主题模型LSI相关介绍

在语义索引(Latent Semantic Indexing,以下简称LSI)，有的文章也叫Latent Semantic Analysis（LSA）。其实是一个东西，后面我们统称LSI，它是一种简单实用的主题模型。LSI是基于奇异值分解（SVD）的方法来得到文本的主题的。而SVD及其应用我们也可以简单的举例，比如奇异值分解原理与降纤中的应用和矩阵分解在协同过滤算法推荐中的应用。

LSI是最早出现的主题模型了，它的算法原理很简单，一次奇异值分解就可以得到主题模型，同时解决词义的问题，非常漂亮。但是LSI有很多不足，导致它在当前实际的主题模型中已基本不再使用。

主要的问题有：

1） SVD计算非常的耗时，尤其是我们的文本处理，词和文本数都是非常大的，对于这样的高维度矩阵做奇异值分解是非常难的。

2） 主题值的选取对结果的影响非常大，很难选择合适的k值。

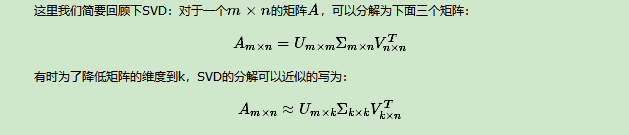
3） LSI得到的不是一个概率模型，缺乏统计基础，结果难以直观的解释。

对于问题1），主题模型非负矩阵分解（NMF）可以解决矩阵分解的速度问题。对于问题2），大部分主题模型的主题的个数选取一般都是凭经验的，较新的层次狄利克雷过程（HDP）可以自动选择主题个数。

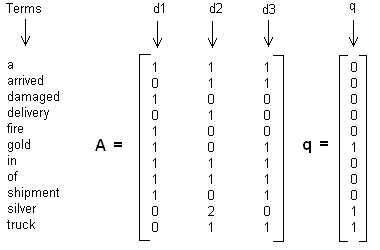
对于问题3），牛人们整出了PLSI和隐含狄利克雷分布(LDA)这类基于概率分布的主题模型来替代基于矩阵分解的主题模型。

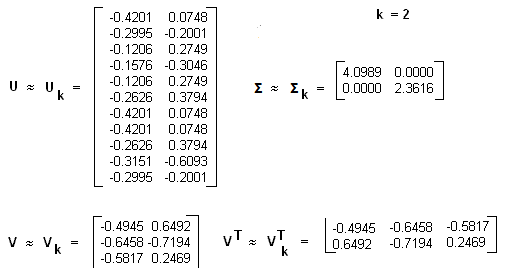
回到LSI本身，对于一些规模较小的问题，如果想快速粗粒度的找出一些主题分布的关系，则LSI是比较好的一个选择，其他时候，如果你需要使用主题模型，推荐使用LDA和HDP。

如果把上式用到我们的主题模型，则SVD可以这样解释：我们输入的有m个文本，每个文本有n个词。而Aij则对应第i个文本的第j个词的特征值，这里最常用的是基于预处理后的标准化TF-IDF值。k是我们假设的主题数，一般要比文本数少。SVD分解后，Uil对应第i个文本和第l个主题的相关度。Vjm对应第j个词和第m个词义的相关度。Σlm对应第l个主题和第m个词义的相关度。



IMG_256





IMG_256