









**2.2 向量空间模型（VSM）**

向量空间模型中将文档表达为一个矢量，看作向量空间中的一个点

(1) 词权重

用来衡量每个词的重要性

(2) 词频(tf)

如果一个词在一个中出现的次数越多，那么这个词在描述这个句子的含义方面贡献度越大。计算方式(两种)：



(3) 逆文档频率(idf)

如果一个词在越多的文档中出现过，那个这个词对某一个文档区别于其他文档的贡献度就越小，可以用逆文档频率(idf)来度量对文档区分的贡献度。

文档频率（df）：对于某个词来说，包含这个词的文档的数量，即称为文档频率。

计算方式：



**2.3概率统计检索模型**

概率统计检索模型(Probabilistic Retrieval Model)是另一种普遍使用的信息检索算法模型，它应用文档与查询相关的概率来计算文档与查询的相似度。

通常利用检索单元作为线索，通过统计得到每个检索单元在相关的文档集（对应于某询）中出现和不出现的概率以及其在与该查询不相关的文档集中出现和不出现的概率，最终，利用这些概率值，计算文档与查询的相似度。