数据预处理

数据结构及类型 — 理论与实践工作册

Python 编程语言

实践能力强化工作册

姓名：_____ 学号：_____

日期：_____

第一部分 学习目标

完成本工作册后，你将能够：

- 理解 Python 中基本数据类型的特征和应用场景
- 掌握列表（list）、元组（tuple）、字典（dict）、集合（set）的创建与操作
- 能够进行数据类型的转换与判断
- 掌握常用数据结构在数据预处理中的实际应用
- 能够选择合适的数据结构解决实际问题

第二部分 知识回顾

2.1 Python 基本数据类型概览

Python提供了多种内置数据类型，以下是常用类型的分类总结：

分类	类型	示例	特点
数值型	int, float, complex	42, 3.14, 1+2j	不可变
序列型	str, list, tuple	"abc", [1,2], (1,2)	有序、可索引
映射型	dict	{"a": 1}	键值对、可变
集合型	set, frozenset	{1, 2, 3}	无序、不重复
布尔型	bool	True, False	真/假值

2.2 可变与不可变类型

理解可变性是数据预处理的基础：

特性	可变类型	不可变类型
代表类型	list, dict, set	int, str, tuple, frozenset
能否修改	✓ 可以原地修改	✕ 修改会创建新对象
哈希表键	✕ 不能作为字典键	✓ 可以作为字典键

第三部分 填空练习

请在横线处填写正确的内容：

1. Python中的基本数值类型包括整数型 _____、浮点型 _____、和复数型 _____。

2. 列表是 _____（可变/不可变）的序列，而元组是 _____（可变/不可变）的序列。

3. 使用 _____ 函数可以查看一个变量的数据类型。

4. 字典的键必须是 _____（可变/不可变）类型，而值可以是 _____ 类型。

5. 集合（set）的两个核心特点是 _____ 和 _____。

6. 将字符串 "123" 转换为整数的函数是 _____，将整数转换为字符串的函数是 _____。

7. 列表推导式的基本语法是 _____。

8. 字典推导式的基本语法是 _____。

第四部分 代码阅读练习

阅读以下代码，写出每段代码的输出结果：

1. 预测以下代码的输出：

```
a = [1, 2, 3, 4, 5]
b = a
b.append(6)
print(a)
print(a is b)
```

输出结果：

提示： 思考：为什么修改b会影响b？这与可变对象的引用机制有关。

2. 预测以下代码的输出：

```
data = {'name': 'Alice', 'age': 25, 'scores': [88, 92, 76]}
print(type(data))
print(type(data['scores']))
print(len(data))
print(data.get('gender', 'Unknown'))
```

输出结果：

3. 预测以下代码的输出：

```
s1 = {1, 2, 3, 4, 5}
s2 = {3, 4, 5, 6, 7}
print(s1 & s2)       # 交集
print(s1 | s2)       # 并集
print(s1 - s2)       # 差集
print(s1 ^ s2)       # 对称差集
```

输出结果：

4. 预测以下代码的输出：

```
x = (1, 2, [3, 4])
x[2].append(5)
print(x)


# 以下代码会发生什么？
# x[0] = 10
```

输出结果及解释：

第五部分 编程实践练习

请根据要求编写Python代码，将代码写在空白区域：

练习 5.1 数据类型探索

要求：创建不同类型的变量，并使用 type() 和 isinstance() 进行类型检查。

```
# 任务：创建以下变量并打印其类型
# 1. 一个整数变量
# 2. 一个浮点数变量
# 3. 一个字符串变量
# 4. 一个布尔变量
# 5. 一个 None 值

# 使用 isinstance() 检查变量是否为数值类型
# 提示：isinstance(var, (int, float))

# 在此处编写代码：
```

你的代码：

练习 5.2 列表操作练习

要求：完成以下列表操作任务。

```
students = ['Alice', 'Bob', 'Charlie', 'David', 'Eve']
scores = [88, 92, 76, 95, 83]


# 任务 1：将两个列表合并为字典 {name: score}


# 任务 2：找出分数大于90的学生姓名（使用列表推导式）


# 任务 3：计算平均分并将低于平均分的成绩替换为平均分


# 任务 4：按分数从高到低排序并输出学生姓名和分数
```

你的代码：

练习 5.3 字典的应用

要求：使用字典处理学生信息数据。

```
# 给定以下学生数据
students_data = [
    {'name': 'Alice', 'age': 20, 'major': 'CS', 'gpa': 3.8},
    {'name': 'Bob', 'age': 21, 'major': 'Math', 'gpa': 3.5},
    {'name': 'Charlie', 'age': 20, 'major': 'CS', 'gpa': 3.9},
    {'name': 'David', 'age': 22, 'major': 'Physics', 'gpa': 3.2},
    {'name': 'Eve', 'age': 21, 'major': 'CS', 'gpa': 3.7},
]


# 任务 1：统计每个专业的学生人数


# 任务 2：找出GPA最高的学生信息


# 任务 3：使用字典推导式创建 {name: gpa} 的新字典


# 任务 4：按专业分组并计算各专业的平均GPA
```

你的代码：

练习 5.4 集合运算实践

要求：使用集合解决实际问题。

```
# 场景：分析两次考试的参加人员
exam1 = {'Alice', 'Bob', 'Charlie', 'David', 'Eve'}
exam2 = {'Bob', 'David', 'Frank', 'Grace', 'Eve'}


# 任务 1：找出两次都参加了的学生


# 任务 2：找出只参加了第一次考试的学生


# 任务 3：找出总共有哪些学生参加了考试


# 任务 4：找出只参加了其中一次考试的学生
```

你的代码：

第六部分 综合挑战

挑战 6.1 数据清洗实战

场景：你收到了一批原始订单数据，需要进行清洗和整理。

```
# 原始订单数据
raw_orders = [
    {'id': '001', 'product': 'Apple', 'price': '5.5', 'qty': '3'},
    {'id': '002', 'product': 'Banana', 'price': 'N/A', 'qty': '5'},
    {'id': '003', 'product': 'apple', 'price': '5.5', 'qty': '2'},
    {'id': '001', 'product': 'Apple', 'price': '5.5', 'qty': '3'},
    {'id': '004', 'product': 'Cherry', 'price': '15.0', 'qty': ''},
    {'id': '005', 'product': 'Durian', 'price': '45.0', 'qty': '1'},
]

# 任务：
# 1. 去除重复订单（根据 id 判断）
# 2. 将 price 和 qty 转换为数值类型，无效值设为 0
# 3. 统一产品名称为首字母大写
# 4. 计算每个订单的总价（price * qty）
# 5. 找出总价最高的订单
```

你的代码：

挑战 6.2 数据结构选择

针对以下场景，请选择最合适的数据结构并说明理由：

1. 场景：存储学生的姓名和对应的成绩

选择的数据结构：_____

理由：

2. 场景：记录一周七天的气温（不需要修改）

选择的数据结构：_____

理由：

3. 场景：统计一篇文章中每个单词出现的次数

选择的数据结构：_____

理由：

4. 场景：存储一组不重复的用户ID

选择的数据结构：_____

理由：

第七部分 选择题

请选择最佳答案：

1. 以下哪个不是Python的不可变类型？

A. int

B. str

C. list

D. tuple

答案：_____ 理由：_____

2. 执行 d = {}; d[[1,2]] = 'test' 会发生什么？

A. 正常赋值

B. TypeError，因为列表不可哈希

C. KeyError

D. ValueError

答案：_____ 理由：_____

3. 以下哪个表达式的结果是 True?

A. type([]) == type(())

B. isinstance(True, int)

C. type(1) == type(1.0)

D. {1, 2} == {2, 1, 2}

答案：_____（可多选）　理由：_____

4. 关于字典的说法，哪个是错误的?

A. 字典的键必须是不可变类型

B. 字典从Python 3.7起保持插入顺序

C. 字典的值可以是任意类型

D. 字典可以使用列表作为键

答案：_____　理由：_____

第八部分 自我评估

请根据完成情况进行自我评估：

评估项目	完全掌握	基本掌握	需要加强
能够正确识别Python的基本数据类型	□	□	□
能够熟练进行列表的创建、访问、修改操作	□	□	□
能够使用字典存储和检索键值对数据	□	□	□
能够运用集合进行去重和集合运算	□	□	□
能够根据场景选择合适的数据结构	□	□	□
能够进行数据类型转换处理	□	□	□
能够使用列表/字典推导式	□	□	□
能够完成综合数据清洗任务	□	□	□

学习心得与反思

请写下你在本次实践中的收获和需要进一步学习的内容：

数据结构与数据预处理——理论与实践工作册

基于 Python 的数据类型、结构及应用场景

第一部分 学习目标

完成本工作册后，你将能够：

- 理解Python中基本数据类型的特征和应用场景
- 掌握列表（list）、元组（tuple）、字典（dict）、集合（set）的创建与操作
- 能够进行数据类型的转换与判断
- 掌握常用数据结构在数据预处理中的实际应用
- 能够选择合适的数据结构解决实际问题

第二部分 知识回顾

2.1 Python 基本数据类型概览

Python提供了多种内置数据类型，以下是常用类型的分类总结：

分类	类型	示例	特点
数值型	int, float, complex	42, 3.14, 1+2j	不可变
序列型	str, list, tuple	"abc", [1,2], (1,2)	有序、可索引
映射型	dict	{"a": 1}	键值对、可变
集合型	set, frozenset	{1, 2, 3}	无序、不重复
布尔型	bool	True, False	真/假值

2.2 可变与不可变类型

理解可变性是数据预处理的基础：

特性	可变类型	不可变类型
代表类型	list, dict, set	int, str, tuple, frozenset
能否修改	✓可以原地修改	✗修改会创建新对象
哈希表键	✗不能作为字典键	✓可以作为字典键

第三部分 填空练习

请在横线处填写正确的内容：

1Python中的基本数值类型包括整数型_____、浮点型_____、和复数型_____。

2列表是_____（可变/不可变）的序列，而元组是_____（可变/不可变）的序列。

3使用_____函数可以查看一个变量的数据类型。

4字典的键必须是_____（可变/不可变）类型，而值可以是_____类型。

5集合（set）的两个核心特点是_____和_____。

6将字符串"123"转换为整数的函数是_____，将整数转换为字符串的函数是_____。

7列表推导式的基本语法是_____。

8字典推导式的基本语法是_____。

第四部分 代码阅读练习

阅读以下代码，写出每段代码的输出结果：

1预测以下代码的输出：

```
a = [1, 2, 3, 4, 5]
b = a
b.append(6)
print(a)
print(a is b)
```

输出结果：

> **提示：**思考：为什么修改b会影响b？这与可变对象的引用机制有关。

2预测以下代码的输出：

```
data = {'name': 'Alice', 'age': 25, 'scores': [88, 92, 76]}
print(type(data))
print(type(data['scores']))
print(len(data))
```

```
print(data.get('gender', 'Unknown'))
```

输出结果：

3预测以下代码的输出：

```
s1 = {1, 2, 3, 4, 5}
s2 = {3, 4, 5, 6, 7}
print(s1 & s2) #交集
print(s1 | s2) #并集
print(s1 - s2) #差集
print(s1 ^ s2) #对称差集
```

输出结果：

4预测以下代码的输出：

```
x = (1, 2, [3, 4])
x[2].append(5)
print(x)
# 以下代码会发生什么？
# x[0] = 10
```

输出结果及解释：

第五部分 编程实践练习

请根据要求编写Python代码，将代码写在空白区域：

练习 5.1 数据类型探索

要求：创建不同类型的变量，并使用type()和isinstance()进行类型检查。

```
# 任务：创建以下变量并打印其类型
# 1. 一个整数变量
# 2. 一个浮点数变量
# 3. 一个字符串变量
# 4. 一个布尔变量
# 5. 一个None值
# 使用isinstance()检查变量是否为数值类型
# 提示：isinstance(var, (int, float))
# 在此处编写代码：
```

你的代码：

练习 5.2 列表操作练习

要求：完成以下列表操作任务。

```
students = ['Alice', 'Bob', 'Charlie', 'David', 'Eve']
scores = [88, 92, 76, 95, 83]
# 任务1：将两个列表合并为字典{name: score}
# 任务2：找出分数大于90的学生姓名（使用列表推导式）
# 任务3：计算平均分并将低于平均分的成绩替换为平均分
# 任务4：按分数从高到低排序并输出学生姓名和分数
```

你的代码：

练习 5.3 字典的应用

要求：使用字典处理学生信息数据。

```
# 给定以下学生数据
students_data = [
{'name': 'Alice', 'age': 20, 'major': 'CS', 'gpa': 3.8},
{'name': 'Bob', 'age': 21, 'major': 'Math', 'gpa': 3.5},
{'name': 'Charlie', 'age': 20, 'major': 'CS', 'gpa': 3.9},
{'name': 'David', 'age': 22, 'major': 'Physics', 'gpa': 3.2},
{'name': 'Eve', 'age': 21, 'major': 'CS', 'gpa': 3.7},
]
# 任务1：统计每个专业的学生人数
# 任务2：找出GPA最高的学生信息
# 任务3：使用字典推导式创建{name: gpa}的新字典
# 任务4：按专业分组并计算各专业的平均GPA
```

你的代码：

练习 5.4 集合运算实践

要求：使用集合解决实际问题。

```
exam1 = {'Alice', 'Bob', 'Charlie', 'David', 'Eve'}
exam2 = {'Bob', 'David', 'Frank', 'Grace', 'Eve'}
# 任务1：找出两次都参加了的学生
# 任务2：找出只参加了第一次考试的学生
# 任务3：找出总共有哪些学生参加了考试
# 任务4：找出只参加了其中一次考试的学生
```

你的代码：

第六部分 综合挑战

挑战 6.1 数据清洗实战

场景：你收到了一批原始订单数据，需要进行清洗和整理。

```
# 原始订单数据
raw_orders = [
{'id': '001', 'product': 'Apple', 'price': '5.5', 'qty': '3'},
{'id': '002', 'product': 'Banana', 'price': 'N/A', 'qty': '5'},
{'id': '003', 'product': 'apple', 'price': '5.5', 'qty': '2'},
{'id': '001', 'product': 'Apple', 'price': '5.5', 'qty': '3'},
{'id': '004', 'product': 'Cherry', 'price': '15.0', 'qty': ''},
{'id': '005', 'product': 'Durian', 'price': '45.0', 'qty': '1'},
]
# 任务：
# 1. 去除重复订单（根据id判断）
# 2. 将price和qty转换为数值类型，无效值设为0
# 3. 统一产品名称为首字母大写
# 4. 计算每个订单的总价（price * qty）
# 5. 找出总价最高的订单
```

你的代码：

挑战 6.2 数据结构选择

针对以下场景，请选择最合适的数据结构并说明理由：

1场景：存储学生的姓名和对应的成绩

选择的数据结构：_____

理由：

2场景：记录一周七天的气温（不需要修改）

选择的数据结构：_____

理由：

3场景：统计一篇文章中每个单词出现的次数

选择的数据结构：_____

理由：

4场景：存储一组不重复的用户ID

选择的数据结构：_____

理由：

第七部分 选择题

请选择最佳答案：

1以下哪个不是Python的不可变类型？

A.int

B.str

C.list

D.tuple

答案：_____ 理由：_____

2执行d = {}; d[[1,2]] = 'test'会发生什么？

A.正常赋值

B.TypeError，因为列表不可哈希

C.KeyError

D.ValueError

答案：_____ 理由：_____

3以下哪个表达式的结果是True？

A.type([]) == type(())

B.isinstance(True, int)

C.type(1) == type(1.0)

D.{1, 2} == {2, 1, 2}

答案：_____（可多选） 理由：_____

4关于字典的说法，哪个是错误的？

A.字典的键必须是不可变类型

B.字典从Python 3.7起保持插入顺序

C.字典的值可以是任意类型

D.字典可以使用列表作为键

答案：_____ 理由：_____

第八部分 自我评估

请根据完成情况进行自我评估：

评估项目	完全掌握	基本掌握	需要加强
能够正确识别Python的基本数据类型	☐	☐	☐
能够熟练进行列表的创建、访问、修改操作	☐	☐	☐
能够使用字典存储和检索键值对数据	☐	☐	☐
能够运用集合进行去重和集合运算	☐	☐	☐
能够根据场景选择合适的数据结构	☐	☐	☐
能够进行数据类型转换处理	☐	☐	☐
能够使用列表/字典推导式	☐	☐	☐
能够完成综合数据清洗任务	☐	☐	☐

学习心得与反思

请写下你在本次实践中的收获和需要进一步学习的内容：