统计学习方法读书笔记

61518122 丁自民

**读书进度**

第17章第1节

**读书总结**

1. 潜在语义分析是一种无监督学习方法，主要用于文本第话题分析，通过矩阵分解发现文本与单词之间的基于话题的语义关系。
2. 传统方法以单词向量表示文本的语义内容，以单词向量空间的度量表示文本之间的语义相似度。潜在语义分析旨在解决精准表示语义的问题。具体方法就是，用大量的文本进行学习。
3. 单词向量空间：行表示单词，列表示文本，元素值表示单词的权重或频数。一个单词在一个文本中出现次数越高，说明该单词越重要；一个单词在其他文本中出现的越少，说明其越能反应出文本的特点。二者抽象为两种“重要度”，其内积为综合重要度。
4. 单词向量空间模型优点：模型简单，计算效率高；因为矩阵通常是稀疏的，所以实际计算量没有那么大。缺点：会被一词多义或多词同义现象干扰。
5. 话题向量空间：行表示话题，列表示文本；
6. 潜在语义分析的工作就是把单词向量空间通过线性变化转化为话题向量空间。

**问题解答**

吴亦珂问题1:

我觉得，话题矩阵就是单词矩阵经过线性变换之后得到的。话题矩阵应该是一个单位正交矩阵，所以单词-话题矩阵与话题-文本矩阵相乘的时候，话题矩阵消掉了

回答黄一凡问题1:

相当于前面一节所说的主成分分析。话题就是单词的主成分。也即话题之间不交叉，全体话题构成一篇文章。

回答王贵涛问题1：

人为选择，相当于取前k个主成分

**读书疑问**

1、如何解决近义词问题？