2 Understanding word2vec

2.a

$$\sigma(\mathbf{x}+c)_j = \frac{e^{z_j+c}}{\Sigma_{k=1}^{K} e^{z_k+c}} = \frac{e^{c}e^{z_j}}{e^{c}\Sigma_{k=1}^{K} e^{z_k}} = \frac{e^{z_j}}{\Sigma_{k=1}^{K} e^{z_k}} = \sigma(\mathbf{x})_j$$

$$Quod.Erat.Demonstrandum$$

2.b

$$-\sum_{w\in Vocab} y_w \log(\hat{y}_w) =$$

$$-y_o \log(\hat{y}_o) - \sum_{w\in Vocab, w\neq o} y_w \log(\hat{y}_w) =$$

$$-\log(\hat{y}_o)$$

2.c

$$\frac{\partial J\left(v_c, o, U\right)}{\partial v_c} =$$

$$-\frac{\partial\left(u_o^T v_c\right)}{\partial v_c} + \frac{\partial\left(\log\left(\sum_w \exp\left(u_w^T v_c\right)\right)\right)}{\partial v_c} =$$

$$-u_o + \frac{1}{\sum_w \exp\left(u_w^T v_c\right)} \frac{\partial\left(\sum_w \exp\left(u_w^T v_c\right)\right)}{\partial v_c} =$$

$$-u_o + \sum_w \frac{\exp\left(u_w^T v_c\right) u_w}{\sum_w \exp\left(u_w^T v_c\right)} =$$

$$-u_o + \sum_w p(O=w|C=c) u_w =$$

$$-u_o + \sum_w \hat{y_w} u_w =$$

$$U(\hat{y} - y)$$

2.d

$$\frac{\partial J\left(v_c, o, U\right)}{\partial u_w} =$$

$$-\frac{\partial\left(u_o^T v_c\right)}{\partial u_w} + \frac{\partial\left(\log\left(\sum_w \exp\left(u_w^T v_c\right)\right)\right)}{\partial u_w}$$

$Either:$

$w = 0:$

$$\frac{\partial J\left(v_c, o, U\right)}{\partial u_w} =$$

$$-v_c + p(O = o|C = c)v_c =$$

$$\hat{y}_w v_c - v_c =$$

$$(\hat{y}_w - 1)v_c$$

$w \neq 0:$

$$\frac{\partial J\left(v_c, o, U\right)}{\partial u_w} =$$

$$0 + p(O = w|C = c)v_c =$$

$$\hat{y}_w v_c$$

$Eventually:$

$$\frac{\partial J\left(v_c, o, U\right)}{\partial U} = v_c(\hat{y} - y)^T$$

2.e

$$\frac{\partial \sigma(x_i)}{\partial x_i} = \frac{1}{(1 + \exp(-x_i))^2}\exp(-x_i) = \sigma(x_i)(1 - \sigma(x_i))$$

$$\frac{\partial \sigma(x)}{\partial x} = \left[\frac{\partial \sigma\left(x_j\right)}{\partial x_i}\right]_{d\times d} = \begin{bmatrix} \sigma'\left(x_1\right) & 0 & \cdots & 0 \\ 0 & \sigma'\left(x_2\right) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma'\left(x_d\right) \end{bmatrix} = \mathrm{diag}(\sigma'(x))$$

2.f

$For\, v_c :$

$$\frac{\partial J_{\text{neg-sample}}}{\partial v_c} =$$

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

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

$For\ u_o : o \notin [w_i]_{i=1}^{K}$

$$\frac{\partial J_{\text{neg-sample}}}{\partial u_o} = (\sigma(u_o^T v_c) - 1)v_c$$

$For\ u_k : k = [1, K] \in \mathbb{N}$

$$\frac{\partial J}{\partial \boldsymbol{u}_k} = -\left(\sigma\left(-\boldsymbol{u}_k^\top \boldsymbol{v}_c\right) - 1\right) \boldsymbol{v}_c = \sigma\left(\boldsymbol{u}_k^\top \boldsymbol{v}_c\right) \boldsymbol{v}_c,$$

$For\ naive\ softmax\ loss\ function :$

$$\frac{\partial J\left(v_c, o, U\right)}{\partial U} = v_c(\hat{y} - y)^T$$

$$\frac{\partial J\left(v_c, o, U\right)}{\partial v_c} = U(\hat{y} - y)$$

$For\ negative\ sampling\ loss\ function :$

$$\frac{\partial J}{\partial \boldsymbol{v}_c} =$$

$$\left(\sigma\left(\boldsymbol{u}_o^\top v_c\right)-1\right)\boldsymbol{u}_o+\sum_{k=1}^{K}\sigma\left(\boldsymbol{u}_k^\top \boldsymbol{v}_c\right)\boldsymbol{u}_k=$$

$$\sigma\left(-\boldsymbol{u}_o^\top v_c\right)\boldsymbol{u}_o+\sum_{k=1}^{K}\sigma\left(\boldsymbol{u}_k^\top \boldsymbol{v}_c\right)\boldsymbol{u}_k$$

$$\frac{\partial J}{\partial \boldsymbol{u}_k}=\sigma\left(\boldsymbol{u}_k^\top \boldsymbol{v}_c\right)\boldsymbol{v}_c : k=[1,K]\in\mathbb{N}$$

$$\frac{\partial J}{\partial \boldsymbol{u}_o}=\left(\sigma\left(\boldsymbol{u}_o^\top \boldsymbol{v}_c\right)-1\right)\boldsymbol{v}_c=\sigma\left(-\boldsymbol{u}_o^\top \boldsymbol{v}_c\right)\boldsymbol{v}_c$$

With this loss function the computation is not related with going over all of the words in vocabulary, V, rather than in the number of samples, K.

2.g.i

$$\frac{\partial J_{skip-gram}(v_c, w_{t-m}, \ldots, w_{t+m}, U)}{\partial U}=$$

$$\sum_{-m\leq j\leq m, j\neq 0}\frac{\partial J\left(v_c, w_{t+j}, U\right)}{\partial U}$$

2.g.ii

$$\frac{\partial J_{skip-gram}(v_c, w_{t-m}, \ldots, w_{t+m}, U)}{\partial v_c}=$$

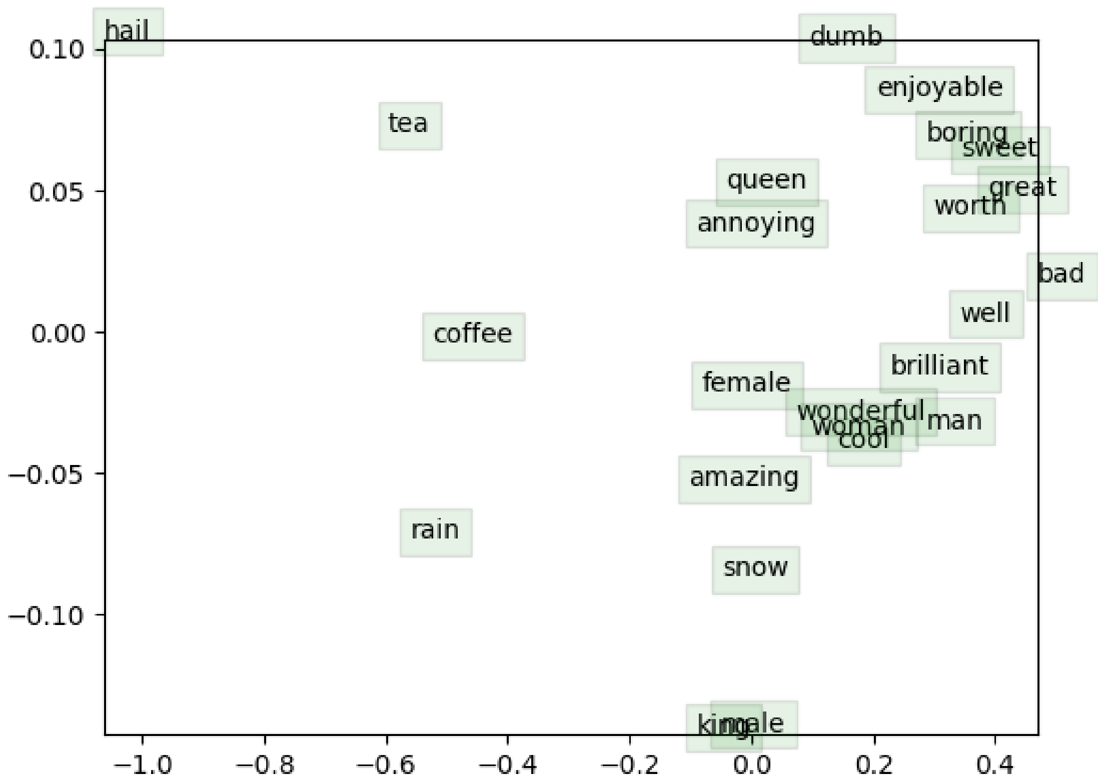
$$\sum_{-m\leq j\leq m, j\neq 0}\frac{\partial J\left(v_c, w_{t+j}, U\right)}{\partial v_c}$$

2.g.iii

$$\frac{\partial J_{skip-gram}(v_c, w_{t-m}, \ldots, w_{t+m}, U)}{\partial v_w} = 0$$

3 Implementing word2vec

3.e



We observe the words {sweet, great} clustered together, and it's reasonable that they can appear in proximity.

The words {coffee,tea} are not clustered together, however this is surprising because they in many cases used interchangeably, or along with each other.

However the clustering didn't catch that the words {man, male} can have similar meanings, as they are afar from each other.