Subject

Date

$J_{naive\text{-}softmax}(v_c, o, U) \rightarrow J_{ns}(v_c, o, U)$ (1b)

$$\frac{\partial J_{ns}(v_c,o,U)}{\partial v_c} = -\frac{\partial \log P(O=o \mid C=c)}{\partial v_c} = -\frac{\partial}{\partial v_c}\log\frac{e^{u_o^T v_c}}{\sum_{w\in V} e^{u_w^T v_c}}$$

$$= -\frac{\partial}{\partial v_c}\Big[\underbrace{\log e^{u_o^T v_c}}_{u_o^T v_c} - \log\sum_{w\in V} e^{u_w^T v_c}\Big] \quad ☆ \quad = -\Big(u_o - \sum_{w\in V} \underbrace{\frac{u_w e^{u_w^T v_c}}{\sum_{w\in V} e^{u_w^T v_c}}}_{P(O=o\mid C=c)u_w}\Big) \quad ☆$$

$$= -u_o + \sum_{w\in V} u_w \underbrace{P(O=w \mid C=c)}_{\text{predicted distro } \hat{y}} = \underbrace{\sum_{w\in V} u_w \hat{y}}_{U\hat{y}} - \underbrace{u_o}_{Uy} = U(\hat{y}-y)$$

---

show that: $-\sum_{w\in V} y_w \log(\hat{y}_w) = -\log(\hat{y}_o)$ (1a)

$$-\sum_{w\in V} y_w \log(\hat{y}_w) = -\Big(y_o \log(\hat{y}_o) + \overbrace{\sum_{w\in V, w\neq o} y_w \log(\hat{y}_w)}^{\text{zero}}\Big) = -y_o \log(\hat{y}_o)$$

$\forall w\in V \quad \frac{\partial J_{ns}(v_c,o,U)}{\partial u_w} \longrightarrow$ from ☆ $\rightarrow \forall w\in V \quad \frac{-\partial}{\partial u_w}\Big(u_o^T v_c - \log\sum_{w\in V} e^{u_w^T v_c}\Big)$

$w \neq o \longrightarrow \text{zero} + \overset{v_c}{P}(O=w \mid C=c) = v_c \hat{y}_w$ from ☆

$w = o \longrightarrow -v_c + v_c P(O=w \mid C=c) = -v_c + v_c \hat{y}_w = v_c(-1+\hat{y}_w)$

$$\frac{\partial J_{ns}(v_c,o,U)}{\partial U} = v_c(\hat{y}-y)^T$$ (1d) (1c)

$\sigma(x) = \frac{1}{1+e^{-x}} = (1+e^{-x})^{-1} \rightarrow$ ~~$\sigma(x) = ...^{-1}, ... = 1+e^{-x}$~~ (1e)

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

$$= \frac{1}{1+e^{-x}}\,\frac{e^{-x}}{1+e^{-x}} = \sigma(x)\frac{e^{-x}+(1-1)}{1+e^{-x}} = \sigma(x)\Big[\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\Big] =$$

$$\sigma(x)(1-\sigma(x))$$

Scanned by CamScanner

Subject

Date

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

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

۱۴

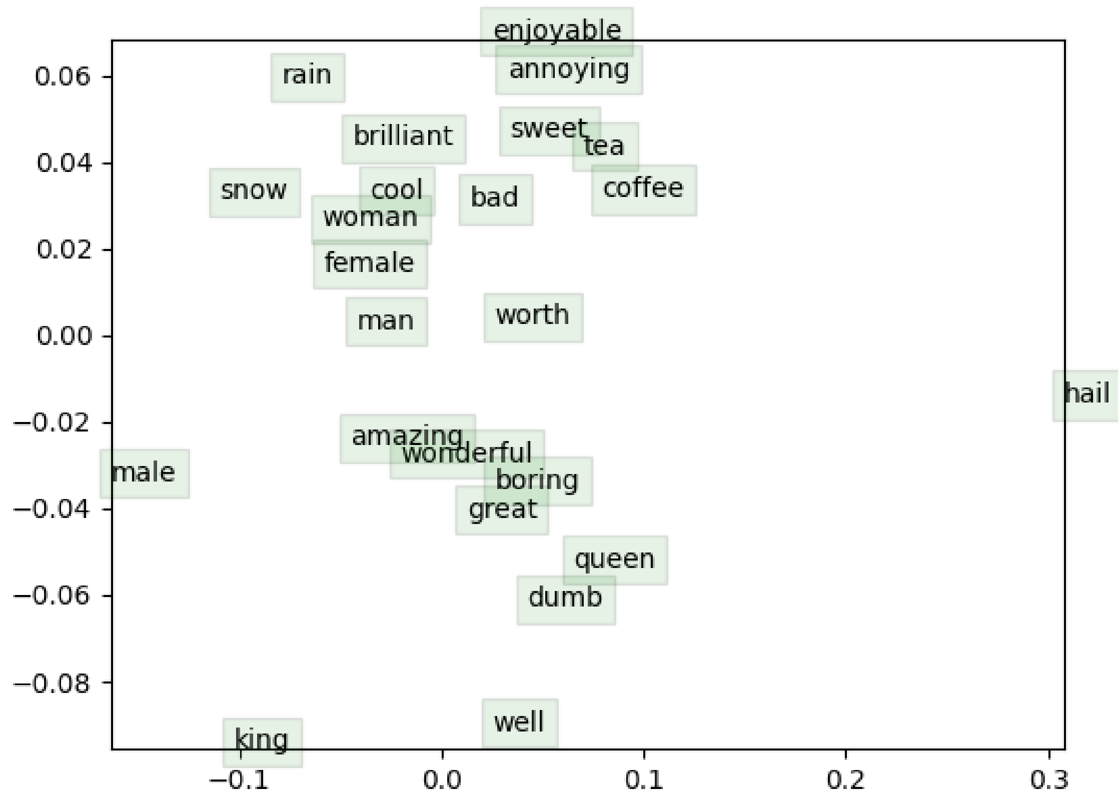
$$\frac{\partial J_{neg}}{\partial u_o} = v_c\left(\sigma(u_o^T v_c) - 1\right)$$

$$\frac{\partial J_{neg}}{\partial u_k} = v_c\left(1 - \sigma(-u_k^T v_c)\right) = v_c\, \sigma(u_k^T v_c)$$

در softmax پیچیدگی محاسبات با سایز V (تعداد کلمات) رابطه دارد، اما در neg این رابطه برقرار نیست و می‌توان الگوریتم را در زمان معقول روی دایره لغات گسترده‌تر اجرا کرد.

PaPCO

Scanned by CamScanner



In this plot, we can see that similar words are mapped near together. (Their vectors are similar)

Woman, female, man

Enjoyable, annoying

Sweet, tea, coffee

Amazing, wonderful, boring, great

We can see, the words in each of these groups are having similar contexts.

Also this is not perfect, since for example king and queen are far from each other, or male is not near man.