Python快速入门

嵩 天

实例3：中文词语统计

嵩 天

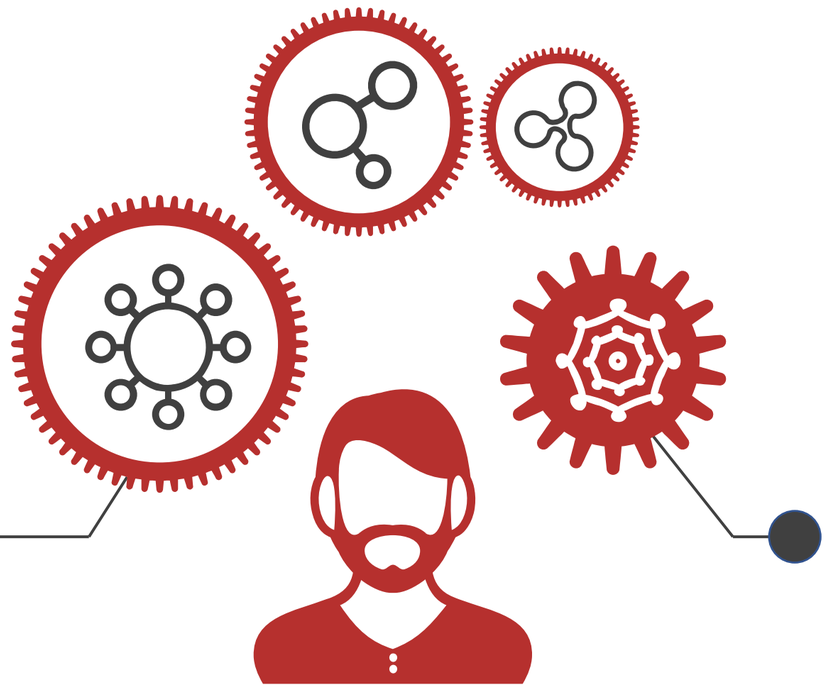
Python快速入门

单元开篇

Python123

单元开篇

(1) "中文词语统计"需求分析



(2) "中文词语统计"实例编写

实例3：中文词语统计

单元开篇

聊天记录词语统计

网络文章词语统计

新闻文章词语统计

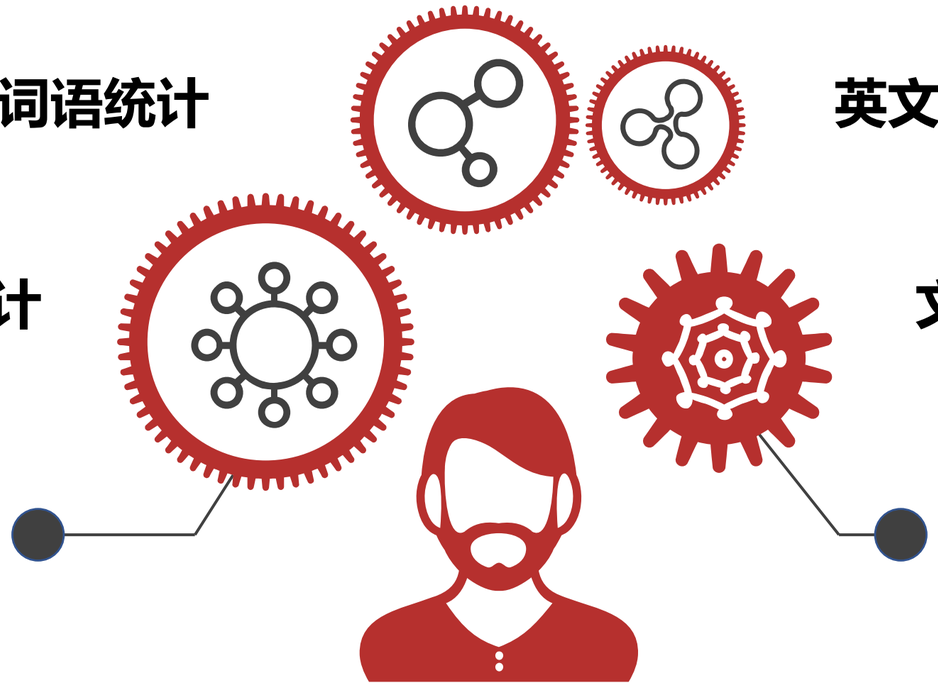
英文文本词语统计

经典名著词语统计

文本重点信息概要

政府文件词语统计

写作风格分析



实例3：中文词语统计

Python快速入门

"中文词语统计"需求分析

Python123

程序需求

统计中文词语出现次数

- 以政府一号文件为例，统计出现的中文词语数量
- 按照一定标准输出，如出现次数等
- 需要解决中文分词问题，如：这是一门好课 -> 这是 一门 好课

程序需求

统计中文词语出现次数

- 以每年政府一号文件为例

程序需求

统计中文词语出现次数

- 输入：2018年一号文件.txt
- 输出：出现次数超过50次的词语，不包括换行

Python快速入门

"中文词语统计"实例编写

Python123

代码演示

```
#WordsCount.py
import jieba
f = open("2018年一号文件.txt", "r", encoding="utf-8")
txt = f.read()
f.close()
ls = jieba.lcut(txt)
d = {}
for w in ls:
    d[w] = d.get(w, 0) + 1
for k in d:
    if d[k] >= 50 and k != "\n":
        print('"{}"出现{}次'.format(k, d[k]))
```

代码演示

```
#WordsCount.py
import jieba
f = open("2018年一号文件.txt", "r", encoding="utf-8")
txt = f.read()
f.close()
ls = jieba.lcut(txt)
d = {}
for w in ls:
    d[w] = d.get(w, 0) + 1
for k in d:
    if d[k] >= 50 and k != "\n":
        print('"{}"出现{}次'.format(k, d[k]))
```

注释

代码演示

```
#WordsCount.py
import jieba
f = open("2018年一号文件.txt", "r", encoding="utf-8")
txt = f.read()
f.close()
ls = jieba.lcut(txt)
d = {}
for w in ls:
    d[w] = d.get(w, 0) + 1
for k in d:
    if d[k] >= 50 and k != "\n":
        print('"{}"出现{}次'.format(k, d[k]))
```

引入外部功能库

代码演示

```
#WordsCount.py
import jieba
f = open("2018年一号文件.txt", "r", encoding="utf-8")
txt = f.read()
f.close()
ls = jieba.lcut(txt)
d = {}
for w in ls:
    d[w] = d.get(w, 0) + 1
for k in d:
    if d[k] >= 50 and k != "\n":
        print('"{}"出现{}次'.format(k, d[k]))
```

打开文件

关闭文件

代码演示

```
#WordsCount.py
import jieba
f = open("2018年一号文件.txt", "r", encoding="utf-8")
txt = f.read()
f.close()
ls = jieba.lcut(txt)
d = {}
for w in ls:
    d[w] = d.get(w, 0) + 1
for k in d:
    if d[k] >= 50 and k != "\n":
        print('"{}"出现{}次'.format(k, d[k]))
```

读入文本

代码演示

```
#WordsCount.py
import jieba
f = open("2018年一号文件.txt", "r", encoding="utf-8")
txt = f.read()
f.close()
ls = jieba.lcut(txt)
d = {}
for w in ls:
    d[w] = d.get(w, 0) + 1
for k in d:
    if d[k] >= 50 and k != "\n":
        print('"{}"出现{}次'.format(k, d[k]))
```

中文分词

代码演示

```
#WordsCount.py
import jieba
f = open("2018年一号文件.txt", "r", encoding="utf-8")
txt = f.read()
f.close()
ls = jieba.lcut(txt)
d = {}
for w in ls:
    d[w] = d.get(w, 0) + 1
for k in d:
    if d[k] >= 50 and k != "\n":
        print('"{}"出现{}次'.format(k, d[k]))
```

建立字典

代码演示

```
#WordsCount.py
import jieba
f = open("2018年一号文件.txt", "r", encoding="utf-8")
txt = f.read()
f.close()
ls = jieba.lcut(txt)
d = {}
for w in ls:
    d[w] = d.get(w, 0) + 1
for k in d:
    if d[k] >= 50 and k != "\n":
        print('"{}"出现{}次'.format(k, d[k]))
```

利用字典
词语统计

代码演示

```
#WordsCount.py
import jieba
f = open("2018年一号文件.txt", "r", encoding="utf-8")
txt = f.read()
f.close()
ls = jieba.lcut(txt)
d = {}
for w in ls:
    d[w] = d.get(w, 0) + 1
for k in d:
    if d[k] >= 50 and k != "\n":
        print('"{}"出现{}次'.format(k, d[k]))
```

遍历结果
设置条件
打印输出

注意事项

相比其他编程语言

- 每行后没有分号;
- 没有begin, end, {, }等表示代码归属的元素，只用缩进表达代码所属关系
- 变量直接使用，无需类型声明
- import可以引入外部功能库

Python 123

Thank you