文本相似度计算基础算法BM25

原创 纳纳 网络安全探秘 2019-09-28

接着文本相似度计算基础算法（一），继续介绍BM25算法。

BM25是基于TF-IDF算法的改进方法，也是一种统计学的方法，算法简单易懂。相似度计算方法如下：

$$\mathrm{Score}(\mathrm{Q},\mathrm{d}) = \sum_{i}^{n} W_i \cdot R(q_i, d)$$

其中，d表示相关文档，Q是query查询中的所有的词，q_i是查询语句中的一个关键词词，W_i是这个词的权重，$R(q_i,d)$是这个词和文档的相关度值的查询权重，n是Q的词总数。

W_i默认是IDF的值，上一节已经介绍过，计算公式如下：

$$\mathrm{IDF}(q_i) = log\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}$$

其中，N表示所有文档数，$n(q_i)$表示包含查询词q_i的文档数，0.5是避免$n(q_i)$为0的情况。大致的意思是关键词出现频率越小说明越重要。

$R(q_i,d)$的计算公式如下：

$$R(q_i,d) = \frac{f_i \cdot (k_1+1)}{f_i + K} \cdot \frac{qf_i \cdot (k_2+1)}{qf_i + k_2}$$

$$\mathrm{K} = k_1 \cdot \left(1 - b + b \cdot \frac{dl}{avgdl}\right)$$

其中k1,k2，b都是调节因子，一般k1=2,k2=1,b=0.75，其中fi表示关键字出现的次数，dl是文档的长度，avgdl是文档的平均长度（为了防止K过大，归一化处理）,qfi是关键字在query查询Q中出现的次数。
$R(q_i,d)$乘积左边表示关键词在文档中的关系，右边表示关键词在查询语句中的关系。一般情况下，关键词在查询语句只会出现一次，所以qfi可以看成1，公式右边经过简化后也等于1，忽略不计。

相似度计算总公式经过分析后，最终计算式如下：

$$\text{Score}(\text{Q},\text{d}) = \sum_{i}^{n} IDF(q_i) \cdot \frac{f_i \cdot (k_1+1)}{f_i + k_1 \cdot \left(1 - b + b \cdot \frac{dl}{avgdl}\right)}$$

从上面公式可以看出，影响BM25公式的因数有：

1. idf，idf得分越高，分值越高。
2. tf，tf得分越高，分值越高。
3. dl，文档长度越长，分值越低。