Assignment 2

a) Show that the naïve-softmax loss given in equation (2) is the same as the cross entropy loss between y and yhat i.e. show that $-\sum_{w\epsilon\ vocab} y_w \log(yhat_w) = -\log(y_o)$

Ans) y_w = 1 if w = o

= 0 otherwise

Therefore $-\sum_{w\epsilon\ vocab} y_w \log(yhat_w) = -\log(y_o)$

b) Compute the partial derivative of $J_{naive\text{-}softmax}$ (V_c, o, U) with respect to V_c. Please write your answer in terms of y, yhat and U.

Ans) = -U_o + $yhat \cdot U$

c) Compute the partial derivative of $J_{naive\text{-}softmax}$ ($V_{c,}$ o, U) with respect to each of the outside word vectors U_w. There will be two cases when w=o, the true outside word vector and when $w \neq o$ for all other words. Please write your answer in terms of y, yhat and V_c.

Ans) $\frac{\partial J}{\partial U} = yhat \cdot V_c$

When U = o $\frac{\partial J}{\partial U_o} += -V_c$

d) The sigmoid function is given by $\sigma(x) = \frac{1}{1+e^{-x}}$; compute the derivative of $\sigma(x)$ with respect to x.

Ans) $\frac{d\sigma(x)}{dx} = \sigma(x)(1 - \sigma(x))$

e) Now we shall consider the negative sampling loss, which is an alternative to the naïve softmax loss. Assume that K negative samples (words) are drawn from the vocabulary. For simplicity of notation we shall refer to them as w_1, w_2, …, w_k and their outside vectors as U_1, U_2, …, U_k. Note that $o \notin \{w_1, w_2, \dots, w_k\}$. For a center word c and an outside word o, the negative sampling loss function is given by $J_{neg-sample}(V_c, o, U) = -\log(\sigma(U_o{}^T V_c) - \sum_{k=1}^{K} \log(\sigma(-U_k{}^T V_c))$ for a sample w_1, ..., w_k where σ is the sigmoid function. Repeat parts b and c for the negative sampling loss function.

Ans) $\frac{\partial J}{\partial V_c} = -\{1 - \sigma(U_o{}^T V_c)\}U_o + \sum_{k=1}^{K}\{1 - \sigma(-U_k{}^T V_c)\}U_k$

$\frac{\partial J}{\partial U_o} = -V_c\{1 - \sigma(U_o{}^T V_c)$ $\frac{\partial J}{\partial U_k} = V_c\{1 - \sigma(-U_k{}^T V_c)\}$

f) Suppose the center word c = w_t and the context window is (w_{t-m}, …, w_{t-1}, w_t, w_{t+1}, …, w_{t+m}), where m is the context window size. Recall that for the skip gram version of word2vec, the total loss for the context window is $J_{skip-gram}(V_c, w_{t-m}, \dots, w_{t+m}, U) = \sum_{-m\leq j\leq m, j\neq 0} J(V_c, w_{t+j}, U)$. Here $J(V_c, w_{t+j}, U)$ represents an arbitrary loss term for the center word c = w_t and outside word w_{t+j}.

$J(V_c, w_{t+j}, U)$ could be $J_{\text{naive-softmax}}$ or $J_{\text{neg-sample}}$, depending on your implementation. Write down three partial derivatives

i) $dJ_{\text{skip-gram}}$ (V_c, w_{t-m}, …, w_{t+m}, U)/dU

ii) $dJ_{\text{skip-gram}}$ (V_c, w_{t-m}, …, w_{t+m}, U)/dV_c

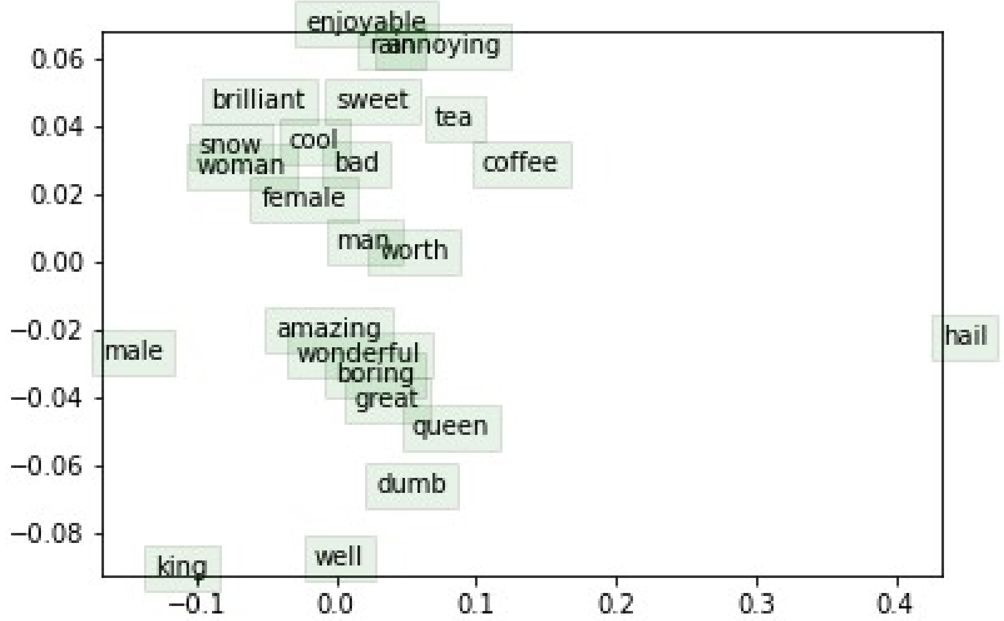
iii) $dJ_{\text{skip-gram}}$ (V_c, w_{t-m}, …, w_{t+m}, U)/dV_w when w $\neq$ c

Write your answers in terms of dJ (V_c, w_{t+j}, U)/dU and dJ (V_c, w_{t+j}, U)/dV_c

Ans) $dJ_{\text{skip-gram}}$(V_c, w_{t-m}, …, w_{t+m}, U)/dU = $\sum_{-m\leq j\leq m, j\neq 0} \mathrm{dJ}(\mathrm{Vc}, w_{t+j}, \mathrm{U})/\mathrm{dU}$

$dJ_{\text{skip-gram}}$ (V_c, w_{t-m}, …, w_{t+m}, U)/dV_c = $\sum_{-m\leq j\leq m, j\neq 0} \mathrm{dJ}(\mathrm{Vc}, w_{t+j}, \mathrm{U})/\mathrm{d}V_c$

$dJ_{\text{skip-gram}}$(V_c, w_{t-m}, …, w_{t+m}, U)/dV_w = 0



- From the plot we can see that closely related words like amazing and wonderful, tea and coffee, etc. are close to each other.
- It can be also seen that if you draw a line starting and male and ending at female, and another line starting at king and ending at queen, both these lines are parallel to each other and have roughly the same distance from each other.
- However some words that have no correlation are close to each other like wonderful and boring, man and worth.