

python

The Python logo, consisting of two interlocking snakes, one blue and one yellow, is positioned below the word "python".

```
import turtle
turtle.setup(650,350,200,200)
turtle.penup()
turtle.fd(-250)
turtle.pendown()
turtle.pensize(25)
turtle.pencolor("purple")

for i in range(4):
    turtle.circle(40, 80)
    turtle.circle(-40, 80)
    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**

(cmd命令行) pip install jieba

```
命令提示符
98%
99%
99%
99%
99%
99%
99%
99%
100%
7.3MB 61kB/s
Installing collected packages: jieba
  Running setup.py install for jieba ... done
Successfully installed jieba-0.39

C:\Users\Tian Song>
```

jieba分词的原理

Jieba分词依靠中文词库

- 利用一个中文词库，确定汉字之间的关联概率
- 汉字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 |
|--|---|
| <code>jieba.lcut(s)</code> | <p>精确模式，返回一个列表类型的分词结果</p> <pre>>>>jieba.lcut("中国是一个伟大的国家")</pre> <pre>['中国', '是', '一个', '伟大', '的', '国家']</pre> |
| <code>jieba.lcut(s, cut_all=True)</code> | <p>全模式，返回一个列表类型的分词结果，存在冗余</p> <pre>>>>jieba.lcut("中国是一个伟大的国家",cut_all=True)</pre> <pre>['中国', '国是', '一个', '伟大', '的', '国家']</pre> |

jieba库常用函数

| 函数 | 描述 |
|---------------------------------------|---|
| <code>jieba.lcut_for_search(s)</code> | <p>搜索引擎模式，返回一个列表类型的分词结果，存在冗余</p> <pre>>>>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']</pre> |
| <code>jieba.add_word(w)</code> | <p>向分词词典增加新词w</p> <pre>>>>jieba.add_word("蟒蛇语言")</pre> |

jieba分词要点

jieba.lcut(s)

