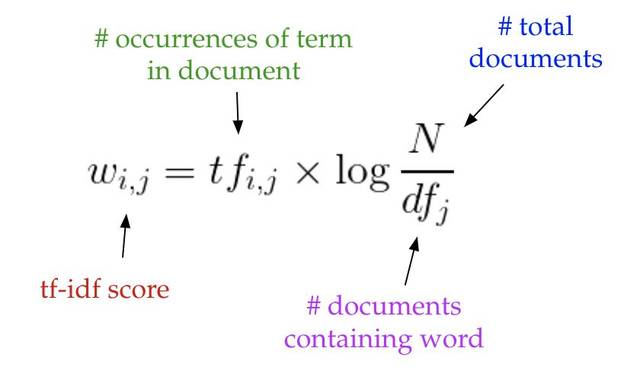
LSA

（LSA）



直观地说，术语出现在文档中的频率越高，则其权重越大；同时，术语在语料库中出现的频率越低，其权重越大。

一旦拥有文档-术语矩阵 A，我们就可以开始思考潜在主题。问题在于：A 极有可能非常稀疏、噪声很大，并且在很多维度上非常冗余。因此，为了找出能够捕捉单词和文档关系的少数潜在主题，我们希望能降低矩阵 A 的维度。

这种降维可以使用截断 SVD 来执行。SVD，即奇异值分解，是线性代数中的一种技术。该技术将任意矩阵 M 分解为三个独立矩阵的乘积：M=U\*S\*V，其中 S 是矩阵 M 奇异值的对角矩阵。很大程度上，截断 SVD 的降维方式是：选择奇异值中最大的 t 个数，且只保留矩阵 U 和 V 的前 t 列。在这种情况下，t 是一个超参数，我们可以根据想要查找的主题数量进行选择和调整。

：

降维是LSA分析的核心。具体做法，是对词项文档矩阵（也叫Term-Document矩阵，以词项(terms)为行, 文档(documents)为列,这里记为C）做SVD分解

\

设C一共有t行d列, 矩阵的元素为词项的TF-IDF（词频term frequency&ndash;反文档频率inverse document frequency）值。然后把的r个对角元素的前k个保留（最大的k个保留）, 后面最小的r-k个奇异值置0, 得到\，最后计算一个近似的分解矩阵\，则是在最小二乘意义下C的最佳逼近，从而得到原始词项文档矩阵的一个低秩逼近矩阵。

通过在SVD分解近似，我们将原始的向量转化成一个低维隐含语义空间中，起到了特征降维的作用。每个奇异值对应的是每个“语义”维度的权重，将不太重要的权重置为0，只保留最重要的维度信息，去掉一些“噪音”,因而可以得到文档的一种更优表示形式。

PLSA：

Hofmann于1999年在SIGIR&#39;99提出了基于概率统计的PLSA模型，并且用EM算法学习模型参数。在传统文本分类或聚类模型中，一篇文章（document）只有一个topic（姑且翻译成主题）,而Hofmann提出，一篇文章可以有多个topic，这就是自然语言处理领域颇具开创性的Topic Model的思想。

另外，

4. pLSA的不足

1）

LDA：

LDA

LDA算法全称为Latent Dirichlet Allocation，LLDA算法LDA的参数规模只取决于Topic个数和字典中term总数。LDA在document到topic一层引入了Dirichlet分布，这是它优于PLSA的地方，使得模型参数的数量不会随着语料库的扩大而增多。

LSA的核心思想是降维，只是基于SVD分解的LSA方法是借助于矩阵分解降维，而PLSA是借助于引进Topic层降维。