

$$(\sigma(u_o^T v_c) - 1)v_c$$

and for u_k is

$$(\sigma(-u_k^T v_c) - 1)v_c$$

.

This derivative is easier to compute because we don't need to compute the partial derivative for all words in vocabulary.

(f) (i) $\sum_{j \in window} \frac{\partial J(v_c, w_j, U)}{\partial U}$

(ii) $\sum_{j \in window} \frac{\partial J(v_c, w_j, U)}{\partial v_c}$

(iii) 0

After training, final loss is 9.367644.
The results of word vectors are as shown in the belowing picture.

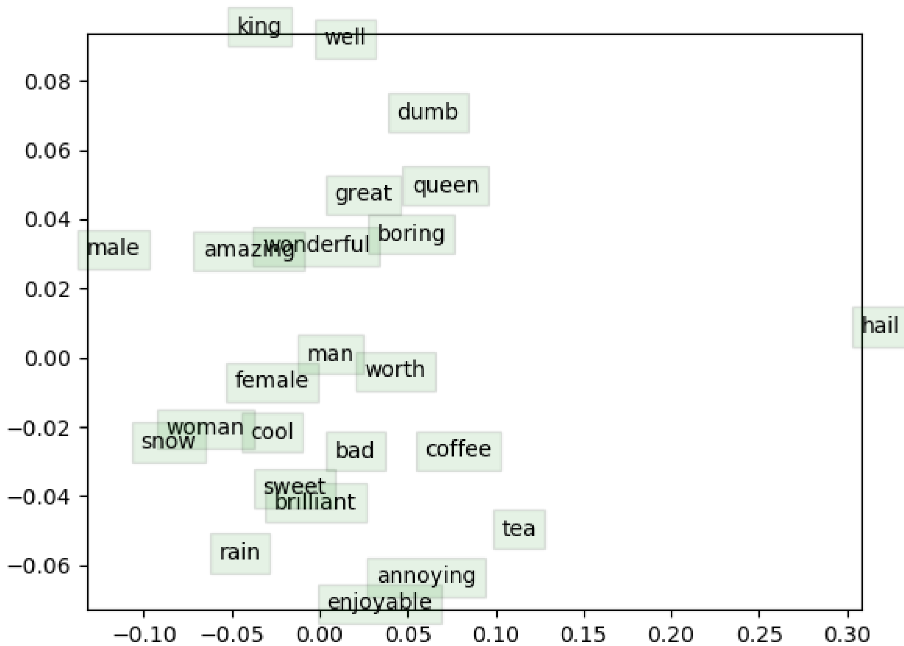

Figure 1: word vectors

We can see that

1. Some words like amazing, wonderful and boring are clustered well
2. Some words like hail, snow and rain are not.
3. It's interesting to see that the relative distance between male to king and female to queen is equal.