word2Vec

刘绪成

YNU

December 1, 2020

Overview

Model of CBOW

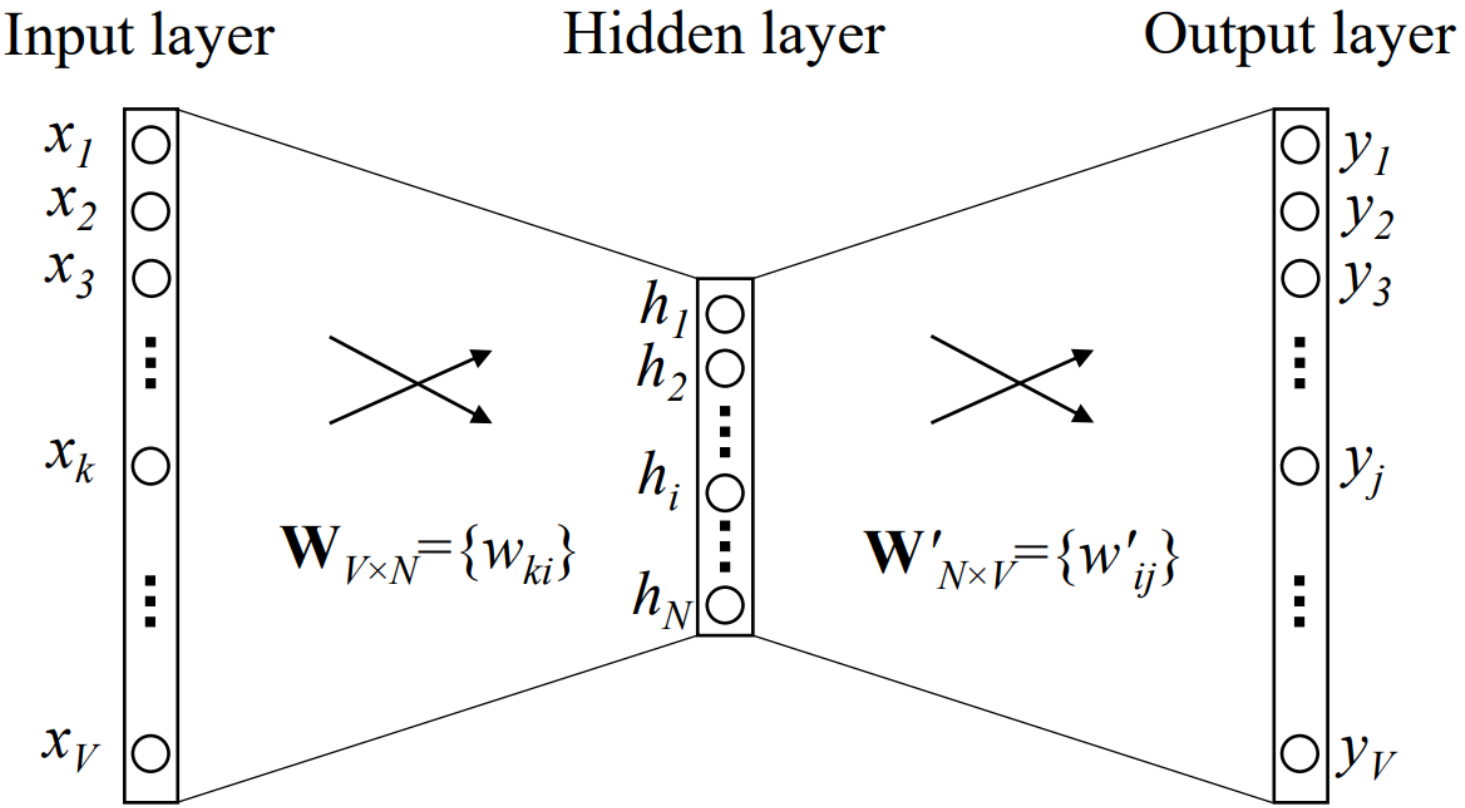


Figure 1: A simple CBOW model with only one word in the context

Derivation

随机初始化 W，由于输入向量 x 是 One-hot Vector ， 隐藏层 h 就是 W 的某一行。

$$\mathbf{h} = \mathbf{W}^T\mathbf{x} = \mathbf{W}^T_{(k,\cdot)} := \mathbf{v}^T_{w_I} \tag{1}$$

从隐藏层到输入出层：

$$u_j = \mathbf{v}'_{\mathbf{w_j}}{}^T\mathbf{h} \tag{2}$$

使用 softmax 求输入层第 j 个神经元的值

$$p(w_j \mid w_I) = y_j = \frac{\exp(u_j)}{\sum_{j'=1}^{V}\exp(u_{j'})} \tag{3}$$

将公式 (1) 和 (2) 代入

$$p(w_j \mid w_I) = \frac{\exp\left(\mathbf{v}'_{w_j}\mathbf{v}_{w_I}\right)}{\sum_{j'=1}^{V}\exp\left(\mathbf{v}'_{w_{j'}}\mathbf{v}_{w_I}\right)} \tag{4}$$

Derivation:hidden->output

反向传播从输出层到隐藏层：我们的训练目标是最大化公式 (4)，即在实际输出 W_O 的条件概率

$$\begin{aligned} \max p\left(w_O \mid w_I\right) &= \max y_{j^*} \\ &= \max \log y_{j^*} \\ &= u_{j^*} - \log \sum_{j'=1}^{V} \exp\left(u_{j'}\right) := -E \end{aligned} \tag{5}$$

$$\frac{\partial E}{\partial u_j} = y_j - t_j := e_j \tag{6}$$

$t_j = 1(j = j^*)$ 其他情况下 $t_j = 0$

Derivation:hidden->output

E 对 W^{prime} 第 j 列, 第 i 行的偏微分

$$\frac{\partial E}{\partial w'_{ij}} = \frac{\partial E}{\partial u_j} \cdot \frac{\partial u_j}{\partial w'_{ij}} = e_j \cdot h_i \tag{7}$$

使用随机梯度下降更新

$$w'^{(\text{new})}_{ij} = w'^{(\text{old})}_{ij} - \eta \cdot e_j \cdot h_i \tag{8}$$

Derivation:input->hidden

从隐藏层-> 输入层:

$$\frac{\partial E}{\partial h_i} = \sum_{j=1}^{V} \frac{\partial E}{\partial u_j} \cdot \frac{\partial u_j}{\partial h_i} = \sum_{j=1}^{V} e_j \cdot w'_{ij} := \mathrm{EH}_i \tag{9}$$

$$h_i = \sum_{k=1}^{V} x_k \cdot w_{ki} \tag{10}$$

$$\frac{\partial E}{\partial w_{ki}} = \frac{\partial E}{\partial h_i} \cdot \frac{\partial h_i}{\partial w_{ki}} = \mathrm{EH}_i \cdot x_k \tag{11}$$

$$\frac{\partial E}{\partial \mathbf{W}} = \mathbf{x} \otimes \mathrm{EH} = \mathbf{x}\mathrm{EH}^T \tag{12}$$

输入向量的更新:

$$\mathbf{v}_{w_I}^{(\mathrm{new})} = \mathbf{v}_{w_I}^{(\mathrm{old})} - \eta \mathrm{EH}^T \tag{13}$$

Model of CBOW

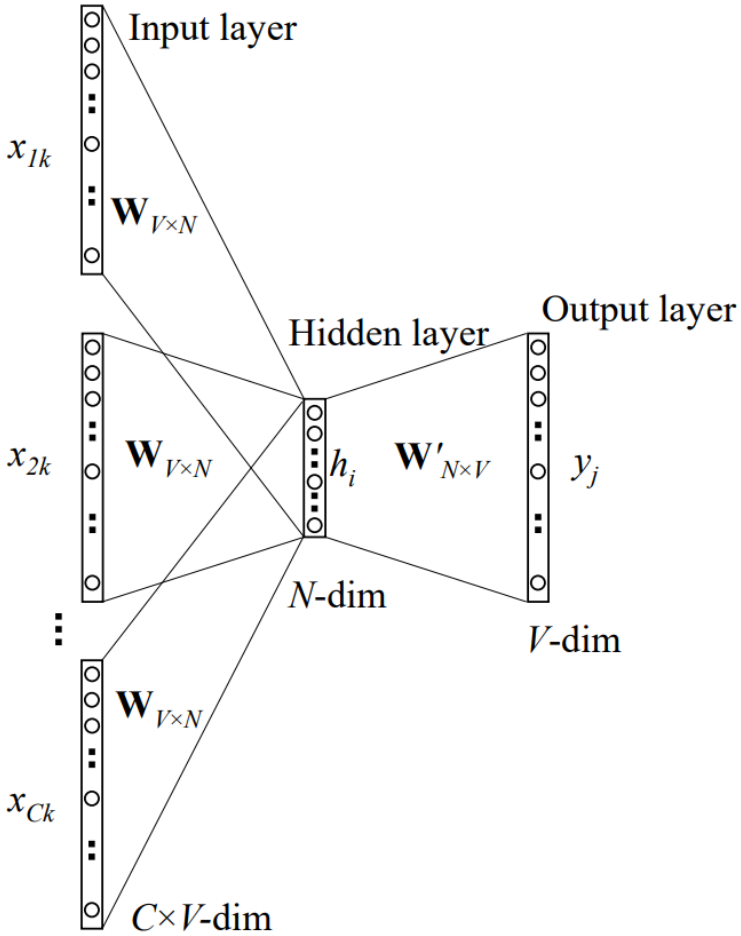


Figure 2: Continuous bag-of-word model

Points

用输入层-> 隐藏层的 W^T 与平均向量 x 的积作为输入

$$\begin{aligned} h &= \frac{1}{C} W^T (x_1 + x_2 + ... + x_C) \\ &= \frac{1}{C} (v_{\omega_1} + v_{\omega_2} + ... + v_{\omega_C}) \end{aligned} \tag{14}$$

$$\begin{aligned} E &= -\log p(\omega_O | \omega_{I,1}, ..., \omega_{I,C}) \\ &= -u_{j_*} + \log \sum_{j'=1}^{V} \exp(u_{j'}) \\ &= -{v'_{\omega_O}}^T \cdot h + \log \sum_{j'=1}^{V} \exp({\omega'_j}^T \cdot h) \end{aligned} \tag{15}$$

Figure

$$v'^{(new)}_{\omega_j} = v'^{(old)}_{\omega_j} - \eta \cdot e_j \cdot h \qquad for \quad j = 1, 2, ...V \tag{16}$$

$$v'^{(new)}_{\omega_{I,c}} = v'^{(old)}_{\omega_{I,c}} - \frac{1}{C} \cdot \eta \cdot EH^T \qquad for \quad j = 1, 2, ...C \tag{17}$$

Model of Skip-gram

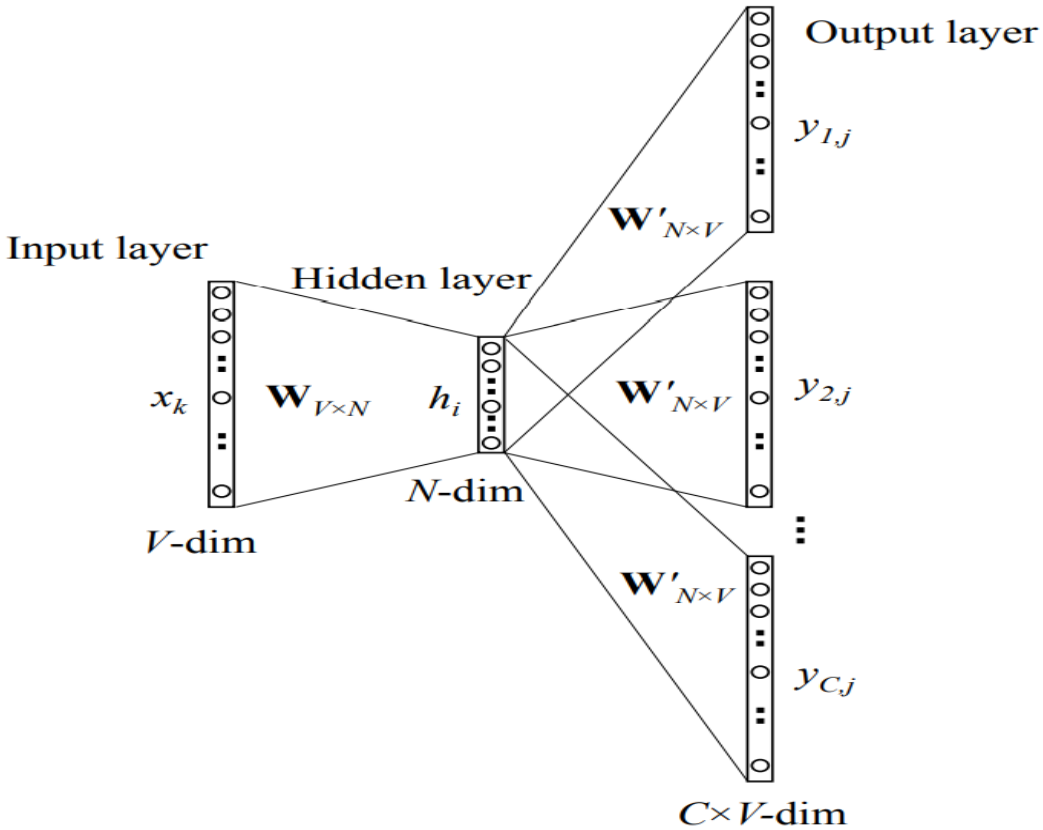


Figure 3: The skip-gram model.

Points

$$
\begin{aligned}
E &= -\log p\left(w_{O,1}, w_{O,2}, \cdots, w_{O,C} \mid w_I\right) \\
&= -\log \prod_{c=1}^{C} \frac{\exp\left(u_{c,j_c^*}\right)}{\sum_{j'=1}^{V} \exp\left(u_{j'}\right)} \\
&= -\sum_{c=1}^{C} u_{j_c^*} + C \cdot \log \sum_{j'=1}^{V} \exp\left(u_{j'}\right)
\end{aligned} \tag{18}
$$

$$
\frac{\partial E}{\partial u_{c,j}} = y_{c,j} - t_{c,j} := e_{c,j} \tag{19}
$$

$$
EI_j = \sum_{c=1}^{C} e_{c,j} \tag{20}
$$

$$
\frac{\partial E}{\partial w'_{i,j}} = \sum_{c=1}^{C} \frac{\partial E}{\partial u_{c,j}} \cdot \frac{\partial u_{c,j}}{w'_{i,j}} = EI_j \cdot h_i \tag{21}
$$

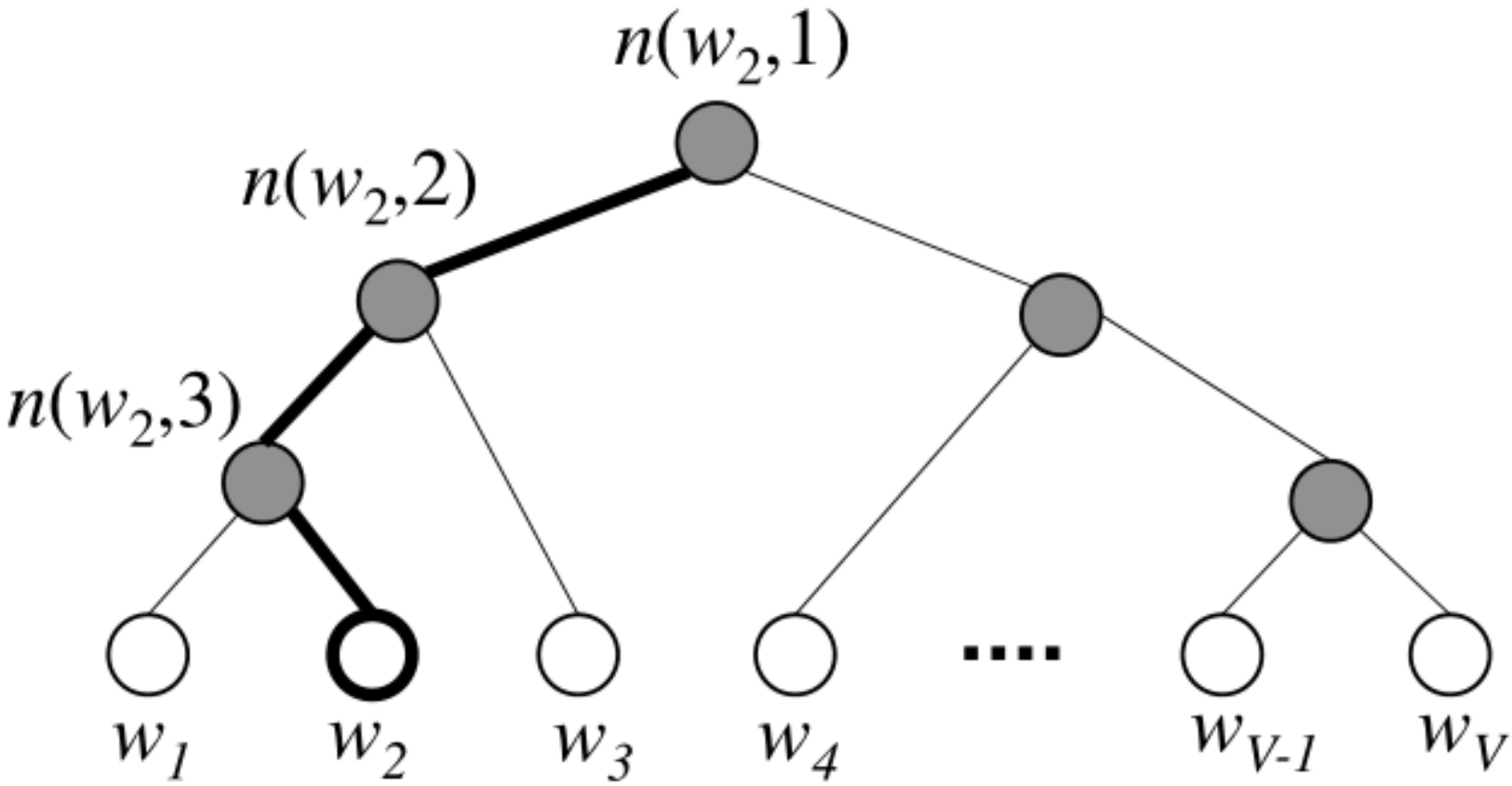
Points

$$v'^{(new)}_{w_j} = v'^{(old)}_{w_j} - \eta \cdot EI_j \cdot h \qquad for\ j = 1, 2, 3..., V. \tag{22}$$

$$V'^{(new)}_{w_j} = V'^{(old)}_{w_j} - \eta \cdot EI^T \tag{23}$$

$$EH_i = \sum_{j=1}^{V} EI_j \cdot w'_{ij} \tag{24}$$

Hierarchical Softmax



Derivation

$$p(w = w_O) = \prod_{j=1}^{L(w)-1} \sigma([\![n(w, j+1) = ch(n(w, j))]\!] \cdot {v'_{n(w,j)}}^T \cdot h) \tag{25}$$

$$[\![x]\!] = \begin{cases} 1 & if\ x\ is\ true \\ -1 & otherwise \end{cases} \tag{26}$$

$$p(n, left) = \sigma({v'_n}^T \cdot h) \tag{27}$$

$$p(n, right) = 1 - \sigma({v'_n}^T \cdot h) = \sigma(-{v'_n}^T \cdot h) \tag{28}$$

$$\begin{aligned} p(w_2 = w_O) &= p(n(w_2, 1), left) \cdot (n(w_2, 2), left) \cdot (n(w_2, 3), left) \\ &= \sigma({v'_n(w_2, 1)}^T h) \cdot \sigma({v'_n(w_2, 2)}^T h) \cdot \sigma({v'_n(w_2, 3)}^T h) \end{aligned} \tag{29}$$

$$\sum_{i=1}^{V} p(w_i = w_O) = 1 \tag{30}$$

Derivation

$$[\![\cdot]\!] := [\![n(w, j+1) = ch(n(w, j))]\!] \tag{31}$$

$$v'_j := v'_{n(w,j)} \tag{32}$$

$$E = -\log p(w = w_O | w_I) = -\sum_{j=1}^{L(w)-1} \log \sigma([\![\cdot]\!] v'^T_j h) \tag{33}$$

$$\begin{aligned}
\frac{\partial E}{\partial \mathbf{v}'_j \mathbf{h}} &= \left(\sigma\left([\![\cdot]\!] \mathbf{v}'^T_j \mathbf{h}\right) - 1\right)[\![\cdot]\!] \\
&= \begin{cases} \sigma\left(\mathbf{v}'^T_j \mathbf{h}\right) - 1 & ([\![\cdot]\!] = 1) \\ \sigma\left(\mathbf{v}'^T_j \mathbf{h}\right) & ([\![\cdot]\!] = -1) \end{cases} \\
&= \sigma\left(\mathbf{v}'^T_j \mathbf{h}\right) - t_j
\end{aligned} \tag{34}$$

$$\frac{\partial E}{\partial \mathbf{v}'_j} = \frac{\partial E}{\partial \mathbf{v}'_j \mathbf{h}} \cdot \frac{\partial \mathbf{v}'_j \mathbf{h}}{\partial \mathbf{v}'_j} = \left(\sigma\left(\mathbf{v}'^T_j \mathbf{h}\right) - t_j\right) \cdot \mathbf{h} \tag{35}$$

Derivation

$$\mathbf{v}_j'^{(\text{ new })} = \mathbf{v}_j'^{(\text{ old })} - \eta \left(\sigma\left(\mathbf{v}_j'^{T}\mathbf{h}\right) - t_j\right) \cdot \mathbf{h} \tag{36}$$

$$\begin{aligned} \frac{\partial E}{\partial \mathbf{h}} &= \sum_{j=1}^{L(w)-1} \frac{\partial E}{\partial \mathbf{v}_j'\mathbf{h}} \cdot \frac{\partial \mathbf{v}_j'\mathbf{h}}{\partial \mathbf{h}} \\ &= \sum_{j=1}^{L(w)-1} \left(\sigma\left(\mathbf{v}_j'^{T}\mathbf{h}\right) - t_j\right) \cdot \mathbf{v}_j' := \quad \mathrm{EH} \end{aligned} \tag{37}$$

Negative Sampling

$$E = -\log \sigma({v'_{w_O}}^T \mathbf{h}) - \sum_{w_j \in \mathcal{W}_{neg}} \log(-{v'_{w_j}}^T \mathbf{h}) \tag{38}$$

$$\begin{aligned} \frac{\partial E}{\partial \mathbf{v}'^T_{w_j} \mathbf{h}} &= \begin{cases} \sigma\left(\mathbf{v}'_{w_j} \mathbf{h}\right) - 1 & \text{if } w_j = w_O \\ \sigma\left(\mathbf{v}'_{w_j} \mathbf{h}\right) & \text{if } w_j \in \mathcal{W}_{\text{neg}} \end{cases} \\ &= \sigma\left(\mathbf{v}'_{w_j} \mathbf{h}\right) - t_j \end{aligned} \tag{39}$$

$$\frac{\partial E}{\partial \mathbf{v}'_{w_j}} = \frac{\partial E}{\partial \mathbf{v}'_{w_j} \mathbf{h}} \cdot \frac{\partial {\mathbf{v}'_{w_j}}^T \mathbf{h}}{\partial \mathbf{v}'_{w_j}} = \left(\sigma\left(\mathbf{v}'_{w_j} \mathbf{h}\right) - t_j\right) \mathbf{h} \tag{40}$$

$${\mathbf{v}'_{w_j}}^{(new)} = {\mathbf{v}'_{w_j}}^{(old)} - \eta \left(\sigma\left(\mathbf{v}'_{w_j} \mathbf{h}\right) - t_j\right) \mathbf{h} \tag{41}$$

Negative Sampling

$$
\begin{aligned}
\frac{\partial E}{\partial \mathbf{h}} &= \sum_{w_j \in w_O \cup \mathcal{W}_{\text{neg}}} \frac{\partial E}{\partial {\mathbf{v}'_{w_j}}^T \mathbf{h}} \cdot \frac{\partial {\mathbf{v}'_{w_j}}^T \mathbf{h}}{\partial \mathbf{h}} \\
&= \sum_{w_j \in w_O \cup \mathcal{W}_{\text{neg}}} \left(\sigma\left(\mathbf{v}'_{w_j} \mathbf{h}\right) - t_j\right) \mathbf{v}'_{w_j} \\
&:= \mathrm{EH}
\end{aligned}
\tag{42}
$$

References

Xin Rong (2016)

word2vec Parameter Learning Explained

arXiv e-prints arXiv:1411.2738.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a)

Efficient estimation of word representations in vector space

arXiv preprint arXiv:1301.3781.

Le, Quoc and Mikolov, Tomas (2014)

Distributed Representations of Sentences and Documents

Proceedings of the 31st International Conference on International Conference on Machine Learning II–1188–II–1196.

Example

python	科比	云南
perl,0.880822062492370 6	杜兰特,0.73625	贵州,0.72855
lua,0.856161892414093	布莱恩特,0.72571	云南省,0.70297
smalltalk,0.845733880996704 1	奥尼尔,0.72011	昆明,0.69951
javascript,0.840437710285186 8	罗德曼,0.71371	西双版纳,0.67357
脚本语言,0.839746654033660 9	欧尼尔,0.70804	腾冲,0.66212
僵尸	特朗普	开心
丧尸,0.631321668624877 9	川普,0.78675	放暑假,0.53373
狼人,0.550813555717468 3	奥巴马,0.77438	华之里,0.53014
吸血,0.540710330009460 4	尼克松,0.71462	欢乐,0.51938
幽灵,0.530379533767700 2	小布什,0.70919	欢笑,0.51552
奇鸭,0.499460101127624 5	希拉里,0.70734	天天,0.51292

Example

伦敦美国英国	king women man
伦敦之于英国，如美国之于	king 之于 man，如 women 之于
纽约,0.70653	narai,0.48565
纽约市,0.63568	queen,0.47201
芝加哥,0.57615	kings,0.46423
美国纽约,0.57061	scotland,0.46213
白天太阳晚上	武松红楼梦水浒传
白天之于晚上，如太阳之于	武松之于水浒传，如红楼梦之于
月亮,0.48442	林黛玉,0.72075
不西沉,0.467524	贾宝玉,0.70228
白昼,0.46403	王熙凤,0.69696
永昼,0.46176	大观园,0.68577

The End