jieba模块常用方法及使用

|  |  |
| --- | --- |
| **方法** | **使用说明** |
| jieba.cut() | 分词，返回一个 可迭代的 generator |
| jieba.lcut() | 分词，返回一个list |
| jieba.cut\_for\_search | 用于搜索引擎构建倒排索引的分词，返回一个 可迭代的 generator |
| jieba.lcut\_for\_search | 用于搜索引擎构建倒排索引的分词，返回一个list |
| jieba.Tokenizer(dictionary=DEFAULT\_DICT) | 新建自定义分词器，可用于同时使用不同词典 |
| jieba.load\_userdict(file\_name) | 添加指定文件名的词典 |
| add\_word(word, freq=None, tag=None) | 向词典中添加一个词 |
| del\_word(word) | 在词典中删除一个词 |
| suggest\_freq(segment, tune=True) | 调节单个词语的词频，使其能（或不能）被分出来 |
| jieba.analyse.extract\_tags(sentence, topK=20, withWeight=False, allowPOS=()) | 基于 TF-IDF 算法的关键词抽取 |
| jieba.analyse.TFIDF(idf\_path=None) | 新建 TFIDF 实例 |
| jieba.analyse.set\_idf\_path(file\_name) | 关键词提取所使用逆向文件频率（即IDF）文本语料库可以切换成自定义语料库的路径 |
| jieba.analyse.set\_stop\_words(file\_name) | 关键词提取所使用停止词（即Stop Words）文本语料库可以切换成自定义语料库的路径 |
| jieba.posseg.POSTokenizer(tokenizer=None) | 新建自定义分词器 |
| jieba.tokenize() | 返回词语在原文的起止位置 |