数据预处理

字符串与字符处理

实践工作册

编程语言：Python

姓名：_____　　学号：_____

班级：_____　　日期：_____

目录

第一章　字符串基础操作 3

第二章　字符串索引与切片 6

第三章　字符串常用方法 9

第四章　字符串格式化 13

第五章　正则表达式基础 16

第六章　数据清洗实战 20

第七章　综合挑战题 24

参考答案 27

答案

第一章参考答案

练习1

```
# 输出: Python 数据预处理
# 输出: 6
# 输出: PythonPythonPython
# 输出: True
```

练习2

```
# str(100) -> "100", str类型
# int("256") -> 256, int类型
# float("3.14") -> 3.14, float类型
# str(True) -> "True", str类型
# bool("") -> False, bool类型
# bool("0") -> True, bool类型
# bool("False") -> True, bool类型
```

第二章参考答案

练习1

```
# s[0] -> "D"
# s[-1] -> "g"
# s[5] -> "P"
# s[0:4] -> "Data"
# s[5:] -> "Preprocessing"
# s[:4] -> "Data"
# s[::2] -> "Dt rpoesg"
# s[::-1] -> "gnissecorperP ataD"
# s[5:17:3] -> "Pors"
```

练习2

```
# (1) s[0:5]
# (2) s[7:12]
# (3) s[::-1]
# (4) s[::2]
```

练习3 身份证解析

```
region = id_number[0:6]
year = id_number[6:10]
month = id_number[10:12]
day = id_number[12:14]
gender_code = int(id_number[16])
# 输出: {"地区编码":"110105", "出生日期":"1990年03月07日", "性别":"男"}
```

第三章参考答案

练习1

```
# "HELLO WORLD"
# "hello world"
# "Hello World"
# "Hello Python"
# ["Hello", "World"]
# "Data Pre Processing"
# "Data Science"
# "***hello***"
# 2
# 3
```

练习2

```
text = text.strip()
text = text.lower()
text = text.replace("data", "数据")
words = text.split()
```

```
result = "_".join(words)
```

练习3 邮箱解析

```
def parse_email(email):
    parts = email.split("@")
    return {"用户名": parts[0], "域名": parts[1]}
```

练习4 CSV解析

```
def parse_csv_line(line):
    fields = []                          # 결과를 저장할 빈 리스트 생성
    raw_fields = line.split(",")         # 쉼표를 기준으로 문자열 분할

    for field in raw_fields:             # 분할된 각 항목을 반복
        cleaned_field = field.strip()    # 앞뒤 공백 제거
        fields.append(cleaned_field)     # 정제된 항목을 리스트에 추가

    return fields                        # 최종 리스트 반환
```

第四章参考答案

练习1

```
print(f"商品：{product}，单价：{price:.2f}元")
print(f"数量：{quantity:03d}")
print(f"折扣：{discount:.1%}")
print(f"总价：{total:,.2f}元")
```

第五章参考答案

练习1

```
# 手机号: r"1\d{10}"
# 邮箱: r"[\w.+-]+@[\w-]+\.[\w.]+"
# 日期: r"\d{4}-\d{2}-\d{2}"
```

```
# 中文: r"[\u4e00-\u9fff]+"
```

练习2

```
phone = re.findall(r"1\d{10}", text)
email = re.findall(r"[\w.+-]+@[\w-]+\.[\w.]+", text)
date = re.findall(r"\d{4}-\d{2}-\d{2}", text)
```

练习3 数据提取

```
def extract_and_sum(text):
    numbers = re.findall(r"\d+\.?\d*", text)
    nums = [float(n) for n in numbers]
    return {"numbers": nums, "sum": sum(nums)}
```

练习4 敏感信息脱敏

```
def desensitize(text):
    text = re.sub(r"(1\d{2})\d{4}(\d{4})", r"\1****\2", text)
    text = re.sub(r"([\w])[\w.]*@", r"\1***@", text)
    return text
```

— 工作册完 —