TextRank原理与实践

# 1、原理

TextRank采用图的思想，将文档中的词表示成一张无向有权图，词为图的节点，词之间的联系紧密程度体现为图的边的权值。计算词的权重等价于计算图中节点的权重。提取关键字，等价于找出图中权重排名TopK的节点。

TextRank背后体现的思想为：与其他词关联性强的词，越重要。通俗一点就是：围着谁转，谁就重要。就像大家基本都会围着领导转一样。

# 2、TF-IDF与TextRank算法的比较

1. 从算法原理上来看，基础都是词频统计，只是TD-IDF通过IDF来调整词频的权值，而TextRank通过上下文的连接数来调整词频的权值。TextRank通过滑动窗口的方式，来实现词的位置对词的权值的影响。
2. TD-IDF计算简单，运行性能更好。

TextRank基于图计算来提取关键词，需要进行迭代，速度比TF-IDF慢，但不需要通过输入多 篇文档来提取关键词。

TextRank是把一篇文档构建成无向图，图中的节点就是词语，图上的边就是共现词之间的连 接。

共现词通过滑动窗口来确定，共现词之间用边相连，而边上的权重可以使用共现词的相似 度。

jieba提供了用TextRank提取关键词的函数，但是边的权重是词共现的频率。

这样做其实比较粗糙，我们可以使用基于词向量计算的相似度得分作为权重，来进行迭代。

我这不知不觉又给自己安排了任务，文章名字都想好了：

《TextRank提取关键词：我也是改过jieba源码的人！》

TextRank也需要先去掉停用词。

参考：

<https://mp.weixin.qq.com/s/7sbMdNxbIvp_mST7PB4KUw>

<https://mp.weixin.qq.com/s/xOAfDclb49RaqIWN6UOtPA>