## ****1、TF-IDF算法介绍****

 TF-IDF（term frequency–inverse document frequency，词频-逆向文件频率）是一种用于信息检索（information retrieval）与文本挖掘（text mining）的常用加权技术。

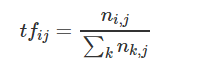
TF-IDF是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

TF-IDF的主要思想是：如果某个单词在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。

1）TF是词频(Term Frequency)

 词频（TF）表示词条（关键字）在文本中出现的频率。

 这个数字通常会被归一化(一般是词频除以文章总词数), 以防止它偏向长的文件。

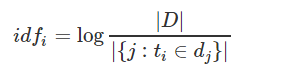


其中 ni,j 是该词i在文件 j 中出现的次数，分母则是文件 j 中所有词汇出现的次数总和；

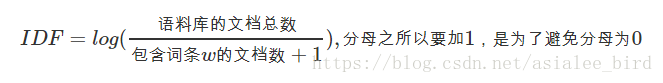
（2） IDF是逆向文件频率(Inverse Document Frequency)

 逆向文件频率 (IDF) ：某一特定词语的IDF，可以由总文件数目除以包含该词语的文件的数目，再将得到的商取对数得到。

如果包含词条t的文档越少, IDF越大，则说明词条具有很好的类别区分能力。



 其中，|D| 是语料库中的文件总数。 |{j:ti∈dj}| 表示包含词语 ti 的文件数目（即 ni,j≠0 的文件数目）。如果该词语不在语料库中，就会导致分母为零，因此一般情况下使用 1+|{j:ti∈dj}|



（3）TF-IDF实际上是：TF \* IDF

 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。

注：  TF-IDF算法非常容易理解，并且很容易实现，但是其简单结构并没有考虑词语的语义信息，无法处理一词多义与一义多词的情况。

## ****2、TF-IDF应用****

     （1）搜索引擎；（2）关键词提取；（3）文本相似性；（4）文本摘要

## ****3、TF-IDF算法的不足****

TF-IDF 采用文本逆频率 IDF 对 TF 值加权取权值大的作为关键词，但 IDF 的简单结构并不能有效地反映单词的重要程度和特征词的分布情况，使其无法很好地完成对权值调整的功能，所以 TF-IDF 算法的精度并不是很高，尤其是当文本集已经分类的情况下。

在本质上 IDF 是一种试图抑制噪音的加权，并且单纯地认为文本频率小的单词就越重要，文本频率大的单词就越无用。这对于大部分文本信息，并不是完全正确的。IDF 的简单结构并不能使提取的关键词， 十分有效地反映单词的重要程度和特征词的分布情 况，使其无法很好地完成对权值调整的功能。尤其是在同类语料库中，这一方法有很大弊端，往往一些同类文本的关键词被盖。

TF-IDF算法实现简单快速，但是仍有许多不足之处：

（1）没有考虑特征词的位置因素对文本的区分度，词条出现在文档的不同位置时，对区分度的贡献大小是不一样的。

（2）按照传统TF-IDF，往往一些生僻词的IDF(反文档频率)会比较高、因此这些生僻词常会被误认为是文档关键词。

（3）传统TF-IDF中的IDF部分只考虑了特征词与它出现的文本数之间的关系，而忽略了特征项在一个类别中不同的类别间的分布情况。

（4）对于文档中出现次数较少的重要人名、地名信息提取效果不佳。