关键词提取算法说明

# 引言

关键字提取是在一篇文档中找到最能体现这篇文档几个关键字，可以帮助读者判断是否对这篇文章感兴趣，也可以用作关键词索引。目前，常用的关键词提取方法有词频统计，将词频最高的词条视为关键字，这种方法就算是过滤掉垃圾词汇，得到的结果也不尽人意，更为科学的方式是采用TF-IDF词条权重计算方法，此算法的主要思想是当一个词条在本篇文档中出现的频率越高，在全部文章中出现的频率越低，这个词条越能代表这篇文档，TF-IDF算法在文本关键词提取中有不错的效果。文档[1]介绍了一种使用Python第三方库scikit-learn工具计算文本TF-IDF值的方法，起到了不错的效果。

本文以校园信息关键词为例，介绍使用TF-IDF算法的文本关键字提取方案，并给出了算法评价。

# 算法介绍

* 1. 概述

针对中文文本的关键词提取是在文章词条基础上进行的，因此首先需要将文章句子切割成词条。将文章切割为词条后就可以使用Python第三方库scikit-learn工具计算文本TF-IDF值，然后统计词条的TF-IDF值选择前几个作为文章的关键字，本文选择的是前五个。

关键字提取可以分为三步：分词、TF-IDF值计算、关键词统计。

* 1. 分词

请参考[2]。

* 1. TF-IDF值计算

参考文档[2]给出了TF-IDF值的概念和算法。

* 1. 关键词统计

得到文章中每个词条的TF-IDF值后，统计每篇文章的TF-IDF值的前五位作为这篇文章的关键词。

# 算法评价和改进

经过测试，本算法可以给出一篇文章的关键字，但是还不能过滤掉一些没有意义的名词，比如地名、官职名等，所以给出的结果不尽人意。

下一步的改进方法是过滤掉官职、地名等一些没有意义的词。但这样做的工作量很大，需要思考如果用更高效的算法过滤掉没有实际意义的词。

# 参考文档

[1]使用scikit-learn工具计算文本TF-IDF值<http://www.voidcn.com/article/p-bbabkmsv-pt.html>.

[2]分类算法说明