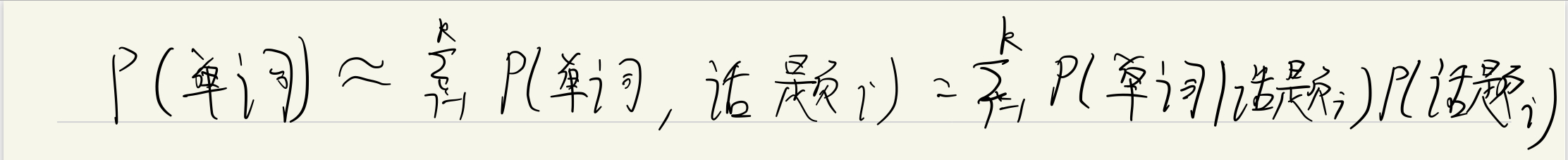
**第十七章读书报告**

09118223 吴亦珂

读书进度：第十七章前半部分

1. **问题列表**

（我提出）P326页上说X约等于TY，是否可以这么理解

讨论结果：可以这么理解。实际上该模型就是将单词在文中的概率拆成多个话题并且分开统计。利用条件概率计算得出。

（我提出）为什么通过奇异值分解的方法就能够直接将左奇异矩阵作为话题矩阵，后面两项作为文本在话题空间的表示？这么做有什么根据吗？

讨论结果：相当于如果需要k个话题，那就认为通过截断奇异值分解的方式，构造出一个k维矩阵来解决。这样，根据SVD定义，这两者乘积可看成原矩阵的近似，同时这两者维嘟也刚好对应，因此可以将左奇异矩阵作为话题矩阵，后面两项作为文本在话题空间的表示。同时由于做奇异矩阵是正交矩阵，不同列之间做内积为0，说明不同话题之间差异很大。因此这么做虽然直接从语义角度分析并不一定正确，但是确实符合实际的需要。

（别人提出）潜在语义分析与主成分分析之间存在什么关系？研究思路是否有差异？  
我觉得潜在语义分析算法利用了SVD来进行。就是SVD是实现这个算法的一个手段,相当于对于将空间的基进行了相应的变换，由原来的的单词向量空间转变为话题向量空间。

（别人提出）求解单词话题矩阵可以使用单词文本矩阵的奇异值分解，除此之外，是否可以使用其他类型的方法直接构造单词话题矩阵关系而不使用分解的方法。  
讨论结果：关键词抽取或者主题抽取的相应算法还有很多，例如LDA也是一种主题模型。

（别人提出）觉得这一章节所谓单词和话题上的相似某种意义上来说就是与文本其它部分的相对关系相似，可以做这样的理解吗？  
讨论结果：我觉得书上的相似使用向量之间的内积来衡量的，其实如果两者含有的话题越相近，乘积项会有更少的项为0，导致乘积越大，同时如果两个文章同时集中在一些单词话题，那么对应值较大也回到值乘积较大。因此我认为这种方法可以说明相似度达则两者含有的相同的单词、话题也就越多，或者说两者集中的单词话题越多，这样一来就可以在一定程度上说明两者属于同一类，统一方向，相似度也就越大。

1. **下周读书计划**

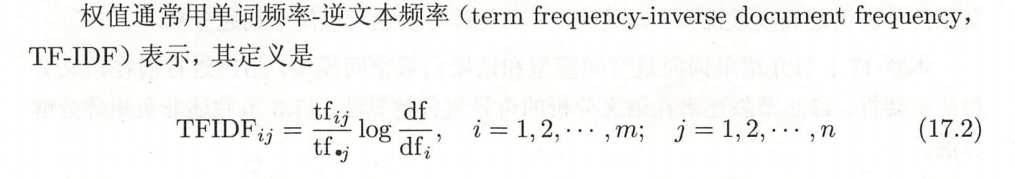
下周计划阅读完第十七章的内容。

1. **读书收获**

第十七章前半部分主要介绍了潜在语义分析的有关概念以及通过SVD 分解矩阵从而获得话题向量空间的方法。

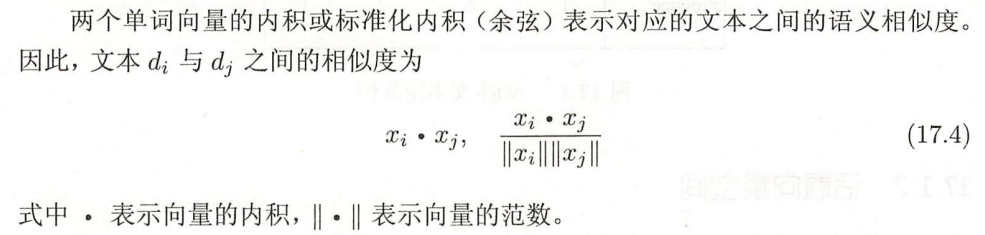
1. 单词向量空间模型：对于一个文本，每一维代表一个单词在文本中出现的频数或权值。

权值表示形式：



1. 单词文本矩阵：每一列代表一个文本，每一行代表一个单词，所以矩阵记录了各个文本中所有单词出现的频率。

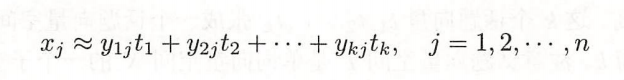
文本的语义相似度：

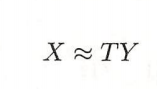


缺点：处理不了一次多义以及同义词的问题。

1. 话题向量空间：每一行代表一个单词，每一列代表一个话题。矩阵中元素表示每一个单词在该话题中的权重。
2. 话题文本矩阵：每一行代表一个单词，每一列代表一个画图，矩阵中元素表示每一个话题在文本中出现的比率。
3. 从单词向量空间到话题向量空间的线性变换：

两个空间可以通过一个表达式关联起来：



即

1. 潜在语义分析算法：

矩阵奇异值分解算法：

1. 将文本数据表示成单词文本矩阵。
2. 对该矩阵进行截断奇异值分解。如果要去k个话题，就进行k维阶段奇异值分解。
3. 将左奇异矩阵作为话题矩阵，后面两项作为文本在话题空间的表示。

即