word2vec gradients Charlie Colley 9/29/17

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

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

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

Thus $\frac{\partial \sigma(x)}{\partial x}(x) = \sigma(x)(1-\sigma(x))$.

Objective function:
$\arg\max_{V_c, V_w} = \Sigma_{(w,c)\in D} \log(\sigma(\langle v_c, v_w \rangle)) + \Sigma_{(w,c)\in D'} \log(\sigma(-\langle v_c, v_w \rangle))$

$$\nabla_{v_c}(\langle v_c, v_w \rangle)_k = \frac{\partial}{\partial (v_c)_k} \sum_i (v_c)_i (v_w)_i = \frac{\partial}{\partial (v_c)_k} (\sum_{i \neq k} (v_c)_i (v_w)_i + (v_c)_k (v_w)_k) = (v_w)_k$$

by symmetry, the gradient $\nabla_{v_w} \langle v_c, v_w \rangle$ is v_c. Let $B \in \mathbb{R}^{n \times m}$.

$$\nabla_B(\langle v_c, v_w \rangle_B)_{kl} = \frac{\partial}{\partial B_{kl}} \sum_{ij} B_{ij} (v_c)_i (v_w)_i =$$

$$\frac{\partial}{\partial B_{kl}} (\sum_{i,j \neq k} (v_c)_i (v_w)_i + B_{kl}(v_c)_k (v_w)_l) = (v_c)_l (v_w)_k = (v_c \otimes v_w)_{kl}$$

Then the gradient of $\log(\sigma(\langle v_c, v_w \rangle))$ with respect to v_c is

$$\frac{\partial}{\partial (v_c)_k}(\log(\sigma(\langle v_c, v_w \rangle))) = (\frac{1}{\sigma(\langle v_c, v_w \rangle)})(\sigma(\langle v_c, v_w \rangle)(1-\sigma(\langle v_c, v_w \rangle)))(v_w)_k$$

$$= (1-\sigma(\langle v_c, v_w \rangle)))(v_w)_k$$

again by symmetry we can show that the gradient with respect to v_w is similarly
$\nabla_{v_w}(log(\sigma(\langle v_c, v_w \rangle))) = (1-\sigma(\langle v_c, v_w \rangle))v_c$