1. Written : Understanding Word2Vec

(a) $y_w = \begin{cases} 1 & (w=0) \\ 0 & (w\neq 0) \end{cases}$

$$-\sum_{w\in Vocab} y_w \log \hat{y}_w = -y_0 \log \hat{y}_0 - \sum_{w\neq 0} y_w \log \hat{y}_w = -\log \hat{y}_0$$

$$J_{naive-softmax} = -\log \frac{\exp(u_0^T v_c)}{\sum_{w\in Vocab} \exp(u_w^T v_c)} = -\log \hat{y}_0$$

(b)
$$\frac{\partial J_{naive-softmax}}{\partial v_c} = \frac{\partial}{\partial v_c}\left(-\log \frac{\exp(u_0^T v_c)}{\sum_{w\in Vocab} \exp(u_w^T v_c)}\right) = \frac{\partial}{\partial v_c}\left(-\log \exp(u_0^T v_c) + \log \sum_{w\in Vocab} \exp(u_w^T v_c)\right)$$

$$= -\frac{1}{\exp(u_0^T v_c)} \cdot \exp(u_0^T v_c) \cdot u_0 + \frac{1}{\sum_{w\in Vocab} \exp(u_w^T v_c)} \cdot \left(\sum_{w\in Vocab} \exp(u_w^T v_c) \cdot u_w\right)$$

$$= -u_0 + \sum_{w\in Vocab}\left(\frac{\exp(u_w^T v_c)}{\sum_{w\in Vocab} \exp(u_w^T v_c)} \cdot u_w\right) = -u_0 + \sum_{w\in Vocab} \hat{y}_w \cdot u_w$$

$$= -y_0 u_0 + \sum_{w\in Vocab} \hat{y}_w u_w = \sum_{w\in Vocab} -y_w \cdot u_w + \hat{y}_w \cdot u_w$$

$$= \sum_{w\in Vocab} \underbrace{(\hat{y}_w - y_w)}_{scalar} \cdot \underbrace{u_w}_{vector\ (d\times 1)} \in \mathbb{R}^{d\times 1}$$

$U = [u_1, u_2, \cdots, u_0, \cdots u_{|Vocab|}] \in \mathbb{R}^{d\times |Vocab|}$
$y \in \mathbb{R}^{|Vocab|\times 1}$, $\hat{y} \in \mathbb{R}^{|Vocab|\times 1}$

$$\therefore \frac{\partial J_{naive-softmax}}{\partial v_c} = U\cdot(\hat{y}-y)$$

(c)
$$\frac{\partial J_{naive-softmax}}{\partial u_w} = \frac{\partial}{\partial u_w}\left(-\log\left(\frac{\exp(u_0^T v_c)}{\sum_w \exp(u_w^T v_c)}\right)\right) = \frac{\partial}{\partial u_w}\left(-u_0^T v_c + \log \sum_w \exp(u_w^T v_c)\right)$$

i) $w=0$

$$\frac{\partial}{\partial u_w}\left(-u_0^T v_c + \log \sum_w \exp(u_w^T v_c)\right) = -v_c + \frac{\exp(u_w^T v_c)}{\sum_w \exp(u_w^T v_c)} \cdot v_c = -v_c + \hat{y}_w \cdot v_c = (\hat{y}_w - 1)\cdot v_c$$

ii) $w\neq 0$

$$\frac{\partial}{\partial u_w}\left(-u_0^T v_c + \log \sum_w \exp(u_w^T v_c)\right) = 0 + \frac{\exp(u_w^T v_c)}{\sum \exp(u_w^T v_c)} \cdot v_c = \hat{y}_w \cdot v_c$$

$$(\therefore) \quad \frac{\partial J_{naive-softmax}}{\partial u_w} = (\hat{y}_w - y_w)\, v_c \quad \begin{cases} (\hat{y}_w - 1)\cdot v_c & (w=0) \\ \hat{y}_w \cdot v_c & (w\neq 0) \end{cases}$$

(d)
$$\frac{\partial J_{naive-softmax}}{\partial U} = \left[\frac{\partial J}{\partial u_1}, \frac{\partial J}{\partial u_2}, \cdots, \frac{\partial J}{\partial u_{|Vocab|}}\right] = [(\hat{y}_1 - y_1)\cdot v_c, (\hat{y}_2 - y_2)\cdot v_c, \cdots, (\hat{y}_{|Vocab|} - y_{|Vocab|})\cdot v_c]$$

$$= v_c \cdot (\hat{y} - y)^T$$

(e)
$$J_{neg}(v_c, o, U) = -\log \sigma(u_0^T v_c) - \sum_{s=1}^{K} \log \sigma(-u_{w_s}^T v_c)$$

(i)
$$\frac{\partial}{\partial x}\sigma(x) = \frac{\partial}{\partial x}\left(\frac{1}{e^{-x}+1}\right) = e^{-x}\cdot\frac{1}{(e^{-x}+1)^2} = \frac{1}{e^{-x}+1}\cdot\left(1 - \frac{1}{e^{-x}+1}\right) = \sigma(x)\cdot(1-\sigma(x))$$

$$\frac{\partial J}{\partial v_c} = -\frac{1}{\sigma(u_0^T v_c)} \cdot \sigma(u_0^T v_c)\cdot(1-\sigma(u_0^T v_c))\cdot u_0 - \sum_{s=1}^{K} \frac{1}{\sigma(-u_{w_s}^T v_c)} \cdot \sigma(-u_{w_s}^T v_c)(1-\sigma(-u_{w_s}^T v_c))\cdot(-u_{w_s})$$

$$= (\sigma(u_0^T v_c) - 1)\cdot u_0 + \sum_{s=1}^{K} (1-\sigma(-u_{w_s}^T v_c))\cdot u_{w_s} = (\sigma(u_0^T v_c) - 1)\cdot u_0 + \sum_{s=1}^{K} \sigma(u_{w_s}^T v_c)\cdot u_{w_s}$$

$$(*\ \ 1-\sigma(-x)) = 1 - \frac{1}{1+e^{x}} = \frac{e^{x}}{1+e^{x}} = \frac{1}{e^{-x}+1} = \sigma(x))$$

$$\frac{\partial J}{\partial u_0} = -\frac{1}{\sigma(u_0^T v_c)} \cdot \sigma(u_0^T v_c)\cdot(1-\sigma(u_0^T v_c))\cdot v_c = (\sigma(u_0^T v_c) - 1)\cdot v_c$$

$$\frac{\partial J}{\partial u_{w_s}} = -\frac{1}{\sigma(-u_{w_s}^T v_c)} \cdot (1-\sigma(-u_{w_s}^T v_c))\cdot(-v_c) = \sigma(u_{w_s}^T v_c)\cdot v_c$$

$$(\therefore) \begin{cases} \frac{\partial J}{\partial v_c} = (\sigma(u_0^T v_c) - 1)\cdot u_0 + \sum_{s=1}^{K} \sigma(u_{w_s}^T v_c)\cdot u_{w_s} \\ \frac{\partial J}{\partial u_0} = (\sigma(u_0^T v_c) - 1)\cdot v_c \\ \frac{\partial J}{\partial u_{w_s}} = \sigma(u_{w_s}^T v_c)\cdot u_{w_s} \end{cases}$$

(ii) Negative sampling only computes K terms,
which is much smaller than total vocaburary.

(f)
(i) $$\frac{\partial J_{skip}}{\partial U} = \sum_{-m\leq j\leq m,\ j\neq 0} \frac{\partial J(v_c, w_{t+j}, U)}{\partial U}$$

(ii) $$\frac{\partial J_{skip}}{\partial v_c} = \sum_{-m\leq j\leq m,\ j\neq 0} \frac{\partial J(v_c, w_{t+j}, U)}{\partial v_c}$$

(iii) $\frac{\partial J_{skip}}{\partial v_w}$ $(w\neq c) = 0$ (No term of v_w → Derivatives will be zero.)