读书报告

1. （必填）别人提出的问题的理解（选择几个问题罗列，并给出理解）：
2. 问题3：潜在语义分析和主成分分析的异同

自己的理解：主成分分析主要是分析数据的基本结构降维处理的吧，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

1. （必填）读书计划

1、本周完成的内容章节：《统计学习方法》第十七章

2、下周计划：第十八章

三、读书摘要

1、潜在语义分析是一种无监督学习方法，主要用于文本的话题分析，其特点是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。潜在语义分析由Deerwester等于1990年提出，最初应用于文本信息检索，所以被称为语义检索，在推荐系统、图像处理、生物信息学等领域也有广泛应用。

2、文本信息处理中，传统的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这种方法不能明确表示语义的内容，试图从大量的文本数据中发现潜在的话题，以话题向量表示文本的语义内容，以话题向量空间的度量更准确地表示文本之间的语义相似度。这也是话题分析的基本想法。

3、潜在语义分析使用的是非概率的话题分析模型。具体地，将文本集合表示为单词-文本矩阵，对单词-文本矩阵进行奇异值分解，从而得到话题向量空间，以及文本在话题向量，以及文本在话题向量空间的表示。

4、文本信息处理，比如文本信息检索、文本数据挖掘的一个核心问题是对文本的语义内容进行表示，并进行文本之间的语义相似度计算。最简单的方法是利用向量空间模型，也就是单词向量空间模型。向量空间模型的基本思想是，给定一个文本，用一个向量表示该文本的“语义”，向量的每一维对应一个单词，其数值为该单词在该文本中出现的频率或权值；基本假设是文本中所有单词的出现情况表示了文本的语义内容；文本集合中的每个文本都表示为一个向量，存在于一个向量空间；向量空间的度量，或内积或标准化内积表示文本之间的“相似度”。

5、直观上，在两个文本中功通出现的单词越多，其语义内容就越相近，这时，对应的单词向量同不为零的维度就越多，内积就越大，表示两个文本在语义内容上越相似。单词向量空间模型的有点是模型简单，计算效率高。因为单词向量通常是稀疏的，两个向量的内积计算只需要在其同不为零的维度上进行即可。需要的计算很少，可以很高效地完成。单词向量空间模型也有一定的局限性，体现在内积相似度未必能够准确表达两个文本的语义相似度上。因为自然语言的单词具有一词多义性及多词一意性，即同一个单词可以表示多个语义，多个单词可以表示同一个语义，所以基于单词向量的相似度计算存在不精确的问题。

6、两个文本的语义相似度可以体现在两者的话题相似度上，所谓话题，并没有严格的定义，就是指文本所讨论的内容或主题。一个文本一般含有若干个话题。如果两个文本的话题相似，那么两者的语义也应该相似，那么两者的语义应该也相似。话题可以由若干个语义相关的单词表示，同义词可以表示同一个话题，而多义词可以表示不同的话题，而多义词可以表示不同的话题。这样，基于话题的模型可以解决基于单词的模型存在的问题。可以设想定义一种话题向量空间模型，给定一个文本，用话题空间的一个向量表示该文本，该向量的每一分量对应一个话题，其数值为该话题在该文本中出现的权值。用两个向量的内积或标准化内积表示对应的两个文本的语义相似度。话题的个数常常远远小于单词的个数，话题向量空间模型更加抽象。潜在语义分析是构建话题向量的方法，单词向量空间模型与话题向量空间模型可以互为补充，现实中，两者可以同时使用。

6、