人工智能之机器学习

关键词提取（TF-IDF）

产品研发中心 -- 李军

案例分析 – 中国的蜜蜂养殖

- 假定现在有一篇长文《中国的蜜蜂养殖》，我们准备用计算机提取它的关键词

1、一个容易想到的思路，就是找到出现次数最多的词。如果某个词很重要，它应该在这篇文章中多次出现。

2、出现次数最多的词是“的”、“是”、“在”这一类最常用的词。还有“停用词”（stop words），表示对找到结果毫无帮助、必须过滤掉的词、以及标点符号。

3、如果某个词比较少见，但是它在这篇文章中多次出现，那么它很可能就反映了这篇文章的特性，正是我们所需要的关键词。

案例分析 – 中国的蜜蜂养殖

- 第一步采用IK分词或者庖丁分词，然后再计算词频

$$词频(TF) = 某个词在文章中的出现次数$$

考虑到文章有长短之分，为了便于不同文章的比较，进行"词频"标准化。

$$词频(TF) = \frac{某个词在文章中的出现次数}{文章的总词数}$$

或者

$$词频(TF) = \frac{某个词在文章中的出现次数}{该文出现次数最多的词的出现次数}$$

案例分析 – 中国的蜜蜂养殖

- 第二步计算逆文档频率

这时，需要一个语料库（corpus），用来模拟语言的使用环境。

$$\text{逆文档频率(IDF)} = \log\left(\frac{\text{语料库的文档总数}}{\text{包含该词的文档数} + 1}\right)$$

如果一个词越常见，那么分母就越大，逆文档频率就越小越接近0。分母之所以要加1，是为了避免分母为0（即所有文档都不包含该词）。log表示对得到的值取对数。

案例分析 – 中国的蜜蜂养殖

- 第三步计算TF-IDF值

$$\text{TF}-\text{IDF}=\text{词频(TF)}\times\text{逆文档频率（IDF）}$$

可以看到，TF-IDF与一个词在文档中的出现次数成正比，与该词在整个语言中的出现次数成反比，所以提取关键词其实就是计算出文档的每个词的TF-IDF值，然后按降序排列，取排在最前面的几个词。

感谢您的聆听！

Thank you for your time!

凯通科技
www.ctt-net.com

用软件重新定义世界，让世界更加智能互联