cs224-1

#深度学习/nlp/cs224

cs224n-2019-lecture01-wor…
2019年3月28日

1.如何表示一个单词的含义?

含义的定义：

- the idea that is represented by a word, phrase,
- the idea that a person wants to express by using words, signs, etc.
- the idea that is expressed in a work of writing, art, etc.

signifier (symbol) ⟺ signified (idea or thing)

= denotational semantics

独热编码

维度大小为整个语料库，对应单词在自己位置上为1，其余位置为0，所以称为独热编码

缺点

不同的单词向量之间相互正交，无法计算相似度

2.分布式语义

一个单词的含义通过邻近单词来定义

通过出现在类似上下文中的相同位置来判断不同词之间的相似度

词向量

以下等义

- word vectors
- word embeddings
- word representations

3.Word2vec:Overview

word2vec:一种学习词向量的模型框架

理念

- 我们有大量的文本语料
- 待修正的词表中的每一个单词都通过一个向量来表示
- 对于文本中的每一个位置t，都有一个中心单词C和指定大小的上下文O

- 对于C和O，通过每个词向量的相似度来计算给定C情况下O出现的概率(skip-gram)和相反的情况(CBOW)
- 不断地修正词向量使得概率最大化

  目标函数： object function

For each position $t = 1, \dots, T$, predict context words within a window of fixed size m, given center word w_j.

$$\text{Likelihood} = L(\theta) = \prod_{t=1}^{T} \prod_{\substack{-m \le j \le m \\ j \ne 0}} P(w_{t+j} \mid w_t; \theta)$$

θ is all variables to be optimized

sometimes called *cost* or *loss* function

The objective function $J(\theta)$ is the (average) negative log likelihood:

$$J(\theta) = -\frac{1}{T} \log L(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{\substack{-m \le j \le m \\ j \ne 0}} \log P(w_{t+j} \mid w_t; \theta)$$

Minimizing objective function $\Leftrightarrow$ Maximizing predictive accuracy

- 使目标函数最小化：

$$J(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{\substack{-m \le j \le m \\ j \ne 0}} \log P(w_{t+j} \mid w_t; \theta)$$

- 计算条件概率
- 用两个向量来表示两个单词
  - v_w：当单词是中心单词时
  - u_w：当单词属于上下文时
- 对于一个中心单词C和一个上下文单词o求条件概率时：

$$P(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$$

  - 分子：点积比较o和c的相似性

- 分母：对整个词汇进行归一化以给出概率分布

- **softmax**:对任意分布值求概率

$$\operatorname{softmax}(x_i)=\frac{\exp(x_i)}{\sum_{j=1}^{n}\exp(x_j)}=p_i$$

训练模型：通过损失函数来优化参数

计算所有的向量梯度

- theta：代表所有的模型参数
- 小技巧：

$$\frac{\partial \mathbf{x}^T\mathbf{a}}{\partial \mathbf{x}}=\frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}}=\mathbf{a}$$

$$\max\ J'(\theta)=\prod_{t=1}^{T}\prod_{\substack{-m\le j\le m\\ j\ne 0}} p(w_{t+j}\mid w_t;\theta)$$

$$J(\theta)=-\frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-m\le j\le m\\ j\ne 0}}\log p(w_{t+j}\mid w_t)$$

目标函数
最小化

$$p(o\mid c)=\frac{\exp(u_o^T v_c)}{\sum_{w=1}^{V}\exp(u_w^T v_c)}$$

$$\frac{\partial}{\partial v_c} \log \frac{\exp(u_o^T v_c)}{\sum_{w=1}^{V} \exp(u_o^T v_c)}$$

$$\frac{\partial}{\partial v_c} \log \exp(u_o^T v_c) - \frac{\partial}{\partial v_c} \log \sum_{w=1}^{V} \exp(u_o^T v_c)$$

numerator　　　　　　denominator

$$\frac{\partial}{\partial v_c} \log \exp(u_o^T v_c) = u_o$$

$T \rightarrow$ 语料库

$$\frac{\partial}{\partial v_c} \log \sum_{w}^{T} \exp(u_w^T v_c)$$

$$= \frac{\frac{\partial \sum_{w}^{T} \exp(u_w^T v_c)}{\partial v_c}}{\sum_{w}^{T} \exp(u_w^T v_c)} = \frac{\sum_{w}^{T} u_w^T \exp(u_w^T v_c)}{\sum_{w}^{T} \exp(u_w^T v_c)}$$

$$\frac{\partial}{\partial v_c} \log P(o|c) = u_o - \sum_{x=1}^{V} \underbrace{\boxed{\frac{\exp(u_x^T v_c)}{\sum_{w=1}^{V} \exp(u_w^T v_c)}}}_{p(x|c)} \cdot u_x$$

$$= u_0 - \sum_{x=1}^{V} p(x|c) \cdot u_x$$