python是一种脚本编程语言,采用解释执行

python采用严格的缩进来标明程序的格式框架.缩进指每一行代码开始前的空白区域

注释

单行注释:#

多行注释:

python语言允许采用大写字母,小写字母,数字,下划线_,和汉子等字符及其组合给变量命名,但名字的首字符不能是数字,中间不能出现空格,长度没有限制.

python 33个保留字

字符串类型:

单引号''

双引号""

三个单引号''' '''

continue 跳过本次循环,break跳出本层循环

异常

try:

except<异常类型>:

else:

finally:

函数

def <函数名>(<参数列表>)

<函数体>

return <返回值列表>

函数定义时,可以设计可变数量参数,在参数前增加(*)实现

def f(a,*b)

print(type(b))

```
 for n in b:
  a+=b
 return a
print(f(1,2,3,4,5))
```

return 返回多个值

```
def f (a,b)
 return b,a
s=func("sz",2)
print (s)
```

结果:

```
(2,"sz")
```

数据类型

组合数据类型:

\1. 序列类型:

+ 字符串(str)

+ 元组(tuple)

+ 列表(list)

\2. 集合类型

+ 集合(set)

\3. 映射类型

+ 字典(map)

集合的四种操作:

交集(&),并集(|),差集(-),补集(^)

文件的读写

r只读

w覆盖写

x创建写

a 追加写

b二进制

t文本文档

+读写

文件内容的读写

.readall() 读入整个文件内容,返回一个字符串或字节流*

.read(size=-1) 从文件中读入整个文件内容,如果给出参数,读入前size长度的字符串或字节流

.readline(size=-1)从文件中读入一行内容,如果给出参数,读入该行前size长度的字符串或字节流

.readlines(hint=-1)从文件中读取所有行,以每行为元素形成一个列表,如果给出参数,读入hint行

下列不属于html的Tag的是()

A.title B.a C.class D.head

网络爬虫

网络爬虫的应用一般分为两个步骤:

\1. 通过网络链接获取网页内容

\2. 对获得的网页内容进行处理

排除协议

Robots排除协议,也被称为爬虫协议,他是网站管理者表达是否希望爬虫自动获取网络信息意愿的方法.

这是一个简单的HTML页面,请保存为字符串,完成后面的计算要求.

中国,**你好!**.

世界,**大同!**.

\1. 打印head标签的内容

\2. 获取body标签的内容

\3. 获取id为China的标签对象

\4. 获取并打印HTML页面中的中文字符

```py
import requests
from bs4 import BeautifulSoup

r=''

soup=BeautifulSoup(r)
print(soup.head)

```

```py
print(soup.body)
```

```py
print(soup.find(id='china'))
```

```py
for i in r:
 if('\u4e00' <= i <= '\u9fff'):     #检测单个字符是否为汉字
  list_.append(i)   #是则将单个字符加入列表
print(''.join(list_))  #以空字符作为分割符转换为str形式

```