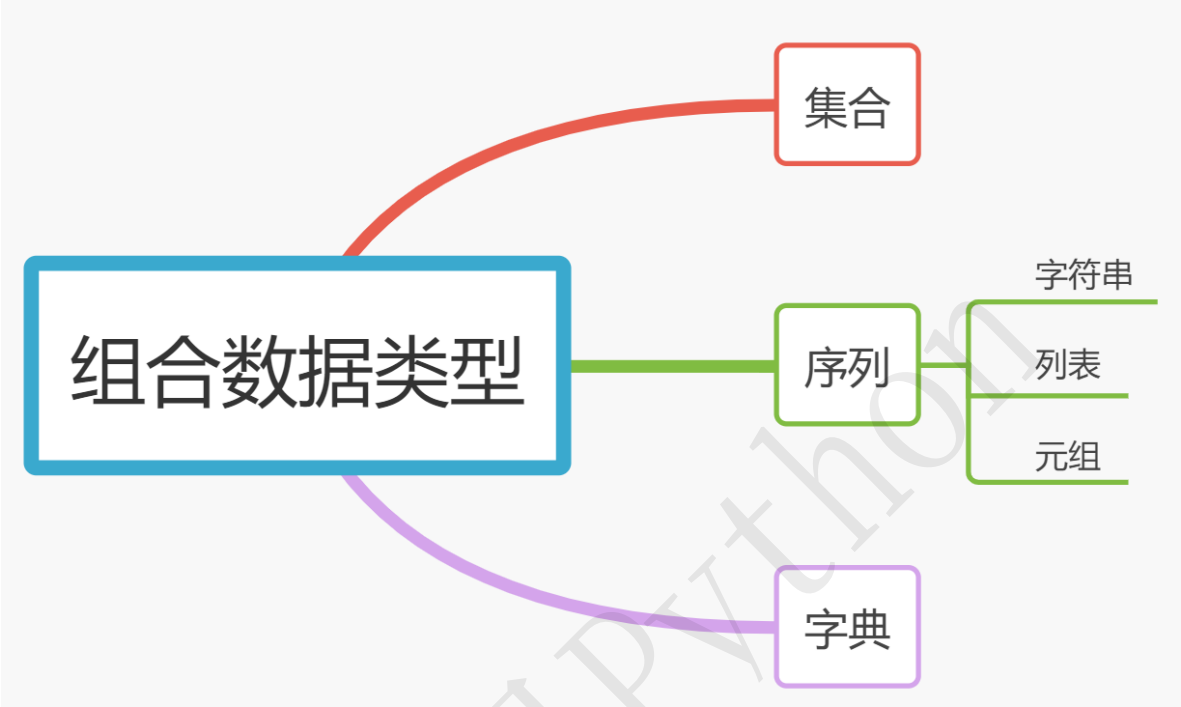
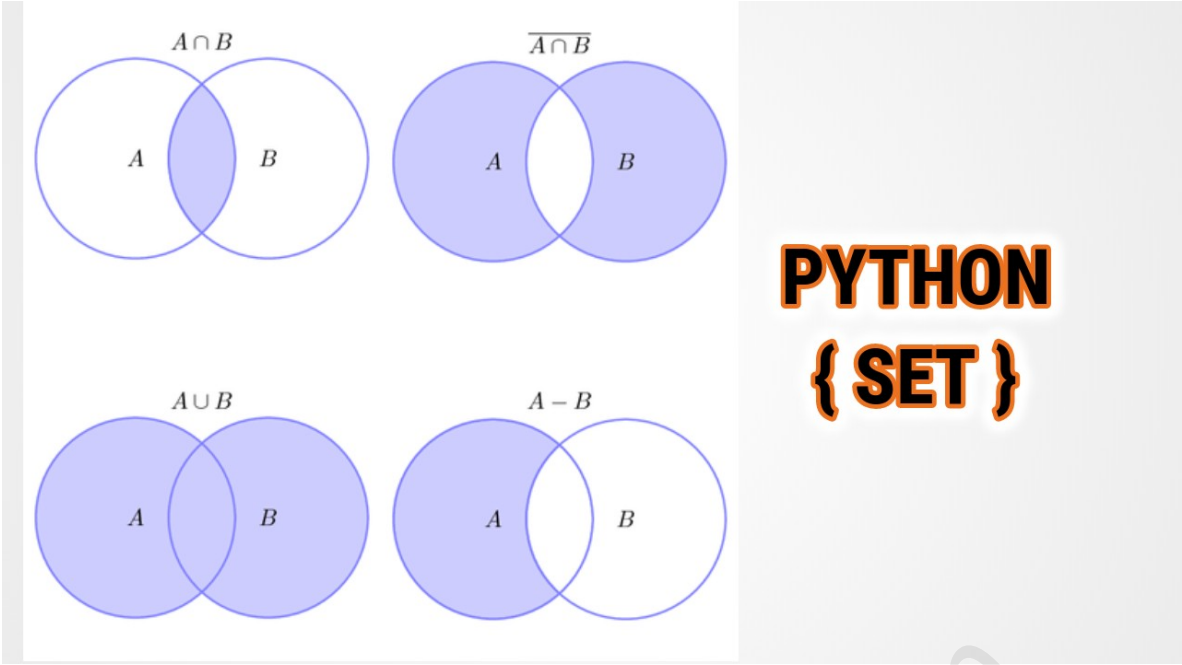
组合数据类型之集合

> 同济子豪兄B站视频专栏：https://space.bilibili.com/1900783
>
> 子豪兄Python交流QQ群：1077638784



集合数据类型



集合是多个元素的无序组合，是组合数据类型中的一种。

组合数据类型包括：集合、序列（列表、元组、字符串）、字典。

Python中的集合和数学中的集合概念是一致的（高中数学必修一第一课）。

- 集合元素之间是无序的。
- 集合的元素是唯一的，不能有重复。

> 集合中的元素 `不能` 是可变数据类型。
>
> 列表(list)是可修改的数据类型，不能作为集合的元素。
>
> 如果集合的元素出现了可变数据类型，比如列表，会报错 `TypeError: unhashable type: 'list'` 。

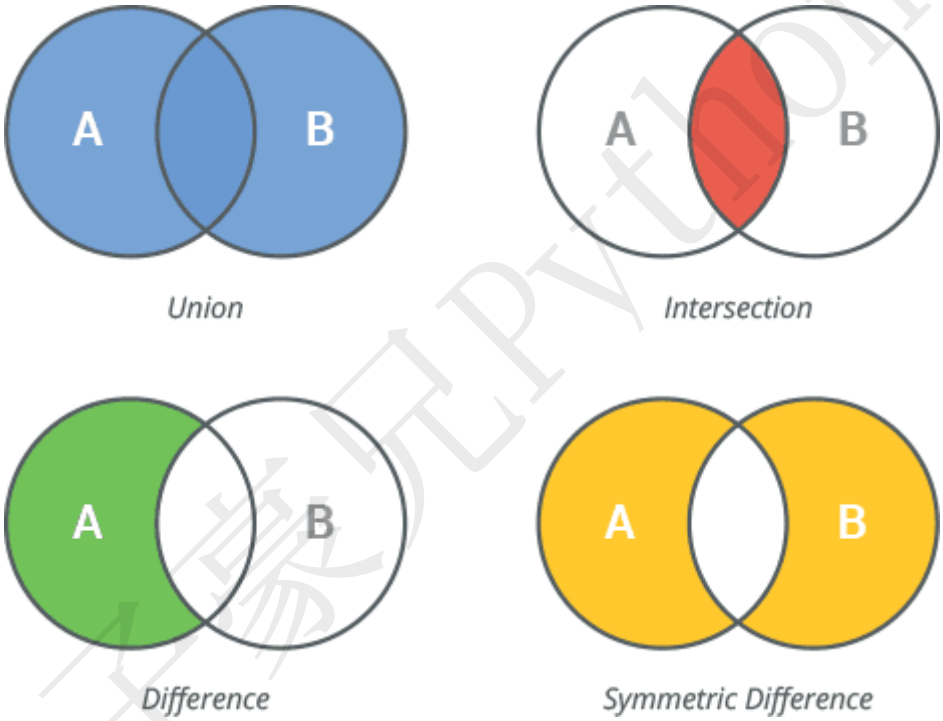
python中的集合用大括号 `{}` 来表示，元素之间用逗号分隔。

```
a = {1,2,3,4,5}

# 如果建立空集合，必须使用set()
a = set()

type(a)
# <class 'set'>

# 新建的是空字典，而不是空集合
a = {}

type(a)
# <class 'dict'>

a = set('python')
# 生成的集合a为{'h', 'n', 'o', 'p', 't', 'y'}

# 遍历集合元素
for each in a:
    print(each)
```

集合操作符

6个基本操作符

操作符	描述
S | T	并集
S & T	交集
S - T	差集
S ^ T	补集
S <= T 或 S < T	返回True/False，判断是否为子集（真子集）
S >= T 或 S> T	返回True/False，判断是否为超集（真超集）




```
S = {'同济','上交','复旦','浙大','川大','厦大','武大','中山'}

E = {'南开','哈工大','清华','北航','同济','复旦','上交','厦大','浙大'}

B = {'清华','北航'}
```

4个增强操作符

操作符	描述
S |= T或 S = S | T	将S更新为并集
S -= T或 S = S - T	将S更新为差集
S &= T或 S = S & T	将S更新为交集
S ^= T 或 S = S ^ T	将S更新为补集

集合处理方法

函数或方法	描述
S.add(x)	如果x不在集合S中，就将x增加到S。如果x已经在S中，那么集合不变。
S.discard(x)	移除S中的元素x，如果x不在集合S中，不报错
S.clear()	清空集合S
S.pop()	随机返回S中的一个元素，并且将改元素从S中移除。如果S是空集，就产生KeyError的异常
S.copy()	返回集合S的一个副本
len(S)	返回S的元素个数
x in S	判断x是否在S中，如果在，返回True，如果不在，返回False
x not in S	判断x是否不在S中，如果不在，返回True，如果在，返回False
set(x)	将其它类型变量x转为集合类型（工厂函数）

集合应用场景

集合的应用场景比较单一，在实际编程中并不特别常用。

包含关系比较

```
"p" in {"p","a","y",123}
# True
```

数据去重

```
ls = [1,2,3,4,5,6,7,7,6,5,4,3,2,1,2,2,2,2,2]
s = set(ls)
print(len(s))
# 7

ls = list(s)
# ls列表为[1, 2, 3, 4, 5, 6, 7]
```

自然语言处理中，要进行“去停词”操作，去掉某些废话词或标点符号，可以用集合构建“废话”语料库，能够保证语料库中的每一个元素都是唯一的。

set与frozenset

集合中的元素 不能 是可变数据类型，但集合本身是可以修改的，比如使用 `S.add(x)` 或 `S.discard(x)` 增删元素。

也就意味着，集合不能作为另一个集合的元素。

如果集合成为了另一个集合的元素，会报错 `TypeError: unhashable type: 'set'` 。

同理，集合同样也不能成为字典的键。

解决这个问题，可以使用 `frozenset()` 工厂函数生成不可变的集合frozenset。

思考题

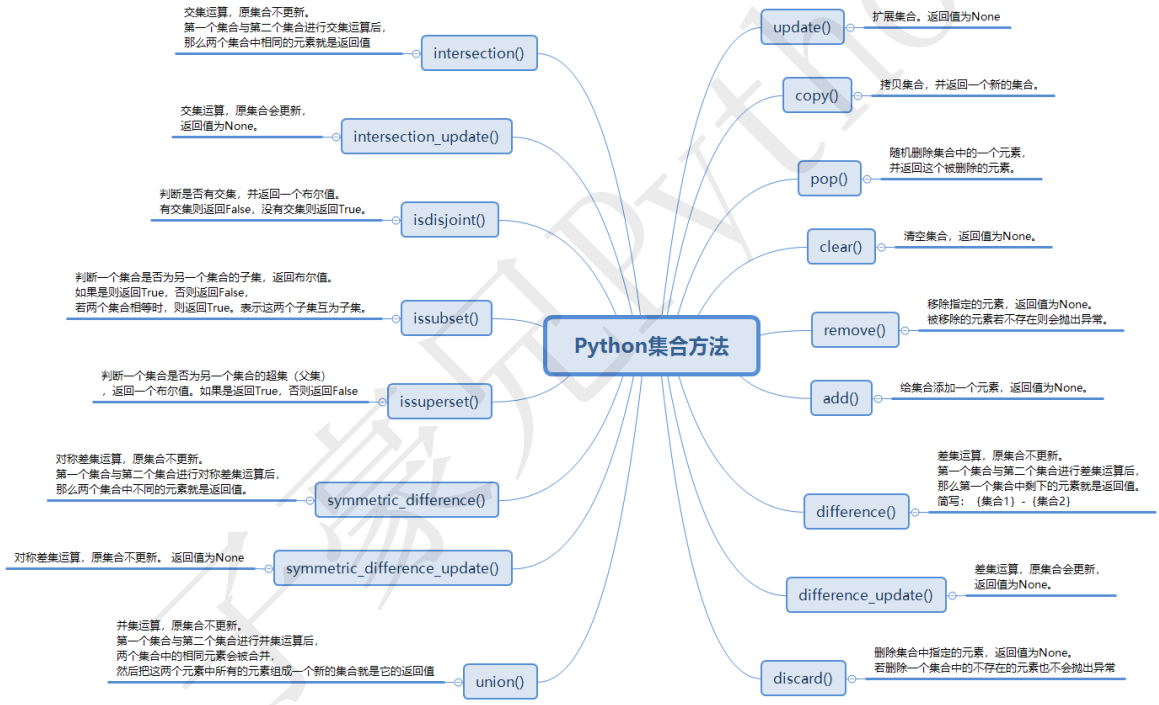
为什么列表（list）不能作为集合的元素？

为什么列表（list）作为集合元素时会报错 `TypeError: unhashable type: 'list'` ？

集合数据类型有哪些应用场景？

参考阅读

https://www.cnblogs.com/jkl1221/p/10742918.html



python中set和frozenset方法和区别

https://www.cnblogs.com/panwenbin-logs/p/5519617.html