一、讨论的问题

1、 P324话题向量空间一节中，提到airplane和aircraft是同义词表示同一个话题。那么计算机如何判断这两个词表示同一话题呢？

计算机通过话题向量矩阵的列向量来确定话题。如果两个列向量非常相似，代表话题相似。

2.LSA的缺陷是什么？

1）LSA可以处理向量空间模型无法解决的一义多词(synonymy)问题，但不能解决一词多义(polysemy)问题。因为LSA将每一个词映射为潜在语义空间中的一个点，也就是说一个词的多个意思在空间中对于的是同一个点，并没有被区分。

2）SVD的优化目标基于L-2 norm 或者 Frobenius Norm 的，这相当于隐含了对数据的高斯分布假设。而 term 出现的次数是非负的，这明显不符合 Gaussian 假设，而更接近 Multi-nomial 分布。

3）特征向量的方向没有对应的物理解释。

4）SVD的计算复杂度很高，而且当有新的文档来到时，若要更新模型需重新训练。

5）没有刻画term出现次数的概率模型。

3.LSA有什么应用？

主要是分类和检索：

1）分类：

a.将词汇表中的字词按意思归类（比如将各种体育运动的名称都归成一类）

b.将文本按主题归类（比如将所有介绍足球的新闻归到体育类）

2）检索：用户提出提问式（通常由若干个反映文本主题的词汇组成），然后系统在数据库中进行提问式和预存的文本关键词的自动匹配工作，两者相符的文本被检出。

4.对原有数据进行降维的原因是什么？

1）低维空间表示可以刻画同义词，同义词会对应着相同或相似的主题。  
2）降维可去除部分噪声。

二、读书计划

本周十七章

下周十八章

三、本章小结：

LSA原理：通过对大量的文本集进行统计分析，从中提取出词语的上下文使用含义。技术上通过SVD分解等处理，消除了同义词、多义词的影响，提高了后续处理的精度。

流程：（1）分析文档集合，建立词汇-文本矩阵A。

（2）对词汇-文本矩阵进行奇异值分解。

（3）对SVD分解后的矩阵进行降维

（4）使用降维后的矩阵构建潜在语义空间