**《机器学习》课程实验报告**

|  |  |
| --- | --- |
| **实验名称:** | 文本转VSM |
| **姓 名:** | 李若昊 |
| **学 号：** | 5120180269 |
| **教 师:** | 赵旭剑 |
| **完成时间:** | 2020-5-24 |

|  |
| --- |
| **一、实验目的和内容**  **将文字转化为向量空间的形式表示，为第三次实验——用K-means为不同文档间的主题相似度比较(聚类)做准备，即进行相似度比较**  **把文本转化成数学模型，是用数学方法处理文本的先决条件，这个过程称为文本特征抽取。向量作为一种基本的数学模型，是文本特征抽取的一种常见方法。**  **实验内容参考资料https://www.bilibili.com/video/BV1Bt411p746?p=17** |
| **二、实验过程**  **a.(预处理)考虑到中文语言的特点，区别于英语，我们应该为我们的语言先做“分词”(作为一次学习实验，本次拿到的材料已经被分好词了)**  **b.** **(预处理)分出来的这一堆词语中有很多词是对最终相似度比较没什么贡献的，例如停用词，应该去除；还有重复出现的，要去重；在向量空间角度来说即是达到一个降维的作用**  **开始选特征词：选它们的方式有多种：自动选择的话，可以采用开方拟和检验方法、n-gram方法或者TF-IDF算法，这次实验学习后者，它是计算出文档的每个词的TF-IDF值，然后按降序排列，取排在最前面的(值最大的)几个词作为特征词。**   1. **先计算出现最多的词(TF)：考虑到多篇文章有长短之分，为了便于之后不同文章的比较，采用相对词频** 2. **再计算IDF：有了每个特征词的TF，还要在词频的基础上对每个词分配一个"重要性"权重。理解为一个重要性调整系数，用来衡量一个词是不是常见词。这个权重就是IDF，它的大小与一个词的常见程度成反比。**     **如果一个词越常见，分母就越大，逆文档频率就越小越接近0。分母之所以要加1，是为了避免所有文档都不包含该词，导致分母为0。**   1. **计算TF-IDF：知道了"词频"（TF）和"逆文档频率"（IDF）以后，将这两个值相乘，即得一个词的TF-IDF值。某个词对文章的重要性越高，它的TF-IDF值就越大。所以，选择排在最前面的几个词，就是这篇文章的关键词。** 2. **归一化处理：为了保证权值不偏向长文档，在TF-IDF值计算出来后，再乘上文档长度的倒数进行归一化处理**   **将N篇文档的每一个特征词的TF-IDF实数值写在矩阵里(用矩阵表示)，写的时候一列代表一篇文档；若每一个文档有M个词语(不一定非得严格是二字词)，那这个矩阵就是一个NxM的矩阵，这个矩阵可以就是一个训练集构成的向量空间模型(VSM)，之后的各种算法就是运行在这个矩阵上的：比如计算某两个列向量的余弦相似度，值越大就表示越两篇文章越相似。** |
| **三、实验结果**    **解释执行有点问题…**  **换了个平台有结果了，后面需要查看一下是怎么回事** |
| **四、实验总结**  **TF-IDF算法的优点是简单快速，结果比较符合实际情况。缺点是，单纯以"词频"衡量一个词的重要性，不够全面，有时重要的词可能出现次数并不多。而且，这种算法无法体现词的位置信息，出现位置靠前的词与出现位置靠后的词，都被视为重要性相同，这是不正确的。（一种解决方法是，对全文的第一段和每一段的第一句话，给予较大的权重。）**  **向量空间模型（VSM）就是把一个语言的问题转化成了一个数学问题，这种思想教会我尝试把一些问题转化为数学问题，用数学的手段来解决，也就是给问题赋予数学上的意义，比如这次试验给文章赋予多维向量的意义。** |