
python
import turtle
turtle.setup(650,350,200,200)
turtle.penup()
turtle.fd(-250)
turtle.pendown()
turtle.pensize(25)
turtle.pencolor("purple")
turtle.seth(-40)
for 1 in range(4):
turtle.circle(40, 80)
turtle.circle(-40, 80)
turtle.circle(40, 80/2)
turtle.fd(40)
turtle.circle(16, 180)
turtle.fd(40 * 2/3)

Python语言程序设计

模块5: jieba库的使用

嵩 天
北京理工大学

jieba库基本介绍

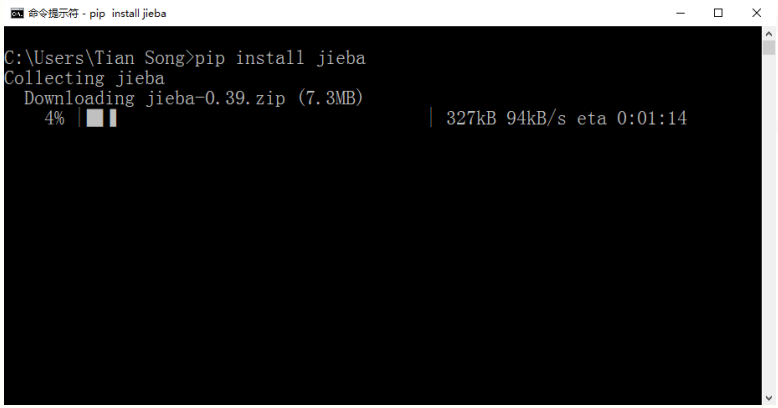
jieba库概述

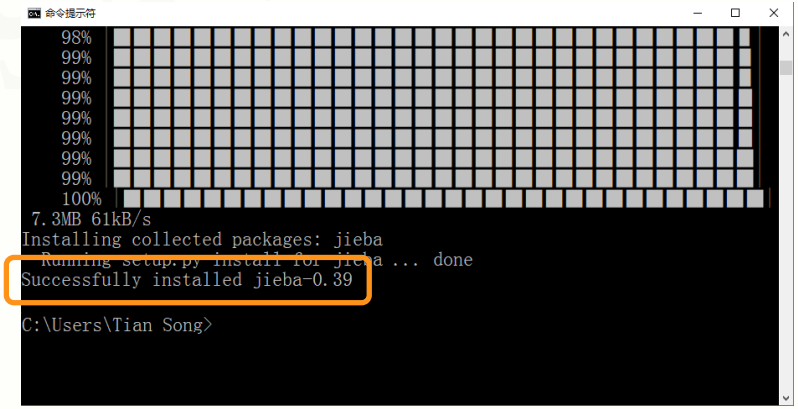
jieba是优秀的中文分词第三方库

- 中文文本需要通过分词获得单个的词语
- jieba是优秀的中文分词第三方库，需要额外安装
- jieba库提供三种分词模式，最简单只需掌握一个函数

jieba库的安装

(cmd命令行) pip install jieba

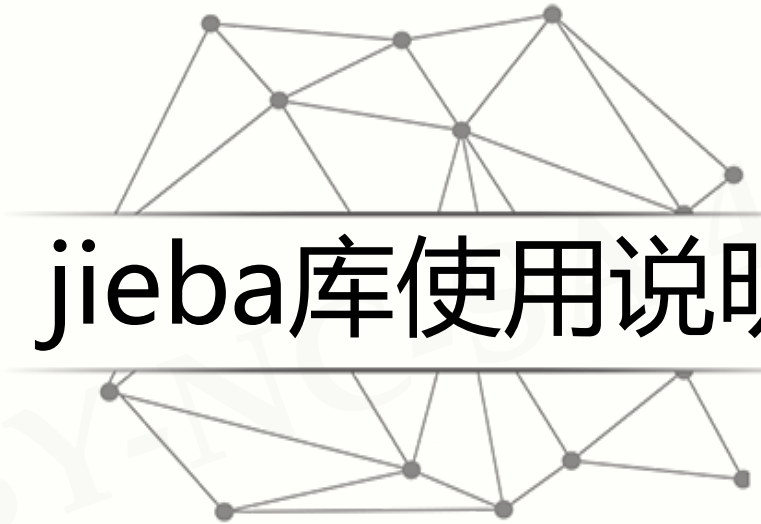




jieba分词的原理

jieba分词依靠中文词库

- 利用一个中文词库，确定中文字符之间的关联概率
- 中文字符间概率大的组成词组，形成分词结果
- 除了分词，用户还可以添加自定义的词组



jieba库使用说明

jieba分词的三种模式

精确模式、全模式、搜索引擎模式

- 精确模式：把文本精确的切分开，不存在冗余单词
- 全模式：把文本中所有可能的词语都扫描出来，有冗余
- 搜索引擎模式：在精确模式基础上，对长词再次切分

jieba库常用函数

函数	描述
jieba.lcut(s)	精确模式，返回一个列表类型的分词结果 >>>jieba.lcut("中国是一个伟大的国家") ['中国', '是', '一个', '伟大', '的', '国家']
jieba.lcut(s, cut_all=True)	全模式，返回一个列表类型的分词结果，存在冗余 >>>jieba.lcut("中国是一个伟大的国家",cut_all=True) ['中国', '国是', '一个', '伟大', '的', '国家']

jieba库常用函数

函数	描述
jieba.lcut_for_search(s)	搜索引擎模式，返回一个列表类型的分词结果，存在冗余 >>>jieba.lcut_for_search(“中华人民共和国是伟大的") ['中华', '华人', '人民', '共和', '共和国', '中华人民共和国', '是', '伟大', '的']
jieba.add_word(w)	向分词词典增加新词w >>>jieba.add_word("蟒蛇语言")

jieba分词要点

jieba.lcut(s)

小花絮

课程的实践平台为何叫Python123?

123 代表了一种符合认知的发展过程

Python123 表达了一种提供最好学习价值的愿望

赶快注册个账号试试吧!