参考链接：

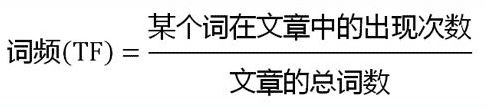
<http://blog.csdn.net/sangyongjia/article/details/52440063>

<http://blog.csdn.net/eastmount/article/details/49898133>

TF-IDF（term frequency–inverse document frequency）是一种用于资讯检索与资讯探勘的常用加权技术。TF-IDF是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。TF-IDF加权的各种形式常被搜寻引擎应用，作为文件与用户查询之间相关程度的度量或评级。

在一份给定的文件里，****词频 (term frequency, TF)**** 指的是某一个给定的词语在该文件中出现的频率。

 词频（Term Frequency，简称TF）表示特征词出现的次数除以文章总词数：



****逆向文件频率 (inverse document frequency, IDF)**** 是一个词语普遍重要性的度量。某一特定词语的IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到。

****IDF为所有文档的数目除以包含该词语的文档数目的对数值****

IMG_256

其中，|D|表示所有文档的数目，|w∈d|表示包含词语w的文档数目。

由于“是”“的”“这”等词经常会出现，故需要IDF值来降低其权值。所谓降维，就是降低维度。具体到文档相似度计算，就是减少词语的数量。常见的可用于降维的词以功能词和停用词为主（如："的"，"这"等），事实上，采取降维的策略在很多情况下不仅可以提高效率，还可以提高精度。如果包含词条w的文档越少，IDF越大，则说明词条w具有很好的类别区分能力。

因为每个词语对实体的贡献度不同，所以需要对这些词语赋予不同的权重。计算词项在向量中的权重方法——TF-IDF。  
     它表示TF（词频）和IDF（倒文档频率）的乘积：

IMG_256

某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的TF-IDF。因此，TF-IDF倾向于过滤掉常见的词语，保留重要的词语。TF-IDF计算权重越大表示该词条对这个文本的重要性越大。