【关于 数据挖掘】那些你不知道的事

摘要

一、什么是文本挖掘?

文本挖掘指的是从文本数据中获取有价值的信息和知识，它是数据挖掘中的一种方法。文本挖掘中最重要最基本的应用是实现文本的分类和聚类，前者是有监督的挖掘算法，后者是无监督的挖掘算法。

二、文本挖掘的作用是什么?

能够从文本数据中获取有价值的信息和知识

三、文本预处理

3.1 中文分词

使用jieba来对文本进行分词处理，它有3类分词模式，即全模式、精确模式、搜索引擎模式：

- 精确模式：试图将句子**最精确地切开**，适合文本分析；
- 全模式：把句子中所有的可以成词的词语都扫描出来, 速度非常快，但是**不能解决歧义**；
- 搜索引擎模式：在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。

> 举例：定位理论认为营销的终极战场在于消费者心智
>
> 【全模式】: 定位/理论/定位理论/认为/营销/的/终极/战场/终极战场/在/于/在于/消费者/心智/消费者心智

> 【精确模式】: 定位理论/认为/营销/的/终极战场/在于/消费者心智
>
> 【搜索引擎模式】: 定位，理论，定位理论，认为，营销，的，终极，战场，终极战场，在于，消费者心智，消费者，心智

3.2 去停用词

对于 文本中 的 停用词需要做处理

- 标点符号：，。！ /、*+-
- 特殊符号：❤️◗웃유♋☮✌☭☢☠✔☑♚▲♪等
- 无意义的虚词：“the”、“a”、“an”、“that”、“你”、“我”、“他们”、“想要”、“打开”、“可以”等

3.3 低频词和高频词处理

- 动机：低频词和高频词对于 后续的文本分析 容易造成影响，比如对于后续的主题模型（LDA、ATM）时使用的，主要是为了排除对区隔主题意义不大的词汇，最终得到类似于停用词的效果。

3.4 计算 N-gram 【这里采用 Bigrams】

- 动机：针对 分词工具 分错 一些新词问题，比如基于词汇之间的共现关系—如果两个词经常一起毗邻出现，那么这两个词可以结合成一个新词，比如“数据”、“产品经理”经常一起出现在不同的段落里，那么，“数据_产品经理”则是二者合成出来的新词，只不过二者之间包含着下划线。

四、文本挖掘

4.1 关键词提取

- 动机：对于文档而言，可以抽取出提取某段文本的关键信息，即关键词来表示文档信息；
- 方法：TF-IDF（termfrequency–inverse document frequency）

> 它用以评估一字/词对于一个文件集或一个语料库中的其中一份文件的重要程度，字/词的重要性会随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

4.2 LDA 主题模型分析

- 动机：如果采用 关键词的分类较为粗糙，人为因素严重，达不到全面效果。
- 方法：LDA 主题模型
- 作用：抽取出语料中的潜在主题以及每个主题下对应的主题词

4.3 情绪分析&LDA主题模型交叉分析

- 动机：对于有些文本，我们需要找到其所表述的情感信息；
- 方法：使用基于深度学习的情绪语义分析模型（该模型有6类情绪，即喜悦、愤怒、悲伤、惊奇、恐惧和中性）

4.4 ATM 模型

- 动机：想了解“文本中各个作家的写作主题，分析某些牛X作家喜欢写哪方面的文章（比如“行业洞察”、“爆品营销”、“新媒体运营”等），以及写作主题类似的作者有哪些；
- 方法：ATM模型（author-topic model）也是“概率主题模型”家族的一员，是LDA主题模型（Latent Dirichlet Allocation ）的拓展，它能对某个语料库中作者的写作主题进行分析，找出某个作家的写作主题倾向，以及找到具有同样写作倾向的作家，它是一种新颖的主题探索方式。

4.5 词向量训练及关联词分析

- 动机：前面的方法无法学习到语义信息问题
- 方法：word2vec
- 介绍：基于深度神经网络的词向量能从大量未标注的普通文本数据中无监督地学习出词向量，这些词向量包含了词汇与词汇之间的语义关系
- 原理介绍：基于词嵌入的Word2vec是指把一个维数为所有词的数量的高维空间嵌入到一个维数低得多的连续向量空间中，每个单词或词组被映射为实数域上的向量。把每个单词变成一个向量，目的还是为了方便计算，比如“求单词A的同义词”，就可以通过“求与单词A在cos距离下最相似的向量”来做到。

4.6 词聚类与词分类

- 动机：需要对某类词 进行聚类；
- 方法：聚类分析
- 思路：运用基于Word2Vec（词向量）的K-Means聚类，充分考虑了词汇之间的语义关系，将余弦夹角值较小的词汇聚集在一起，形成簇群。

4.7 文本分类

文本分类是一种典型的机器学习方法，一般分为训练和分类两个阶段。文本分类一般采用统计方法或机器学习来实现。

4.8 文本聚类

- 类型：无监督方法
- 思路：
    - 首先，文档聚类可以发现与某文档相似的一批文档，帮助知识工作者发现相关知识；
    - 其次，文档聚类可以将一类文档聚类成若干个类，提供一种组织文档集合的方法；
    - 再次，文档聚类还可以生成分类器以对文档进行分类。

4.9 信息检索

主要是利用计算机系统的快速计算能力，从海量文档中寻找用户需要的相关文档。

参考

1. 以虎嗅网4W+文章的文本挖掘为例，展现数据分析的一整套流程