读书报告

09118119 黄一凡

# 自己提出的问题

#### 1.如何直观的理解对单词-文本矩阵进行奇异值分解，可以得到单词-话题矩阵？为什么可以实现？

其实在单词-文本矩阵中也蕴含了话题的隐性表示，我的理解是话题本身是一些单词的组合，而奇异值分解可以得到较为重要的几个单词，因此可以作为话题的表示。

# 别人提出的问题

#### 1.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？

潜在语义分析利用了SVD来进行，SVD是实现这个算法的一个手段。

#### 2.为什么奇异值分解出来的单词-话题矩阵选择的单词表示同一个话题？

其实SVD的问题就在于它得出来的是计算的结果，但是不一定很好解释。它产生的话题本身也是一个抽象的表示，只是在矩阵分解的过程中这样的组合占比重很大， 具有重要性，所以就作为了话题，话题的单词组合有各种各样的，最后呈现的一般是最重要的几个话题。

#### 3.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？

书上的相似使用向量之间的内积来衡量的，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。

#### 4.如何选择话题数，即k的值？

可以参考奇异值分解那一部分中k的选择，可以根据经验或者一定的指标来选择。

# 读书计划

#### 本周所读：

《统计学习方法》17章部分

#### 下周计划：

《统计学习方法》17章剩余