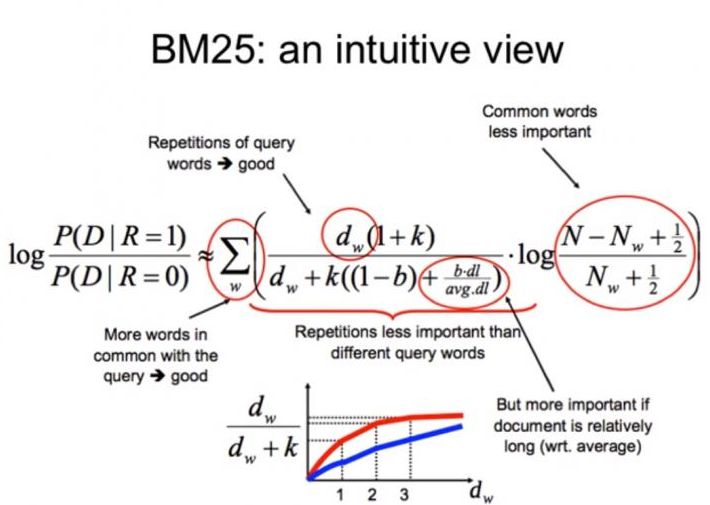
BM25算法原理与实现

# 1 什么是BM25

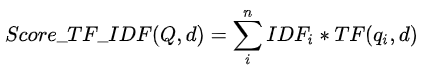
BM25是T基于TF-IDF改进的文本匹配算法（BM是Best Matching的意思），被称为下一代的TF-IDF。

BM25算法通常用来作搜索相关性评分。一句话概况其主要思想：对Query进行语素（term）解析，生成语素qi；然后，对于每个搜索结果D，计算每个语素qi与D的相关性得分，最后，将qi相对于D的相关性得分进行加权求和，从而得到Query与D的相关性得分。



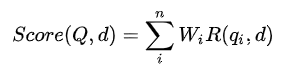
# 2 BM25算法原理

由于BM25算法是TF-IDF算法的优化，所以先回顾一下，如何按照TF-IDF计算相似度？



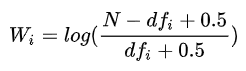
一个用户query与文档库中一个文档的TF-IDF相似度，是将用户输入的query中每一个词与文档库中一篇文档的TF-IDF值求和。

BM25算法与TF-IDf算法十分相似，其公式如下：

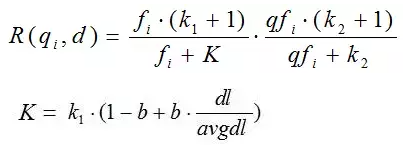


其中，Q为用户输入的query，d是文档库中的一篇文档，n是用户输入的query中包含的总词数，q\_{i}为用户输入的query中的第i个词，W\_{i}为该词的权重，R(q\_{i},d)为该词与文档库中一篇文档的相关性分数。

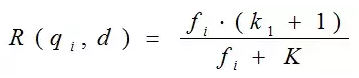
词权重W\_{i}的计算公式如下：



其中，N表示文档库中总文档的个数，dfi表示包含词汇qi的文档的个数。根据W\_{i}的定义可以看出，对于给定的文档集合，包含了qi的文档数越多，qi的权重则越低。也就是说，当很多文档都包含了qi时，qi的区分度就不高，因此使用qi来判断相关性时的重要度就较低。

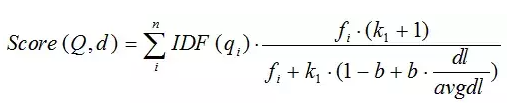


其中，k1，k2，b为调节因子，通常根据经验设置，一般k1=2，b=0.75；fi为qi在d中的出现频率，qfi为qi在Query中的出现频率。dl为文档d的长度，avgdl为所有文档的平均长度。由于绝大部分情况下，qi在Query中只会出现一次，即qfi=1，因此公式可以简化为：



从K的定义中可以看到，参数b的作用是调整文档长度对相关性影响的大小。b越大，文档长度的对相关性得分的影响越大，反之越小。而文档的相对长度越长，K值将越大，则相关性得分会越小。这可以理解为，当文档较长时，包含qi的机会越大，因此，同等fi的情况下，长文档与qi的相关性应该比短文档与qi的相关性弱。

综上，BM25算法的相关性得分公式可总结为：



从上面的分析过程可以看出， W\_{i}相当于TF-IDF算法中的IDF，R(q\_{i},d)相当于TF-IDF算法中的TF；只不过BM25对这两个指标进行了优化。

可以看到，通过使用不同的语素分析方法、语素权重判定方法，以及语素与文档的相关性判定方法，我们可以衍生出不同的搜索相关性得分计算方法，这就为我们设计算法提供了较大的灵活性。

在实际使用BM25算法的过程中，可能出现负值，并且正值都很小。这有可能是因为：（1）文档库库较小；

（2）有一些词汇出现在很多标准问句中，导致其idf小于0，在做累加时，导致相似度一些为负值，并且正值都很小，此时就会出现负数的结果。

因此，在使用BM25算法时，最好先去掉停用词。

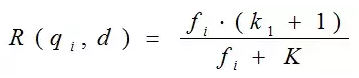
# 3 TF-IDF与BM25对比

传统的TF值理论上是可以无限大的。而BM25与之不同，它在TF计算方法中增加了一个常量k，用来限制TF值的增长极限。下面是两者的公式：

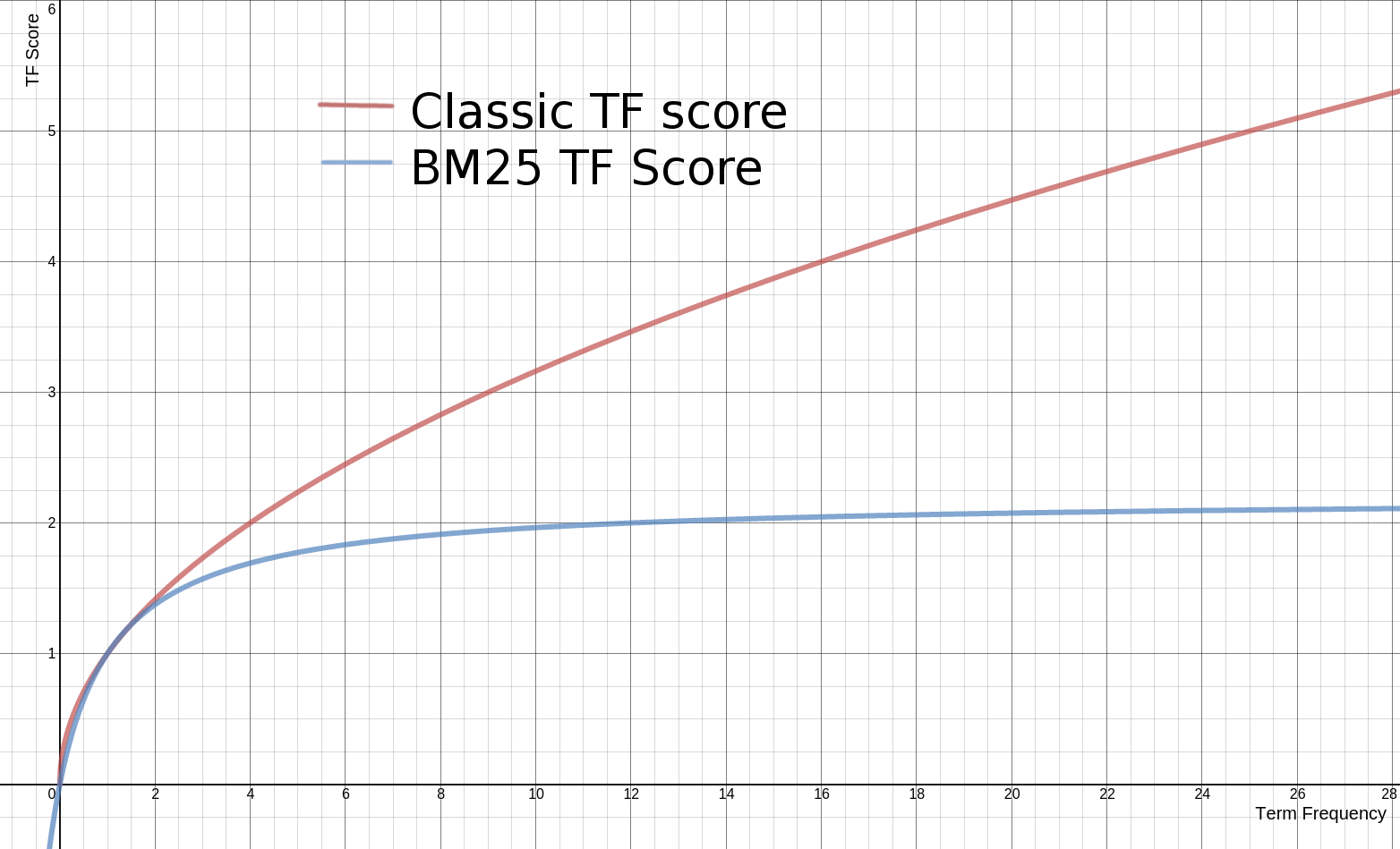
传统TF Score公式如下：

图片

BM25的 TF Score公式如下：

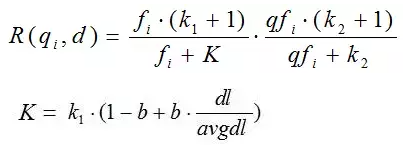


下面是两种计算方法中，词频对TF Score影响的走势图。从图中可以看到，当tf增加时，TF Score跟着增加，但是BM25的TF Score会被限制在0~k+1之间。它可以无限逼近k+1，但永远无法触达它。这在业务上可以理解为某一个因素的影响强度不能是无限的，而是有个最大值，这也符合我们对文本相关性逻辑的理解。在Lucence的默认设置里，k＝1.2，使用者可以修改它。

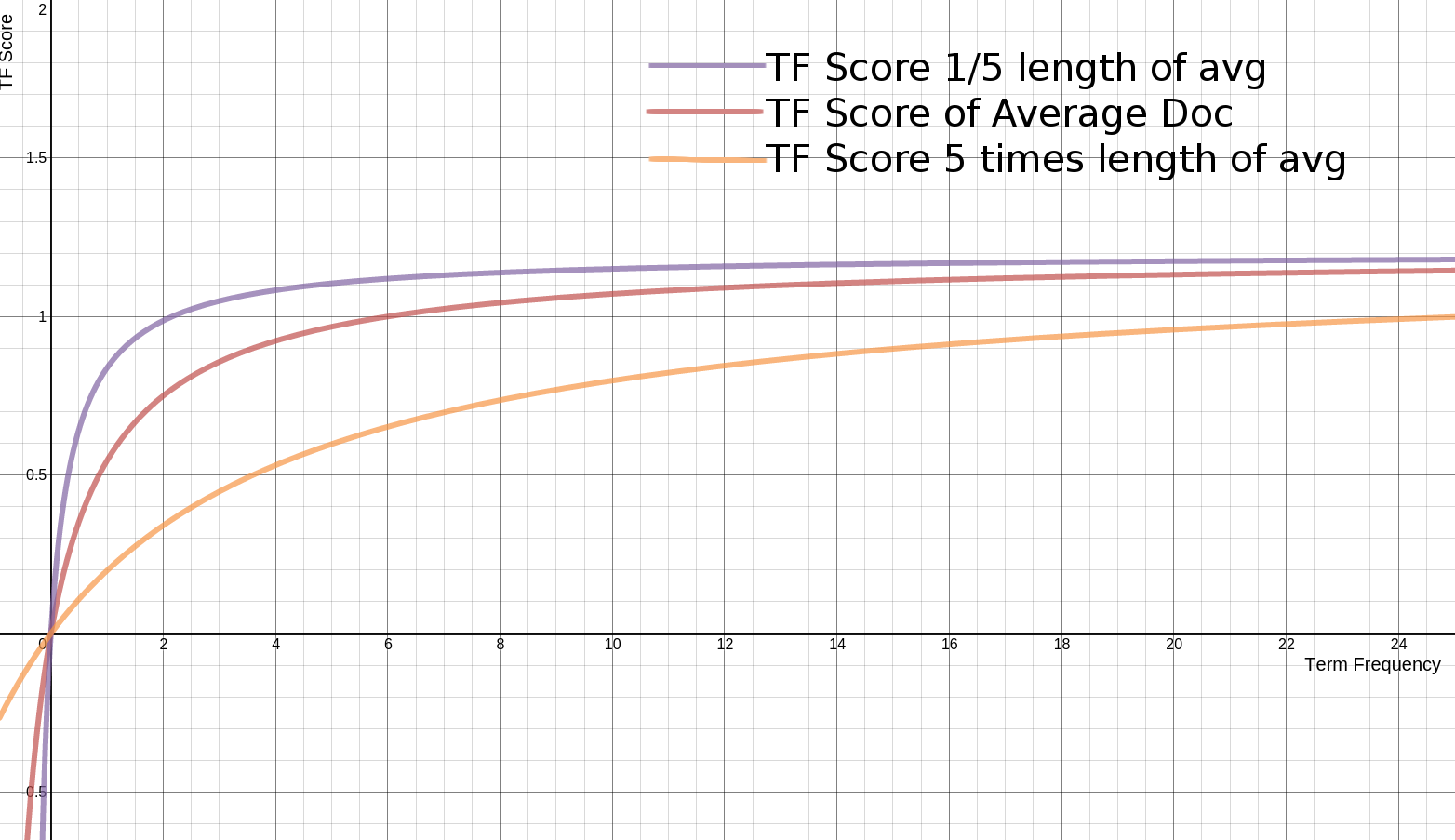


# 4 BM25如何对待文档长度

BM25还引入了平均文档长度的概念，单个文档长度对相关性的影响力与它和平均长度的比值有关系。BM25的TF公式里，除了k外，引入另外两个参数：L和b。L是文档长度与平均长度的比值。如果文档长度是平均长度的2倍，则L＝2。b是一个常数，它的作用是规定L对评分的影响有多大。加了L和b的公式变为：

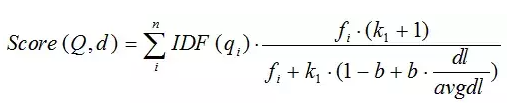


下面是不同L的条件下，词频对TF Score影响的走势图：



从图上可以看到，文档越短，它逼近上限的速度越快，反之则越慢。这是可以理解的，对于只有几个词的内容，比如文章“标题”，只需要匹配很少的几个词，就可以确定相关性。而对于大篇幅的内容，比如一本书的内容，需要匹配很多词才能知道它的重点是讲什么。

上文说到，参数b的作用是设定L对评分的影响有多大。如果把b设置为0，则L完全失去对评分的影响力。b的值越大，L对总评分的影响力越大。此时，相似度最终的完整公式为：



传统的TF-IDF是自然语言搜索的一个基础理论，它符合信息论中的熵的计算原理，虽然作者在刚提出它时并不知道与信息熵有什么关系，但你观察IDF公式会发现，它与熵的公式是类似的。实际上IDF就是一个特定条件下关键词概率分布的交叉熵。

BM25在传统TF-IDF的基础上增加了几个可调节的参数，使得它在应用上更佳灵活和强大，具有较高的实用性。

# 5 实践

参考：

<https://github.com/jpegbert/search_coding/tree/master/bm25>

参考

<https://mp.weixin.qq.com/s/LSHtjnMV2e5lT0drUqRMrw>

<https://my.oschina.net/stanleysun/blog/1617727>

<https://mp.weixin.qq.com/s/TGIPFcmvixntpamWrEJIlA>

<https://mp.weixin.qq.com/s/0oNosNxVW_1ir4g3gTQ1fQ>

<https://mp.weixin.qq.com/s/QPtO78-FbcjkHwbCeebdSg>