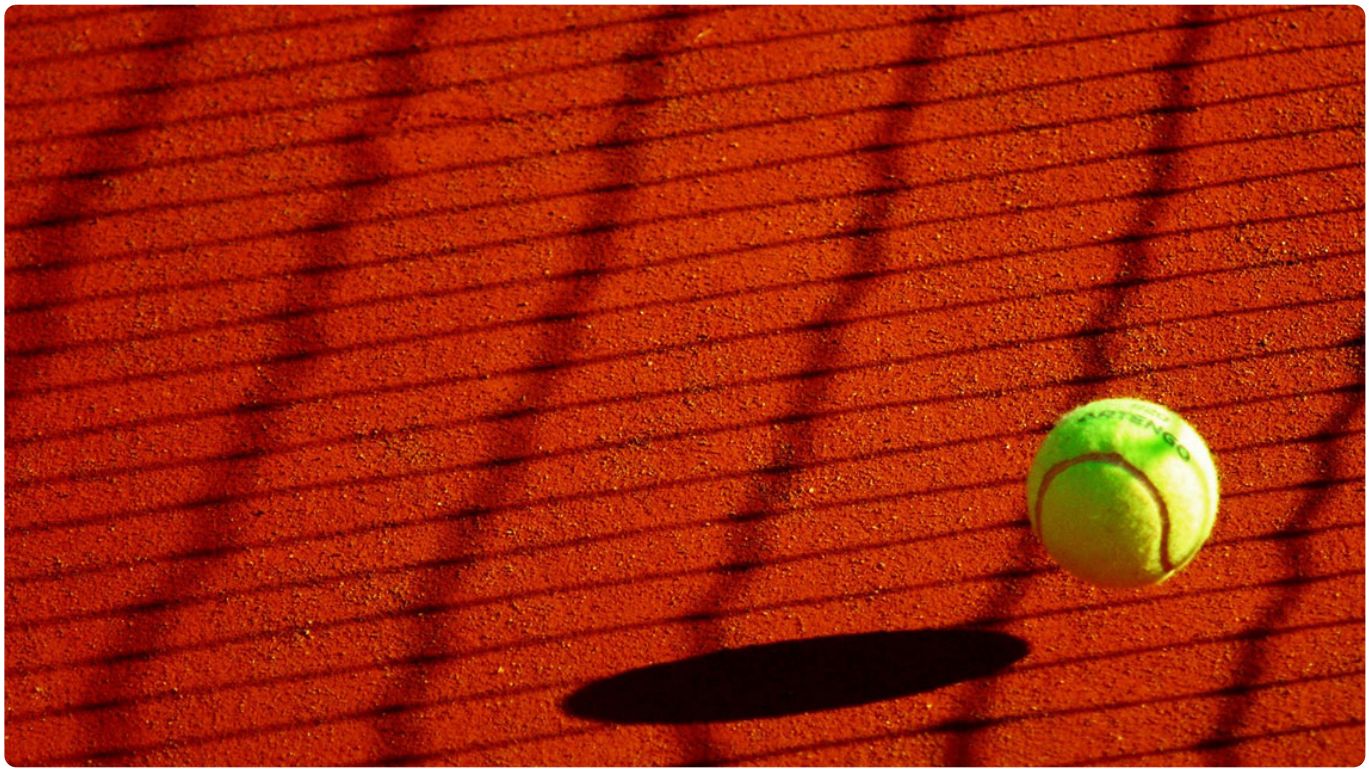
086 | 基础文本分析模型之一：隐语义分析

2018-04-20 洪亮劼

AI技术内参 进入课程 >



讲述：初明明
时长 05:48 大小 2.66M

本周我们分享了文本挖掘中的一个重要工具 LDA（Latent Diriclet Allocation），这是一个出色的无监督学习的文本挖掘模型。

今天，我们沿着文本分析这一方向继续展开。我们首先回到一个最基础的问题，那就是文本分析的基础模型都有哪些，这些最早的模型对后面的发展都有哪些贡献和启发？

带着这些问题，我们一起来看一个叫“**隐语义分析**”（Latent Semantic Indexing）的技术。

隐语义分析的背景

为什么需要隐语义分析呢？隐语义分析到底发挥了怎样的历史性作用呢？

对于数据挖掘而言，文本数据算是大规模数据中，研究人员最早接触到的一类数据了。长久以来，大家都有一种直观的想法，那就是在这些看似没有头绪的文字中，究竟有没有隐含着某些规律呢？我们到底能不能从文字中提取出一些更加有用的结构性的内容呢？

对于文本分析，有一类是基于“显式”的标签来进行的。也就是说，我们可以把文本分析当作是**监督学习的任务**来看待。这一类文本分析的一大特点，往往是针对某一种任务建立分类器，然后对不同类别的文本进行鉴别，从而达到更加深入理解文本的目的。比如，我们需要理解不同情感的文字的时候，通常情况下，我们需要有一个数据集，能够告诉我们哪些文档是“正面情绪”的，哪些是“负面情绪”的。

然而，并不是所有的文本分析任务都是建立在有数据标签的基础之上。实际上，对于绝大多数文本数据而言，我们事先是并没有标签信息的。那么，**在没有标签信息的场景下，如何对文本提取关键信息就成为了研究人员长期面对的一个关键挑战。**

如果我们用今天的眼光来看，**隐语义分析的核心其实就是用无监督的方法从文本中提取特性**，而这些特性可能会对原文本的深层关系有着更好的解释。

其实，从 20 世纪 80 年代发展出来的隐语义分析，一直到今天利用深度学习技术来对文本的内涵进行分析，其实质都是一样的，都是看如何能够用无监督的方法提取文本特性，一个重要的区别当然是在提取办法的差异上。

隐语义分析

对隐语义分析的一个简单直白的解释就是：**利用矩阵分解的概念对“词 - 文档矩阵”（Term-Document Matrix）进行分解。**

在前面介绍推荐系统的时候，我们已经看到了矩阵分解可以认为是最直接的一种针对矩阵数据的分析方式。

那么，为什么我们需要对矩阵进行分解呢？

这里面的一个隐含的假设就是，**“词 - 文档矩阵”是一个稀疏矩阵。**什么意思？意思就是从大规模的文字信息来说，文字服从一个叫“**幂定律**”（Power Law Distribution）的规律。那就是绝大多数的单词仅出现很少的次数，而少数的单词会出现在很多文档中。我们也可以理解成一种变形的“20/80”原理，也就是 20% 的单词出现在 80% 的文档中。当

然，文字的幂定理规则的一个直接结果就是“词 - 文档矩阵”是稀疏矩阵。这个矩阵里面有大量的零，代表很多单词都没有出现在那些文档中。

对一个稀疏矩阵，我们往往假设原有的矩阵并不能真正表示数据内部的信息。也就是说，我们认为可能会有一个结构存在于这个矩阵之中。而这个假设，就是我们经常会在矩阵分解这个语境中提到的“**低维假设**”（Low-rank Approximation）。你不必去担心这个低维假设的本质意义，我们只需要理解这个低维假设的核心，就是我们**可以用比较少的维度来表达原来的这个稀疏的矩阵。**

试想我们拥有一个 N 乘 M 的“词 - 文档矩阵”，也就是说我们有 N 个单词，M 个文档。在这个稀疏矩阵的数据中，矩阵分解的基本思想是希望得到一个 N 乘以 K 的单词矩阵，以及一个 K 乘以 M 的文档矩阵。K 是一个事先指定好的参数，这也是矩阵分解的一个核心问题，那就是如何选择这个 K。我们可以看到，这种分解能够还原之前的 N 乘以 M 的“词 - 文档矩阵”。

那么，这两个新的矩阵有什么“含义”呢？人们通过对很多数据的分解以后发现，单词矩阵往往能够把一些在某种语境下的单词给聚拢。比如我们会发现，很多和体育相关的词会聚拢在某个维度下，而很多和金融相关的词会聚拢在另外一个维度下。慢慢地，大家就开始把每一个维度认定为一个“主题”。那么，**基于矩阵分解的隐语义分析其实就是最早的主题模型。**而文档矩阵则描述了不同文档在我们 K 个主题下的强度。

值得注意的是，我们这里为了介绍隐语义模型的实际意义而隐藏了一些实际的技术细节。从历史上看，比较流行的隐语义模型其实是基于“**奇异值分解**”（Singular Value Decomposition），也就是我们常常听到的**SVD 分解。**由于篇幅有限，我们这里就不针对 SVD 分解展开讨论了。即便是 SVD 分解，其核心思想依然是我们刚才讲到的分解出来的主题矩阵。

基于矩阵分解的隐语义模型也有其局限性，最大的一个问题就是分解出来的矩阵本身都是实数，也就是有负数和正数，这也限制了我们真正用这些数来进行一些含义的推断。然而，即便如此，**在很长的一段时间里，基于 SVD 的隐语义模型可以说是标准的无监督文本挖掘的核心算法。**

总结

今天我为你介绍了基于矩阵分解的隐语义模型的相关知识。

一起来回顾下要点：第一，我们聊了聊为什么需要隐语义模型；第二，我们聊了一下基于矩阵分解的隐语义模型的核心思想及其局限。

最后，给你留一个思考题，如果我们要限制矩阵分解的结果是非负数，我们应该怎么做呢？

欢迎你给我留言，和我一起讨论。

上一篇　085 | 针对大规模数据，如何优化LDA算法？

下一篇　087 | 基础文本分析模型之二：概率隐语义分析

精选留言 (2)

写留言

林彦

2018-04-21

迭代更新矩阵值时把梯度下降方法中的加减更新替换成乘除更新，保证初始值和更新步长值都是非负数，则计算出来的矩阵值为非负数

rushui

2018-04-20

nmf 非负矩阵分解

展开