1 Written

- (a) 根据定义可知：

$$y_w = \begin{cases} 1 & \text{if } w = o \\ 0 & \text{otherwise} \end{cases} \Rightarrow \text{LHS} = -\sum_{w \in \text{Vocab}} y_w \log(\hat{y}_w) = -y_o \log(\hat{y}_o) = -\log(\hat{y}_o) = \text{RHS} \tag{1.1}$$

- (b) 本问可参考[slides]中29-31页的推导过程：

$$\begin{aligned}
\frac{\partial J_{\text{naive-softmax}}(v_c, o, U)}{\partial v_c} &= -\frac{\partial}{\partial v_c} \log P(O = o | C = c) \\
&= -\frac{\partial}{\partial v_c} \log \frac{\exp(u_o^\top v_c)}{\sum_{w \in V} \exp(u_w^\top v_c)} \\
&= -\frac{\partial}{\partial v_c} \log \exp(u_o^\top v_c) + \frac{\partial}{\partial v_c} \log \left(\sum_{w \in V} \exp(u_w^\top v_c) \right) \\
&= -\frac{\partial}{\partial v_c} u_o^\top v_c + \frac{1}{\sum_{w \in V} \exp(u_w^\top v_c)} \cdot \frac{\partial}{\partial v_c} \sum_{x \in V} \exp(u_x^\top v_c) \\
&= -u_o + \frac{1}{\sum_{w \in V} \exp(u_w^\top v_c)} \cdot \sum_{x \in V} \frac{\partial}{\partial v_c} \exp(u_x^\top v_c) \\
&= -u_o + \frac{1}{\sum_{w \in V} \exp(u_w^\top v_c)} \cdot \sum_{x \in V} \exp(u_x^\top v_c) \frac{\partial}{\partial v_c} u_x^\top v_c \\
&= -u_o + \frac{1}{\sum_{w \in V} \exp(u_w^\top v_c)} \sum_{x \in V} \exp(u_x^\top v_c) u_x \\
&= -u_o + \sum_{x \in V} \frac{\exp(u_x^\top v_c)}{\sum_{w \in V} \exp(u_w^\top v_c)} u_x \\
&= -u_o + \sum_{x \in V} P(O = x | C = c) u_x \\
&= -u_o + \sum_{x \in V} \hat{y}_x u_x \\
&= -U^\top y + U^\top \hat{y} \\
&= U^\top (\hat{y} - y)
\end{aligned} \tag{1.2}$$

式(1.2)沿用[notes]中的标记，即约定$U \in \mathbb{R}^{|V| \times n}, y \in \mathbb{R}^{|V|}, \hat{y} \in \mathbb{R}^{|V|}$，具体如下：

$$\begin{aligned}
U &= \begin{bmatrix} u_0 & u_1 & \dots & u_{|V|} \end{bmatrix} \\
y &= \begin{bmatrix} 0 & 0 & \dots & 1 & \dots & 0 \end{bmatrix}^\top \\
\hat{y} &= \begin{bmatrix} \hat{y}_0 & \hat{y}_1 & \dots & \hat{y}_o & \dots & \hat{y}_{|V|} \end{bmatrix}^\top
\end{aligned} \tag{1.3}$$

  - (1) 当$\hat{y} = y$时，梯度为零；
  - (2) 我理解这个问题可能是想说梯度值实际刻画的是观测值与真实值之间误差，因此减去这个误差可以地得到更可靠的v_c

- (c) 类似(b)中的推导，我们有：

$$
\begin{aligned}
\frac{\partial J_{\text{naive-softmax}}(v_c, o, U)}{\partial u_w} &= -\frac{\partial}{\partial u_w}\log P(O=o|C=c) \\
&= -\frac{\partial}{\partial u_w}\log\frac{\exp(u_o^\top v_c)}{\sum_{w\in V}\exp(u_w^\top v_c)} \\
&= -\frac{\partial}{\partial u_w}\log\exp(u_o^\top v_c) + \frac{\partial}{\partial u_w}\log\left(\sum_{w\in V}\exp(u_w^\top v_c)\right) \\
&= -\frac{\partial}{\partial u_w}u_o^\top v_c + \frac{1}{\sum_{w\in V}\exp(u_w^\top v_c)}\cdot\frac{\partial}{\partial u_w}\sum_{x\in V}\exp(u_x^\top v_c) \\
&= -\frac{\partial}{\partial u_w}u_o^\top v_c + \frac{1}{\sum_{w\in V}\exp(u_w^\top v_c)}\frac{\partial}{\partial u_w}\exp(u_w^\top v_c) \\
&= -\frac{\partial}{\partial u_w}u_o^\top v_c + \frac{\exp(u_w^\top v_c)}{\sum_{w\in V}\exp(u_w^\top v_c)}v_c \\
&= -\frac{\partial}{\partial u_w}u_o^\top v_c + P(O=w|C=c)v_c \\
&= -\frac{\partial}{\partial u_w}u_o^\top v_c + \hat{y}_w v_c \\
&= \begin{cases}(\hat{y}_w - 1)v_c & \text{if } w = o \\ \hat{y}_w v_c & \text{otherwise}\end{cases} \\
&= (\hat{y}_w - y_w)v_c
\end{aligned} \tag{1.4}
$$

- (d) 根据(c)中的结果，可得：

$$
\frac{\partial J_{\text{naive-softmax}}(v_c, o, U)}{\partial U} = (\hat{y} - y)^\top v_c \tag{1.5}
$$

- (e) 将ReLU激活函数写作分段形式分别求导：

$$
f(x) = \begin{cases}0 & \text{if } x < 0 \\ x & \text{if } x > 0\end{cases} \Rightarrow f'(x) = \begin{cases}0 & \text{if } x < 0 \\ 1 & \text{if } x > 0\end{cases} \tag{1.6}
$$

- (f) 有如下推导：

$$
\sigma'(x) = \frac{e^x(e^x+1) - e^{2x}}{(e^x+1)^2} = \frac{e^x}{(e^x+1)^2} = \sigma(x)(1-\sigma(x)) \tag{1.7}
$$

- (g) 关于负采样的内容可参考[notes]与[slides]的相关章节。
  - (1) 关于v_c的偏导有如下推导：

$$
\begin{aligned}
\frac{\partial J_{\text{neg-sample}}(v_c, o, U)}{\partial v_c} &= -\frac{\partial}{\partial v_c}\log\sigma(u_o^\top v_c) - \sum_{s=1}^{K}\frac{\partial}{\partial v_c}\log\sigma(-u_{w_s}^\top v_c) \\
&= -\frac{\sigma(u_o^\top v_c)(1-\sigma(u_o^\top v_c))}{\sigma(u_o^\top v_c)}\cdot\frac{\partial u_o^\top v_c}{\partial v_c} - \sum_{s=1}^{K}\frac{\sigma(-u_{w_s}^\top v_c)(1-\sigma(-u_{w_s}^\top v_c))}{\sigma(-u_{w_s}^\top v_c)}\frac{\partial - u_{w_s}^\top v_c}{\partial v_c} \\
&= (\sigma(u_o^\top v_c) - 1)u_o + \sum_{s=1}^{K}(1-\sigma(-u_{w_s}^\top v_c))u_{w_s}
\end{aligned} \tag{1.8}
$$

    关于u_o的偏导有如下推导：

$$
\begin{aligned}
\frac{\partial J_{\text{neg-sample}}(v_c, o, U)}{\partial u_o} &= -\frac{\partial}{\partial u_o}\log\sigma(u_o^\top v_c) - \sum_{s=1}^{K}\frac{\partial}{\partial u_o}\log\sigma(-u_{w_s}^\top v_c) \\
&= -\frac{\sigma(u_o^\top v_c)(1-\sigma(u_o^\top v_c))}{\sigma(u_o^\top v_c)}\cdot\frac{\partial u_o^\top v_c}{\partial u_o} \\
&= (\sigma(u_o^\top v_c) - 1)v_c
\end{aligned} \tag{1.9}
$$

    关于u_{w_s}的偏导有如下推导：

$$\begin{aligned}\frac{\partial J_{\text{neg-sample}}(v_c,o,U)}{\partial u_{w_s}} &= -\frac{\partial}{\partial u_{w_s}}\log\sigma(u_o^\top v_c)-\sum_{k=1}^{K}\frac{\partial}{\partial u_{w_s}}\log\sigma(-u_{w_k}^\top v_c)\\ &= -\sum_{k=1}^{K}\frac{\sigma(-u_{w_k}^\top v_c)(1-\sigma(-u_{w_k}^\top v_c))}{\sigma(-u_{w_k}^\top v_c)}\frac{\partial -u_{w_k}^\top v_c}{\partial u_{w_s}}\\ &= (1-\sigma(-u_{w_s}^\top v_c))v_c\end{aligned} \tag{1.10}$$

- (2) 观察$(1.8)(1.9)(1.10)$三个偏导解析式，显然可以重用的部分是：

$$\sigma(u_o^\top v/_c)-1 \text{ and } 1-\sigma(-u_{w_s}^\top v_c), s=1,2,\dots,K \tag{1.11}$$

  写成要求的矩阵形式即为：

$$\sigma(U_{o,\{w_1,\dots,w_K\}}^\top v_c)-\mathbf{1} \tag{1.12}$$

- (3) 从$(b)(c)$的结果来看，同样有可以重用的部分（$\hat{y}-y$），但是(b)的梯度需要计算矩阵与向量的乘法，耗时较长，而(g)中的三个结果使用重用部分后本质上是标量与向量相乘的运算，自然要高效很多。

- (h) 本问与(g)的区别在于，负采样可能采样到重复的单词，因此结果与式(1.10)稍有区别：

$$\begin{aligned}\frac{\partial J_{\text{neg-sample}}(v_c,o,U)}{\partial u_{w_s}} &= -\frac{\partial}{\partial u_{w_s}}\log\sigma(u_o^\top v_c)-\sum_{k=1}^{K}\frac{\partial}{\partial u_{w_s}}\log\sigma(-u_{w_k}^\top v_c)\\ &= -\sum_{k=1}^{K}\frac{\sigma(-u_{w_k}^\top v_c)(1-\sigma(-u_{w_k}^\top v_c))}{\sigma(-u_{w_k}^\top v_c)}\frac{\partial -u_{w_k}^\top v_c}{\partial u_{w_s}}\\ &= \sum_{k=1}^{K}\mathbf{1}_{w_k=w_s}(1-\sigma(-u_{w_s}^\top v_c))v_c\end{aligned} \tag{1.13}$$

  其中**1**为指示函数，若$w_k=w_s$取值为一，否则取值为零。

- (i) 在[notes]中可以查阅skip-gram模型的目标函数：

$$\text{minimize}\quad J=-\sum_{j=0,j\neq m}^{2m}u_{c-m+j}^\top v_c+2m\log\sum_{k=1}^{|V|}\exp(u_k^\top v_c) \tag{1.14}$$

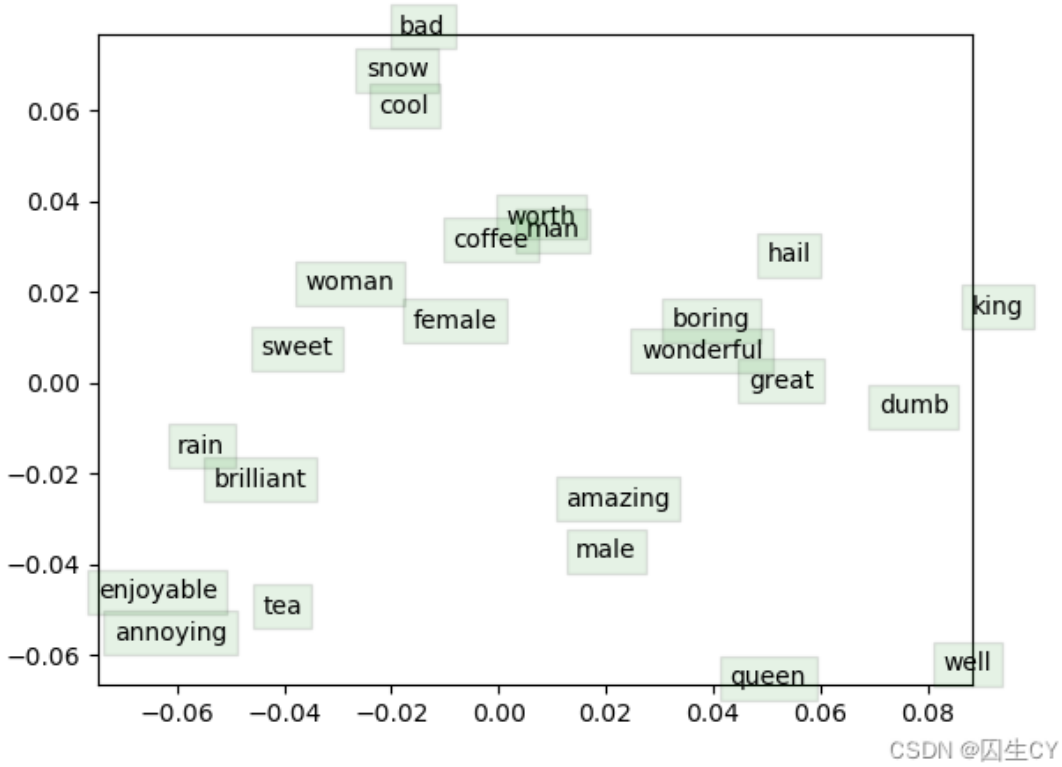
  则要求的三个偏导具有如下形式：

$$\begin{aligned}\frac{\partial J_{\text{skip-gram}}(v_c,w_{t-m},\dots,w_{t+m},U)}{\partial U} &= \sum_{-m\le j\le m,j\neq 0}\frac{\partial J(v_c,w_{t+j},U)}{\partial U}\\ \frac{\partial J_{\text{skip-gram}}(v_c,w_{t-m},\dots,w_{t+m},U)}{\partial v_c} &= \sum_{-m\le j\le m,j\neq 0}\frac{\partial J(v_c,w_{t+j},U)}{\partial v_c}\\ \frac{\partial J_{\text{skip-gram}}(v_c,w_{t-m},\dots,w_{t+m},U)}{\partial v_w} &= 0\end{aligned}$$

-

2 Coding

- (a) 这里第(4)小问可能有点问题，始终无法通过测试，但是在往年的代码里是可以通过测试的，感觉或许是测试代码写得有些问题，因为只有skip-gram的负采样损失函数这一项无法通过测试，其他都是正确的。
  - (1) 参考written部分的(f)
  - (2) 参考written部分的$(b)(c)(d)$
  - (3) 参考written部分的(g)
  - (4) 参考written部分的(i)
- (b) 简单的梯度下降法实现。
- (c) 需要事先下载数据集并解压到utils目录下，运行得到的图片为：


bad
snow
cool
worth
man
coffee
hail
woman
king
female
boring
sweet
wonderful
great
dumb
rain
brilliant
amazing
male
enjoyable
tea
annoying
queen
well
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08