jieba—用于分词的第三方库

安装jieba：

pip install jieba

导入库：

import jieba

三种分词模式：

1.精确模式：将字符串切分为词组，适合文本分析。默认模式

2.全模式：罗列字符串的所有子词组，单字会有重复

3.搜索引擎模式：在精确模式拆分后，再次对长词进行拆分

方法：

load\_userdict(file)：加载自定义词组文件，file可以为文本路径或file-like对象

cut(str,cut\_all=False,hmm=True)：返回切分的词组迭代器，str代表分割的原始字符串，cut\_all代表是否使用全模式，hmm代表是否使用hmm模式

cut\_for\_search(str,hmm=True)：返回切分的词组迭代器，str代表分割的原始字符串，hmm代表是否使用hmm模式。该方法用于搜索引擎模式

lcut(str,cut\_all=False,hmm=True)：与cut类似，但返回列表

lcut\_for\_search(str,hmm=True)：与cut\_for\_search类似，但返回列表

get\_dict\_file()：字典文件路径

del\_word(word)：从字典删除词组

add\_word(word,freq=None,tag=None)：添加词组到字典

suggest\_freq(segment, tune=False)：建议非拆分词组，segment可以为字符串或字符串tuple

Tokenizer(dictionary=None)：新建自定义分词器，默认使用为jieba.dt