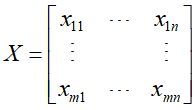
1. **潜在语义分析**

**潜在语义分析(probabilistic latent semantic analysis，PLSA)**是一种利用概率生成模型对文本集合进行话题分析的无监督学习方法。

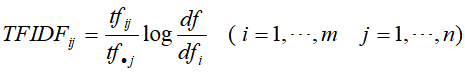
**单词向量空间模型**：即对于一个给定的文本，用一向量表示该文本的语义，向量的每一个维度表示一个单词，其数值为该单词在文本中出现的频数或权数。而模型的基本假设是文本中所有单词的出现情况即表示该文本的语义内容。

关于不同文本之间的语义相似度的度量，则可以用文本在单词向量空间中所表示的向量进行内积，而内积值的大小对应了文本之间语义相似度的情况。

单词向量空间模型的定义，假定n个文本的集合IMG_256，文本集合中的所有单词的IMG_257。将单词在文本中出现的情况用一个单词-文本矩阵表示，记作X

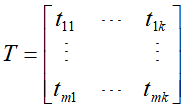


元素IMG_256表示单词IMG_257在文本IMG_258中出现的频数或权值。而权值通常用**单词频率-逆文本频率（term frequency-inverse document frequency，TF-IDF）**表示：



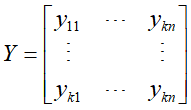
式子中IMG_256是单词IMG_257在文本IMG_258中出现的频数，IMG_259是文本中的单词总数，IMG_260是含有单词的文本数，IMG_261是文本总数n。这个式子能直观表示，一个单词在一个文本中出现的频数越高，这个单词在这个文本中的重要度就越高；一个单词在整个文本集中出现的出现的文本数越少，这个单词就越能表示其所在文本的特点。

**话题向量空间模型**：其先定义了话题向量空间，用来表示话题和单词之间的关系，设单词-话题矩阵为T（共有k个话题）



元素IMG_256表示单词IMG_257在话题IMG_258中出现的权值，权值越大，该单词在此话题中的重要度就越高。

然后，文本在话题空间的表示可定义为话题-文本矩阵Y



这样一来，在单词向量空间的文本IMG_256可以通过它在话题空间中的向量IMG_257近似表示，具体地由k个话题向量以IMG_258为系数的线性组合近似表示：

IMG_256

即IMG_256，而这也正是**潜在语义分析**。在话题向量空间中，文本的相似度可由它们在话题空间中所表示的对应向量的内积IMG_257进行表示。