

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.circle(40, 80/2)
    turtle.fd(40)
    turtle.circle(16, 180)
    turtle.fd(40 * 2/3)
```

实例10: 文本词频统计



嵩 天
北京理工大学





"文本词频统计"问题分析

问题分析

文本词频统计

- 需求：一篇文章，出现了哪些词？哪些词出现得最多？
- 该怎么做呢？

英文文本



中文文本

问题分析

文本词频统计

- 英文文本: *Hamlet* 分析词频

<https://python123.io/resources/pye/hamlet.txt>

- 中文文本: 《三国演义》 分析人物

<https://python123.io/resources/pye/threekingdoms.txt>



"Hamlet英文词频统计"实例讲解

```
#CalHamletV1.py
```

```
def getText():
```

```
    txt = open("hamlet.txt", "r").read()
```

```
    txt = txt.lower()
```

```
    for ch in '!"#$%&()*+,-./:;<=>?@[\\]^_`{|}~':
```

```
        txt = txt.replace(ch, " ")
```

```
    return txt
```

```
hamletTxt = getText()
```

```
words = hamletTxt.split()
```

```
counts = {}
```

```
for word in words:
```

```
    counts[word] = counts.get(word,0) + 1
```

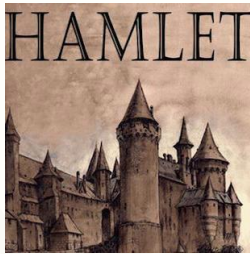
```
items = list(counts.items())
```

```
items.sort(key=lambda x:x[1], reverse=True)
```

```
for i in range(10):
```

```
    word, count = items[i]
```

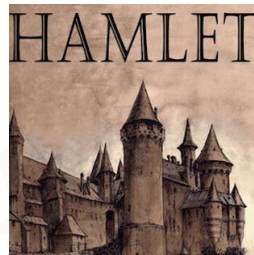
```
    print("{0:<10}{1:>5}".format(word, count))
```



- 文本去噪及归一化
- 使用字典表达词频

>>>

the	1138
and	965
to	754
of	669
you	550
i	542
a	542
my	514
hamlet	462
in	436



- 运行结果由大到小排序
- 观察单词出现次数

准备好电脑，与老师一起编码吧！



"《三国演义》人物出场统计"实例讲解(上)

```
#CalThreeKingdomsV1.py
```

```
import jieba
```

```
txt = open("threekingdoms.txt", "r", encoding="utf-8").read()
```

```
words = jieba.lcut(txt)
```

```
counts = {}
```

```
for word in words:
```

```
    if len(word) == 1:
```

```
        continue
```

```
    else:
```

```
        counts[word] = counts.get(word,0) + 1
```

```
items = list(counts.items())
```

```
items.sort(key=lambda x:x[1], reverse=True)
```

```
for i in range(15):
```

```
    word, count = items[i]
```

```
    print("{0:<10}{1:>5}".format(word, count))
```



- 中文文本分词
- 使用字典表达词频

>>>

曹操 953

孔明 836

将军 772

却说 656

玄德 585

关公 510

丞相 491

二人 469

不可 440

荆州 425

玄德曰 390

孔明曰 390

不能 384

如此 378

张飞 358



- 中文文本分词
- 使用字典表达词频

准备好电脑，与老师一起编码吧！



"《三国演义》人物出场统计"实例讲解(下)

《三国演义》人物出场统计

将词频与人物相关联，面向问题

词频统计



人物统计

```
#CalThreeKingdomsV2.py
import jieba
txt = open("threekingdoms.txt", "r", encoding="utf-8").read()
excludes = {"将军", "却说", "荆州", "二人", "不可", "不能", "如此"}
words = jieba.lcut(txt)
counts = {}
for word in words:
    if len(word) == 1:
        continue
    elif word == "诸葛亮" or word == "孔明曰":
        rword = "孔明"
    elif word == "关公" or word == "云长":
        rword = "关羽"
    elif word == "玄德" or word == "玄德曰":
        rword = "刘备"
    elif word == "孟德" or word == "丞相":
        rword = "曹操"
    else:
        rword = word
    counts[rword] = counts.get(rword, 0) + 1
for word in excludes:
    del counts[word]
items = list(counts.items())
items.sort(key=lambda x: x[1], reverse=True)
for i in range(10):
    word, count = items[i]
    print("{0:<10}{1:>5}".format(word, count))
```



- 中文文本分词
- 使用字典表达词频
- 扩展程序解决问题

>>>

曹操	1451
孔明	1383
刘备	1252
关羽	784
张飞	358
商议	344
如何	338
主公	331
军士	317
吕布	300



- 根据结果进一步优化

隆重发布《三国演义》人物出场顺序前20:

曹操、孔明、刘备、关羽、张飞、吕布、赵云、孙权、
司马懿、周瑜、袁绍、马超、魏延、黄忠、姜维、马岱、
庞德、孟获、刘表、夏侯惇

准备好电脑，与老师一起编码吧！



"文本词频统计"举一反三

```
#CalThreeKingdomsV2.py
import jieba
txt = open("threekingdoms.txt", "r", encoding="utf-8").read()
excludes = {"将军", "却说", "荆州", "二人", "不可", "不能", "如此"}
words = jieba.lcut(txt)
counts = {}
for word in words:
    if len(word) == 1:
        continue
    elif word == "诸葛亮" or word == "孔明":
        rword = "孔明"
    elif word == "关公" or word == "云长":
        rword = "关羽"
    elif word == "玄德" or word == "玄德曰":
        rword = "刘备"
    elif word == "孟德" or word == "丞相":
        rword = "曹操"
    else:
        rword = word
    counts[rword] = counts.get(rword, 0) + 1
for word in excludes:
    del counts[word]
items = list(counts.items())
items.sort(key=lambda x: x[1], reverse=True)
for i in range(10):
    word, count = items[i]
    print("{0:<10}{1:>5}".format(word, count))
```



- 中文文本分词
- 使用字典表达词频
- 扩展程序解决问题

举一反三

应用问题的扩展

- 《红楼梦》、《西游记》、《水浒传》 ...
- 政府工作报告、科研论文、新闻报道 ...
- 进一步呢？ 未来还有词云...

