Keras深度学习入门与实战

讲师：日月光华

词嵌入（embedding）

讲师：日月光华　keras答疑群：863291391

日月光华网易云课堂

文本向量化的方法

传统机器学习中：　tf-idf 算法

独热编码 one-hot　或者　k-hot

散列编码

文本词嵌入（word embedding）

词嵌入

词嵌入是单词的一种数值化表示方式，

一般情况下会将一个单词映射到一个高维的向量中（词向量）来代表这个单词

词嵌入

‘机器学习’ 表示为 [1,2,3]

‘深度学习’ 表示为 [2,3,3]

‘日月光华’ 表示为 [9,1,3]

对于词向量，我们可以使用余弦相似度在计算机中来判断单词之间的距离

词嵌入

词嵌入实际上是一种将各个单词在预定的向量空间中表示为实值向量的一类技术。

每个单词被映射成一个向量（初始随机化），并且这个向量可以通过神经网络的方式来学习更新。因此这项技术基本集中应用与深度学习领域。

词嵌入

词嵌入用密集的分布式向量来表示每个单词。

这样做的好处在于与one-hot这样的编码对比，使用词嵌入表示的单词向量往往只有几十或者几百个维度。极大的减少了计算和储存量。

词嵌入

词向量表示方式依赖于单词的使用习惯，这就使得具有相似使用方式的单词具有相似的表示形式。

‘首都’ 和 ‘北京’ 是向量空间中很相近的2个词。

‘光华老师’ 和 ‘北京’ 是向量空间离的很远。

词嵌入

词嵌入是从文本语料中学习到的一种将单词表示为预定义大小的实值向量形式。

学习过程一般与某个神经网络的模型任务一同进行，比如文档分类。

词嵌入的方法

Word2Vec

Word2Vec是一种能有效从文本语料库中学习到独立词嵌入的统计方法。

GloVe

GloVe算法是对于word2vec方法的扩展，更为有效。它由斯坦福大学Pennington等人开发。

使用词嵌入

在自然语言处理项目中使用单词嵌入时，你可以选择下面两种方式：

1. 自己学一个词嵌入
2. 使用别人训练好的词嵌入

谢谢大家

讲师：日月光华　keras答疑群：863291391

日月光华网易云课堂