Word2vec

2018年8月1日 12:44

- skip-gram & CBOW (Continuous Bag Of Words) 两个模型

  negative sampling & hierarchical softmax 两种训练方法

- skip-gram

  eg. the man hit his son

  中心词 生成 背景词

  时间窗口大小为2

(1) $$\max \prod_{t=1}^{T} \prod_{-m\le j\le m,\, j\ne 0} P(w^{(t+j)} \mid w^{(t)})$$

Maximum Likehood Estimation

给定中心词

生成背景词 $(w^{(t-2)}, w^{(t-1)}, w^{(t+1)}, w^{(t+2)})$

$$\min -\frac{1}{T} \sum_{t=1}^{T} \sum_{-m\le j\le m,\, j\ne 0} \log P(w^{(t+j)} \mid w^{(t)})$$

(2) $$P(w_o \mid w_c) = \underbrace{\frac{e^{\vec{u}_o^T \vec{v}_c}}{\sum_{i\in V} e^{\vec{u}_i^T \vec{v}_c}}}_{\text{softmax}}$$

- 词典 $V_{0,1,2,\dots,|V|-1}$

  设中心词索引为 c,

  背景词索引为 o,

  $\vec{v}_c$ 是中心词的向量表达式,

  $\vec{u}_o$ 是背景词的向量表达式

- 对于每个词 c, 它在 word2vec 模型中有两个向量表达:

  $\vec{v}_c$ c作为中心词

  $\vec{u}_c$ c作为背景词

(3) $$\frac{\partial \log P(w_o \mid w_c)}{\partial \vec{v}_c} = \vec{u}_o - \sum_{j\in V} \boxed{\frac{e^{\vec{u}_j^T \cdot \vec{v}_c}}{\sum_{i\in V} e^{\vec{u}_i^T \cdot \vec{v}_c}}} \vec{u}_j$$

证:

$$\log P(w_o \mid w_c) = \log e^{\vec{u}_o^T \cdot \vec{v}_c} - \log \sum_{i\in V} e^{\vec{u}_i^T \cdot \vec{v}_c}$$

$$= \vec{u}_o^T \cdot \vec{v}_c - \log \sum_{i\in V} e^{\vec{u}_i^T \cdot \vec{v}_c}$$

$$\frac{\partial \log P(w_o \mid w_c)}{\partial \vec{v}_c} = \vec{u}_o - \frac{\partial \log \sum_{i\in V} e^{\vec{u}_i^T \cdot \vec{v}_c}}{\partial \vec{v}_c}$$

$$= \vec{u}_o - \frac{1}{\sum_{i\in V} e^{\vec{u}_i^T \cdot \vec{v}_c}} \cdot \frac{\partial \sum_{i\in V} e^{\vec{u}_i^T \cdot \vec{v}_c}}{\partial \vec{v}_c}$$

$$= \vec{u}_o - \frac{1}{\sum_{i\in V} e^{\vec{u}_i^T \cdot \vec{v}_c}} \cdot \sum_{j\in V} \left(e^{\vec{u}_j^T \cdot \vec{v}_c} \cdot \vec{u}_j\right)$$

$$= \vec{u}_o - \frac{1}{\sum_{i\in V} e^{\vec{u}_i^T\cdot\vec{v}_c}} \cdot \sum_{j\in V}\left(e^{\vec{u}_j^T\cdot\vec{v}_c}\cdot \vec{u}_j\right)$$

$$= \vec{u}_o - \sum_{j\in V} \boxed{\frac{e^{\vec{u}_j^T\cdot\vec{v}_c}}{\sum_{i\in V} e^{\vec{u}_i^T\cdot\vec{v}_c}}} \cdot \vec{u}_j$$

注意到红框中的式子，所以可简化为：

$$☆\quad \frac{\partial \log P(w_o|w_c)}{\partial \vec{v}_c} = \vec{u}_o - \sum_{j\in V}\left(P(w_j|w_c)\cdot \vec{u}_j\right)$$

计算开销大！

- CBOW

e.g. the man hit his son

hit：中心词

the man, his son：背景词 $\xrightarrow{\text{生成}}$ 中心词

(1) max $\prod_{t=1}^{T} P(w^{(t)} | w^{(t-m)}, \dots, w^{(t-1)}, w^{(t+1)}, \dots, w^{(t+m)})$

⇕

(2) min $-\sum_{t=1}^{T} \log P(w^{(t)} | w^{(t-m)}, \dots, w^{(t-1)}, w^{(t+1)}, \dots, w^{(t+m)})$

(3) $$P(w_c | w_{o_1}, \dots, w_{o_{2m}}) = \frac{e^{\vec{u}_c^T \boxed{(\vec{v}_{o_1} + \dots + \vec{v}_{o_{2m}})/2m}}}{\sum_{i\in V} e^{\vec{u}_i^T(\vec{v}_{o_1}+\dots+\vec{v}_{o_{2m}})/2m}}$$

对 2m 个背景词向量做算术平均

窗口大小为 m

窗口内总共 2m 个词

(4) $$\frac{\partial P(w_c | w_{o_1}, \dots, w_{o_{2m}})}{\partial \vec{v}_{o_i}} = \frac{1}{2m}\left(\vec{u}_c - \sum_{j\in V} \boxed{\frac{e^{\vec{u}_j^T\cdot\vec{v}_c}}{\sum_{i\in V} e^{\vec{u}_i^T\cdot\vec{v}_c}}} \cdot \vec{u}_j\right)$$

证略

$$\frac{\partial P(w_c | w_{o_1}, \dots, w_{o_{2m}})}{\partial \vec{v}_{o_i}} = \frac{1}{2m}\left(\vec{u}_c - \sum_{j\in V} P(w_j | w_c)\cdot \vec{u}_j\right)$$

- 近似训练法
  - 负采样

— 设 w_c 和 w_o [illegible] 近似

① w_c 和 w_o 同时出现在训练数据窗口

② w_c 和第1个噪声词 w_1 不同时出现在 - - -

⋮

w_c 和第k个噪声词 w_k - - - - -