089 | 为什么需要Word2Vec算法?

2018-04-27 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 06:02 大小 2.77M

至此，关于文本分析这个方向，我们已经介绍了 **LDA**（Latent Diriclet Allocation），这是一个出色的无监督学习的文本挖掘模型。还有“**隐语义分析**”（Latent Semantic Indexing），其核心是基于矩阵分解的代数方法。接着，我们分享了“**概率隐语义分析**”（Probabilistic Latent Semantic Indexing），这类模型有效地弥补了隐语义分析的不足，成为了在 LDA 兴起之前的有力的文本分析工具。我们还介绍了**EM**（Expectation Maximization）算法，这是针对隐参数模型最直接有效的训练方法之一。

今天，我们进入文本分析的另外一个环节，介绍一个最近几年兴起的重要文本模型，**Word2Vec**。可以说，这个模型对文本挖掘、自然语言处理、乃至很多其他领域比如网络结构分析（Network Analysis）等都有很重要的影响。

我们先来看 Word2Vec 的一个最基本的形式。

Word2Vec 背景

了解任何一种模型或者算法都需要了解这种方法背后被提出的动机，这是一种**能够拨开繁复的数学公式从而了解模型本质的方法。**

那么，Word2Vec 的提出有什么背景呢？我们从两个方面来进行解读。

首先，我们之前在介绍 LDA 和 PLSA 等隐变量模型的时候就提到过，这些模型的一大优势就是在文档信息没有任何监督标签的情况下，能够学习到文档的“隐含特性”。这也是文档领域“**表征学习**”（Representation Learning）的重要工具。遗憾的是，不管是 LDA 还是 PLSA 其实都是把文档当作“**词包**”（Bag Of Word），然后从中学习到语言的特征。

这样做当然可以产生不小的效果，不过，从自然语言处理，或者是文档建模的角度来说，人们一直都在探讨，如何能够把**单词的顺序**利用到学习表征里。什么意思呢？文档中很重要的信息是单词的顺序，某一个特定单词组合代表了一个词组或者是一个句子，然后句子自然也就代表着某种语义。词包的表达方式打破了所有词组顺序以及高维度的语义表达，因此长期以来被认为并不能真正学习到语言的精华。

然而，在主题模型这个大旗帜下，已经有不少学者和研究员试图把词序和语义给加入到模型中，这些尝试都没有得到很好的效果，或者模型过于复杂变得不适用。于是，大家都期待着新的工具能够解决这方面的问题。

另外一个思路也是从词包发展来的。词包本身要求把一个词表达成为一个向量。这个向量里只有一个维度是 1，其他的维度都是 0。因为每个词都表达成为这样离散的向量，因此词与词之间没有任何的重叠。既然两个离散的向量没有重叠，我们自然也就无法从这个离散的词包表达来推断任何词语的高维度语义。这也是为什么大家会利用主题模型从这个离散的词包中抽取主题信息，从而达到理解高维度语义的目的。

既然我们的目的是从离散的词包中获取更加丰富的信息，那有没有另外的方法或者途径能够达到这个目的呢？一种基本的假设是这样的：如果我们能够从离散的向量里面抽取出每个词组的“连续”（Continuous）信息向量，假设两个词有相近的意思，那么这两个词的联系向量势必就会比较相近，这样我们就能够通过词向量（只不过是连续向量），来得到词汇的高级语义信息。这个假设常常被叫作词的“**分布假设**”（Distributed Assumption）。

了解了以上这两个方面后，我们再来理解 Word2Vec，可能就比较容易明白这个模型究竟想要干什么了。

Word2Vec 模型摘要

首先我们需要说明的是，**Word2Vec 是一种语言模型**，主要是根据当前的语境，来预测下一个单词出现的概率，也就是和我们之前所说的产生式模型相似，看是否能够从模型中产生单词。这和我们介绍的主题模型是不一样的，在这个模型里，我们并没有假定数据（也就是单词）是从某几个主题中产生的。

Word2Vec 的核心思想是，当前的单词是从周边单词的隐含表达，或者说是词向量中产生的。也就是说，每一个单词都依赖于上下文，而这个单词的产生，并不是直接依赖周围单词的离散表达，而是依赖**周边单词的连续表达。**这个连续表达自然是事先不知道的，因此这就是 Word2Vec 模型需要学习的**未知参数。**

在具体的操作上，Word2Vec 有两个不太一样的模型，但是经常被同等程度地使用。我们这里做一个简单的介绍。

第一种模型叫作**Skip-Gram**，或者简称**SG 模型。**这种模型的输入是一个词，输出是这个词周围的词。这样做的目的是，看我们能否用当前的词来预测周围的词。要想让这个任务有很好的表现，当前词的表征必须能够抓住某种语义的信息。具体来说，我们就是用当前词的表征向量，和所有其他词的表征向量做点积，然后再重新归一。这个过程就能够保证，当前词的表征向量和周围词的表征向量相似。这样，也就解决了我们之前提到的，如何能够把词序影响到词的表征向量中。

另外一种模型叫作**Continuous-Bag-of-Word**，有时候简称**CBOW 模型。**这种模型刚好和 SG 是相反的，也就是输入是一组词汇，而希望能够通过这组词汇得到中间某个词的预测。和我们刚才所说的一样，这个模型也是基于我们并不知道的表征向量来达到模型学习的目的。

不管是 SG 还是 CBOW，本质上，就是希望能够利用文章的上下文信息学习到连续空间的词表达，这是 Word2Vec 所有模型的核心。

SG 和 CBOW 在具体的应用中，常常需要比较复杂的训练算法，我们这里就不展开讨论了。如果你有兴趣可以进一步阅读一些论文。

总结

今天我为你介绍了 Word2Vec 模型的基本含义。

一起来回顾下要点：第一，我们介绍了 Word2Vec 这个模型是怎么被开发出来的，它背后有哪些原理；第二，我们讨论了 SG 和 CBOW 这两种非常典型的 Word2Vec 模型。

最后，给你留一个思考题，和 LDA 相比，Word2Vec 好在哪里，又有什么不足的地方？

欢迎你给我留言，和我一起讨论。

上一篇　088 | 基础文本分析模型之三：EM算法

下一篇　090 | Word2Vec算法有哪些扩展模型？

精选留言 (2)

写留言

hayley
2018-07-19

对于短文本，怎么提升word2vec的效果？

展开

arfa

2018-05-11

LDA得到的是主题及主题的词分布，word2vec计算的是是词向量

展开