期中周测试题，你做对了吗？

2021-01-22 刘海丰

成为AI产品经理 进入课程 >

你好，我是海丰。今天，我来公布一下主观题的答案。

我们先来回顾一下题目：

假如，你现在是一家电商平台的产品经理，负责点评系统的产品设计，现在有一个需求是要通过计算将海量评论中的垃圾评论（如，打广告的情况）过滤出来，你会怎么思考和设计产品？

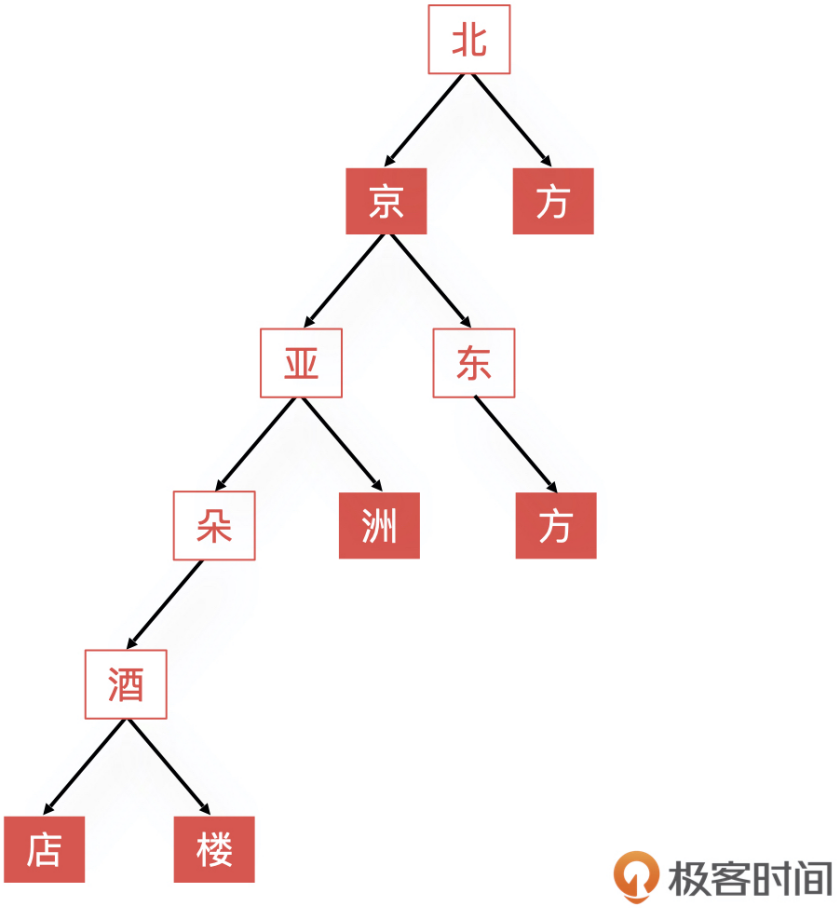
我们知道，用户评论数据都是非结构信息，所以我们首先要做的就是将非结构化数据转化成结构化的。在文本分析中，我们可以使用“词向量”来表示文本中的数据。

举个例子，如果用户评论中出现某些特定词，比如“尊敬的”“您好”“促销”等等，它们很有可能属于垃圾评论。那我们就可以用这些词来构成“词向量”，具体怎么做呢？下面，我分三步来讲。

分词是文本分析的首要工作，“基于字符串匹配”是最简单的分词方法。举个例子，现在有一句短语叫做“北京亚朵酒店”，以及我们事先积累的词库。

首先，我们可以提出短语的第一个字符串“北”，然后将短语中从“北”字开头的后续的内容与词库中收录的词语进行匹配。当匹配到了词库中的词语“北京”后，就可以停止匹配了，“京”字也就作为终止字符。这样，我们就从“北京亚朵酒店”中提出来了第一个词语“北京”。

接着，我们就可以把“北京”这个词从原始短语中删除，从“亚”字重新开始匹配。



根据字符串匹配的方式，在已知业务常见词语的基础之上，我们是可以将评论中的所有词汇都切分出来的，切分的结果如上图所示。

当然，我这里使用“基于字符串匹配”的分词方式讲解是为了让你更容易理解，在实际工作中，我们通常会采用“正逆向最大匹配”的分词算法，以及如果有更为复杂的短语，

第二步：构建训练集和测试集。

解决了分词问题后，垃圾评论分类产品的构建的工作就已经完成一半儿了。接下来，就是构建训练集和测试集了。

首先，我们需要两组评论，一组用于训练，一组用于测试。目前，历史评论的样本数据我们有了，样本数据的标签（正常评论、垃圾评论）我们也有了，那么，只需要确定特征以后，就可以带入分类算法进行训练了。

那么特征是什么呢？其实就是把分词后的每个“词语”在样本中的词频（出现的次数）。

比如说，下面是 5 个评论统计得到的词频统计表。其中“0”表示某个词语在评论中没有出现。“1”则表示某个词语在评论中出现了。

	尊敬的	您好	活动	促销	欢迎	朋友	约会	顾客	分类
1	1	1	0	1	1	0	0	1	垃圾评论
2	1	1	1	1	0	1	0	1	垃圾评论
3	1	0	1	1	1	0	0	0	垃圾评论
频数	3	2	2	3	2	1	0	2	15
4	0	0	1	0	0	1	0	0	正常评论
5	0	0	0	1	0	1	1	0	正常评论
频数	0	0	1	1	0	2	1	0	5

极客时间

第三步：计算概率。

根据刚才得到的表格，我们能够计算出“尊敬的”“你好”这些词语在垃圾评论和正常评论中出现的概率，我把它们总结在了下面的表格里。这个时候，当“顾客”出现在新评论中的，我们就认为它是垃圾评论的概率是 0.2。

频率	0.2	0.13	0.13	0.2	0.13	0.07	0	0.2	垃圾评论
频率	0	0	0.2	0.2	0	0.4	0.2	0	正常评论

极客时间

这样一来，当有了新的评论出现的时候，我们首先对它进行分词，根据概率公式 数学公式: $p = \sum_{i=1}^{8} c_i p_i$ 计算新评论属于垃圾评论的概率。

比如，对于一个含有“尊敬的”、“促销”、“朋友”、“约会”、“顾客”的评论来说。

属于垃圾评论的概率是：

数学公式: $0.67 = 1 \times 0.2 + 0 \times 0.13 + 0 \times 0.13 + 1 \times 0.2 + 0 \times 0.13 + 1 \times 0.07 + 1 \times 0 + 1 \times 0.2$

属于正常评论的概率是：

数学公式: $0.8 = 1 \times 0 + 0 \times 0 + 0 \times 0 + 1 \times 0.2 + 0 \times 0 + 1 \times 0.4 + 1 \times 0.2 + 1 \times 0$

由于 0.8 大于 0.67，所以新评论属于正常评论。

好了，主观题的解题思路就是这样了。那么，期中周马上就要结束了，希望你能尽快巩固好我们所学的内容。我们下节课见！

提建议

下一篇 17 | 模型评估：从一个失控的项目看优秀的产品经理如何评估AI模型？

精选留言 (1)

吴洋
2021-01-25

麻烦问下词“顾客”出现在评论中，可以判断评论为垃圾评论的概率为0.2，为什么像“您好”、“活动”这些的概率是0.13？我想知道我是怎么算错的。。。

展开