方骏-2020年8月2日-读书报告

1. 自己提出的问题的理解：
2. 提出的问题1：潜在语义分析和主成分分析的异同。

讨论后的理解：主成分分析主要是分析数据并做降维处理以达到提取主要信息和方便处理的作用，潜在语义分析主要是用于文本的主题分析。共同点是都会用到奇异值分解，而且本质上，主成分分析和潜在语义分析都是渴望提取出数据的主要信息和减小矩阵的维度，所以两种方法虽然应用方向不同，但是主要思想都是相同的。

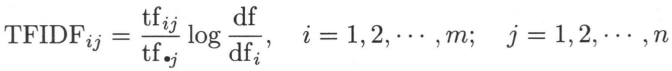
1. 别人提出的问题的理解：
2. 读书计划

1、本周完成的内容章节：《统计机器学习》第十七章

2、下周计划：《统计机器学习》第十八章

四、读书摘要及理解

1、单词向量空间模型通过单词的向量表示文本的语义内容。以单词-文本矩阵为输入，其中每一行对应一个单词，每一列对应一个文本，每一个元素表示单词在文本中的频数或权值。权值通常用单词频率-逆文本频率表示：



2、话题向量空间模型通过话题的向量表示文本的语义内容。有话题-文本矩阵，其中每一行对应一个话题，每一列对应一个文本，每一个元素表示话题在文本中的权值。这个矩阵的每一列向量是话题向量，表示一个文本，两个话题向量的内积或标准化内积表示文本之间的语义相似度。

单词-话题矩阵每一行对应一个单词，每一列对应一个话题，每一个元素表示单词在话题中的权值。

3、潜在语义分析的目标是，找到合适的单词-话题矩阵与话题-文本矩阵，将单词-文本矩阵近似的表示为与的乘积形式：

4、潜在语义分析的算法是奇异值分解。通过对单词-文本矩阵进行截断奇异值分解，得到：



非负矩阵分解也可以用于话题分析。非负矩阵分解将非负的单词-文本矩阵近似分解成两个非负矩阵和的乘积，得到：

矩阵表示话题空间，矩阵是文本在话题空间的表示。

5、非负矩阵分解可以表为以下的最优化问题：

