



文本处理实践
第三讲

TF-IDF : 词的权重计算

1

TF-IDF原理简介

2

利用TF-IDF 进行文本关键词提取

- 什么是TF-IDF？

词的权重计算的重要方法之一。

TFIDF算法建立在这样一个假设之上：对区别文档最有意义的词，应该是那些在文档中出现频率高，而在整个文档集合的其他文档中出现频率少的词语。

TF（term frequency）

例：查找关于“原子能的应用”的网页。在某个一共有一千个词的网页j中，“原子能”、“的”和“应用”分别出现了2次、35次和5次，那么它们的词频就分别是0.002、0.035和0.005。

W_i = “原子能”

- 什么是TF-IDF？

的权重计算的重要方法之一。

STOPWORDS

例。词“的”占了总词频的 80% 以上，而它对确定网页的主题几乎没有用。我们称这种词叫“应删除词”(Stopwords)，也就是说在度量相关性时不应考虑它们的频率。在汉语中，stopwords还有“是”、“和”、“中”、“地”、“得”等。

- 什么是TF-IDF？

权重计算的重要方法之一。

IDF (Inverse document frequency) 逆文档频率

例：在汉语中，“应用”是个很通用的词，而“原子能”是个很专业的词，后者在相关性排名中比前者重要。

因此我们需要给每一个词一个权重，这个权重的设定必须满足下面两个条件：

1. 一个词预测主题能力越强，权重就越大，反之，权重就越小。我们在网页中看到“原子能”这个词，或多或少地了解网页的主题。我们看到“应用”一次，对主题基本上还是一无所知。因此，“原子能”的权重就应该比应用大。
2. Stopwords的权重应该是零。

- 什么是TF-IDF？

词的权重计算的重要方法之一。

IDF

假定一个关键词 w 在 D_w 个网页中出现过， D_w 越大， w 的权重越小，反之亦然。

它的公式为 $\log(D / D_w)$ 其中 D 是全部文档的数量。

例：网页搜索，我们假定中文网页数是 $D = 10$ 亿，“的”在所有的网页中都出现，即 $D_w = 10$ 亿，那么它的 $IDF = \log(10\text{亿}/10\text{亿}) = \log(1) = 0$ 。假如专用词“原子能”在两百万个网页中出现，即 $D_w = 200$ 万，则它的权重 $IDF = \log(500) = 6.2$ 。又假定通用词“应用”，出现在五亿个网页中，它的权重 $IDF = \log(2)$ 则只有 0.7。

- 什么是TF-IDF？

词的权重计算的重要方法之一。

TF-IDF 基本公式

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

其中, n_{ij} 表示词 w_i 在文件 d_j 中出现的次数，分母表示在文件 d_j 中所有字出现的次数之和。

$$IDF_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

- 什么是TF-IDF？

词的权重计算的重要方法之一。

TF-IDF 基本公式

$$IDF_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

$|D|$ 表示语料库中的文档的总数。 $|\{j:t_i \in d_j\}|$ 表示 包含词语 t_i 的文档的数目。如果这个词不在文档中，会导致除数为零。因此，一般情况下使用 $1+ |\{j:t_i \in d_j\}|$ 。这个函数被证明和符合一个特定条件下关键词的概率分布的交叉熵。

- 什么是TF-IDF？

词的权重计算的重要方法之一。

TF-IDF 基本公式

$$TF-IDF_{i,j} = tf_{i,j} \times idf_i$$

某一特定文件内的高频词语，同时又是整个文件集合中的低频词，可以产生高的TF-IDF

2

实验一：利用TF-IDF 进行文本关键词提取

- 数据集
- 处理基本流程
- 处理结果分析

数据集：若干网页

清洗数据：观察数据，根据算法需要，丢弃数据中不重要的信息。

文本分词

计算TF

计算IDF

对计算结果进行排序，输出topK

作业说明：

- 每位学生独立完成
- 针对给定的网络爬取文档，完成文档清洗，提取其中的文本部分，给出每个文档中每个词的TF-IDF值，排序后输出。
- 完成实验报告
- 数据集、说明文档的下载网址：<https://pan.baidu.com/s/1qYo9Xr2>
- 需要自己编码实现，不能调用其他开发包

作业提交：

将以下内容打包(zip或rar)提交

- 结果文件
- 关键代码（调用的软件包如果太大可不提交，做出说明即可）
- 1~3页的作业报告
 - 学号，姓名
 - 实验过程、实验结果分析
 - 其他值得交代的事情
- 提交邮箱：736781877qq.com
- 提交截止时间：2017年3月20日晚12：00