DeepNLP之word2vec (附代码)

原创 小明 小明AI学习 2017-12-02

导读

本文主要通俗地简述word2vec基本原理，为了帮助大家更好的理解文末会附上代码链接

目录

1. 词向量的表示方式
2. word2vec的基本原理及应用
3. word2vec的实现方式

词向量的表示方式主要有两种：

1. one-hot Representation (独热表示)

假设你有一个词典，独热编码用一个高维度的向量来表示词典中的每个词，向量的维度为词典的大小，向量的分量只有一个 1，其他全为 0， 1 位置对应该词在词典中的位置。举个例子：

比如你有词典

```
{ 'I', 'have', 'an', 'apple' }
```

那么有以下表示

```
V( I ) = [ 1, 0, 0, 0 ]
V( have ) = [ 0, 1, 0, 0 ]
V( an ) = [ 0, 0, 1, 0 ]
V( apple ) = [ 0, 0, 0, 1 ]
```

可以看到 one-hot 编码非常简洁，就是一串比特串，而且仅有一个分量为1。

但是想象一下如果词典非常大的话，那 one-hot 维度将是非常庞大的，在DL一些算法中很容易产生维数灾难。而且 one-hot 编码可以认为没有距离的概念，很难发掘词语词之间的关系。

2. Distributed Representation （分布式表示）

相对于 one-hot 编码分布式表示的特点是维度低，通过训练可以隐含语义信息。

Word2vec 是一种分布式表示，它的基本思想是通过训练将某种语言中的每一个词映射成一个固定长度的向量，将所有的向量放在一起形成一个词向量空间，而每一向量为该空间中的一个点，在这个空间上引入"距离"的概念后，那么就可以根据词之间的距离来判断它们之间的相似性了。

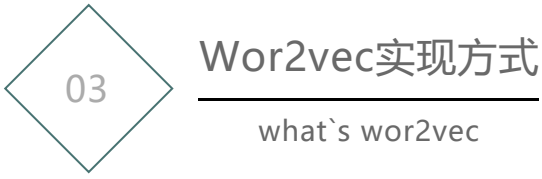
Word2vec在NLP上大有用武之地，比如

1）寻找相似词

2）词的特征扩充

3）关系挖掘等...

在一些DeepNLP任务中，常用word2vec预训练的词向量作为静态的（static）词向量训练模型。



实现word2vec的方法主要有：

1）cbow + hierarchical softmax

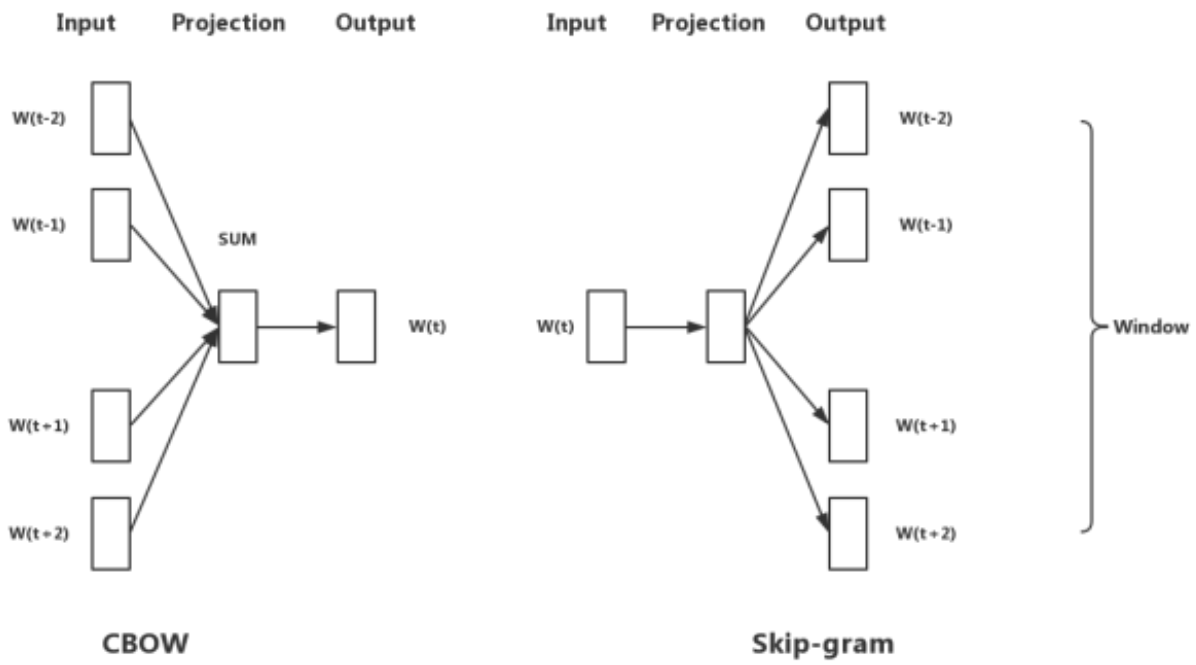
2）cbow + negative sampling

3）skipgram + hierarchical softmax

4）skipgram + negative sampling

(代码给出了上述 2*2 = 4种实现，仅供参考)

CBOW & Skipgram



CBOW（ Continuous Bagof-Words）模型可以简单理解为：用中心词的前后C个词（上下文）来计算中心词出现的概率。

而Skip-Gram 相反，是根据某个词，然后分别计算它前后C个词的各自出现的概率。举个例子：

假设有句子

```
I like to eat apple
```

假设有

1）window=5 (window可以理解为取长度为window一段训练)，

2）中心词为 to

则对于 CBOW 来说有

```
input : ( I, like, eat, apple )
target: ( to )
```

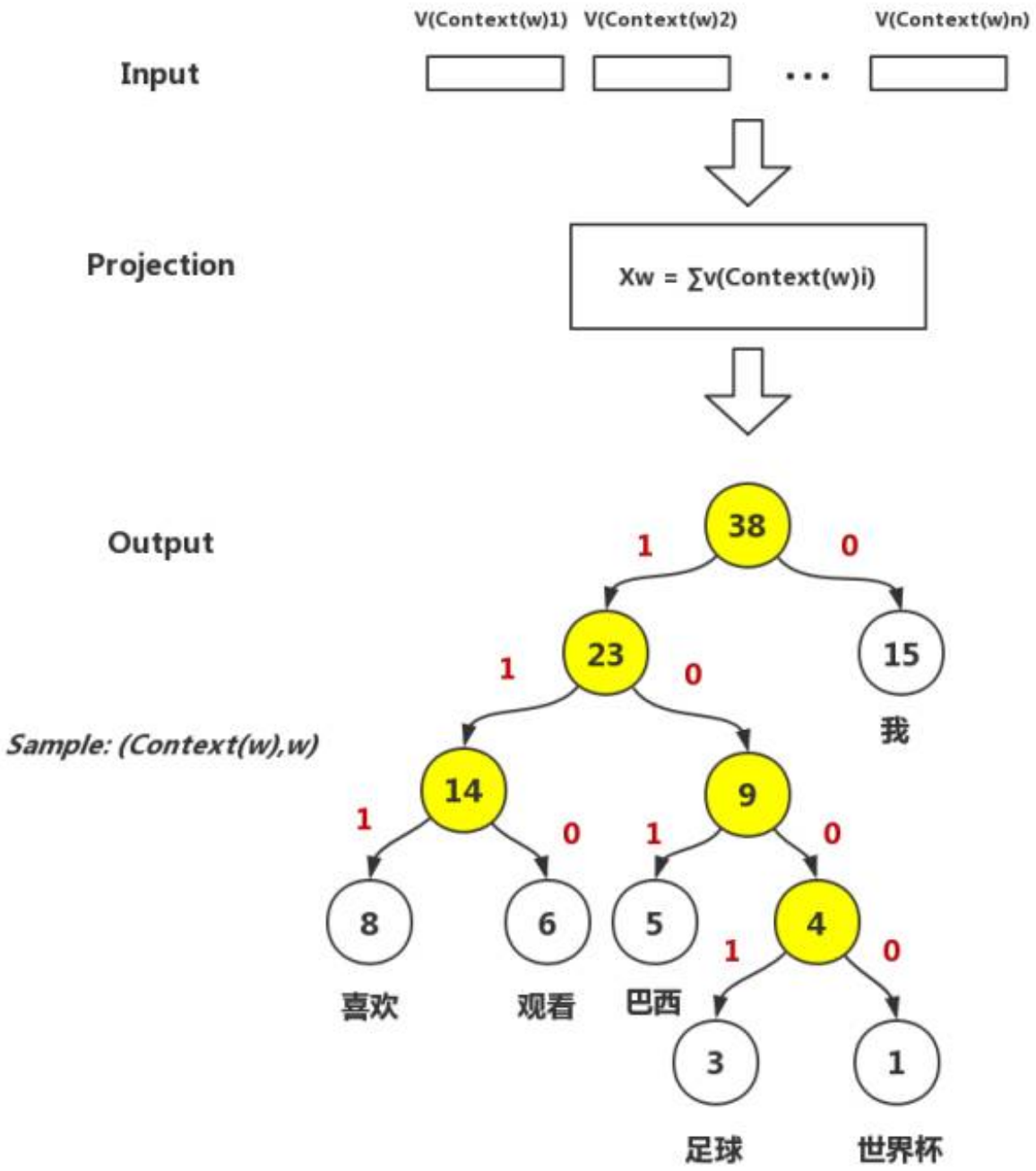
对于Skip-gram来说有

```
( input, target )
=> ( to, I )
=> ( to, like )
=> ( to, eat )
=> ( to, apple )
```

从上面可以知道虽然我们不需要使用标注好的语料去训练word2vec但并不代表这是一种无监督的方式，实际上word2vec的训练过程是有监督的

Hierarchical softmax

分层softmax，采用了Huffman树加速训练过程, 以下是cbow + hierarchical softmax模型（具体数学原理在此不详述，文末会附上推荐阅读）



Negative Sampling

将样本分为正样本和负样本，把语料中的一个词串的中心词替换为别的词，构造语料正样本中不存在的词串作为负样本。在这种策略下，优化目标变为了：最大化正样本的概率，同时最小化负样本的概率。负采样方式也可加速训练速度，且可以获得质量较高的词向量。
（具体数学原理在此不详述，文末会附上推荐阅读）

以下是使用《倚天屠龙记》小说训练的结果。可以看到与张无忌最有关的人物是赵敏。

```
word:  张无忌
similarity     word
0.9638                  赵敏
0.9623                  周芷若
0.9469                  张翠山
0.9246                  殷素素
0.8978                  宋青书
0.8925                  杨不悔
0.8749                  灭绝师太
0.8707                  俞莲舟
0.8665                  谢逊
0.8655                  纪晓芙

赵敏 +  周芷若 =
0.9630                  张无忌
0.9369                  张翠山
0.9353                  杨不悔
0.9324                  殷素素
0.9252                  一眼
0.9163                  宋青书
0.9017                  殷梨亭
0.9004                  纪晓芙
0.8928                  殷离
0.8919                  蛛儿
```

推荐阅读