

# 成都师范学院计算机科学学院

## 实验指导

课 程 名 称 : 数据采集与处理

课 程 类 型 : 选修

课 程 性 质 : 专业基础

上 课 时 间 : 2019 至 2020 学年 第 1 学期  
计算机科学与计划专业 2016 级

授 课 对 象 : 01、02 班

教 师 姓 名 : 李林

计算机科学学院

## 成都师范学院计算机科学学院

### 《数据采集与处理》课程实验（项目实训）指导

实验（项目）一、 股票数据定向 Scrapy 爬虫

实验（项目）二、 使用 Matplotlib 实现国民经济数据的绘制

实验（项目）三、 panda 数据处理实验

实验（项目）四、 基于 wine 数据集的数据分析

## 实验一

### 一、实验名称

股票数据定向 Scrapy 爬虫

### 二、实验目的

- 1) 掌握 Scrapy 爬虫基本使用过程；
- 2) 理解百度网站股票数据基本数据结构；
- 3) 掌握 BeautifulSoup 对 HTML 源码进行解析方法。

### 三、实验要求

- 1) 学生个人独立完成实验；
- 2) 实验后需要撰写实验报告；
- 3) 实验报告应包括原理、步骤、实验结果分析等，重点在实验结果分析。

### 四、实验环境（软硬件条件）

1. 每人 1 台 PC 机；
2. 安装 anaconda 3；
3. visual studio code 或 pycharm 2019。

### 五、实验内容（重难点）

- 1) 构建自己的 Scrapy 爬虫框架程序；
- 2) 构造可接受参数的 Scrapy 爬虫；
- 3) 运行 Scrapy 爬虫；
- 4) 分析爬取信息。

## 六、实验步骤

参考代码：

步骤 1：建立工程和 Spider 模板

```
\>scrapy startproject BaiduStocks
```

```
\>cd BaiduStocks
```

```
\>scrapy genspider stocks baidu.com
```

进一步修改 spiders/stocks.py 文件

步骤 2：编写 Spider

配置 stocks.py 文件 修改对返回页面的处理

```
# -*- coding: utf-8 -*-
import scrapy
import re
class StocksSpider(scrapy.Spider):
    name = "stocks"
    start_urls = ['http://quote.eastmoney.com/stocklist.html']

    def parse(self, response):
        for href in response.css('a::attr(href)').extract():
            try:
                stock = re.findall(r"[s][hz]\d{6}", href)[0]
                url = 'https://gupiao.baidu.com/stock/' + stock + '.html'
                yield scrapy.Request(url, callback=self.parse_stock)
            except:
                continue
```

修改对新增 URL 爬取请求的处理。

```
def parse_stock(self, response):
    infoDict = {}
    stockInfo = response.css('.stock-bets')
    name = stockInfo.css('.bets-name').extract()[0]
    keyList = stockInfo.css('dt').extract()
    valueList = stockInfo.css('dd').extract()
    #此处，请输入代码完成股票网页信息提取，并生产 infoDict 信息

    #结束
```

```
yield infoDict
```

步骤 3: 编写 ITEM Pipelines

```
class BaidustocksInfoPipeline(object):
    def open_spider(self, spider):
        self.f = open('BaiduStockInfo.txt', 'w')

    def close_spider(self, spider):
        self.f.close()

    def process_item(self, item, spider):
        #此处请输入处理代码，将结构写到文件
        ...
    #结束
```

七、实验结果要求和分析

实验结果:

```
{ '今开': '13.57', '成交量': '254.29 万手', '最高': '14.09', '涨停': '15.02', '内盘': '134.08 万手', '成交额': '35.25 亿', '委比': '7.52%', '流通市值': '751.17 亿', '市盈率<sup>MRQ</sup>': '53.29', '每股收益': '0.13', '总股本': '67.16 亿', '昨收': '13.65', '换手率': '4.68%', '最低': '13.57', '跌停': '12.28', '外盘': '144.23 万手', '振幅': '3.81%', '量比': '5.92', '总市值': '928.10 亿', '市净率': '4.58', '每股净资产': '3.01', '流通股本': '54.35 亿', '股票名称': '东方财富 300059' }
```

生成 json 文件，获得股票的基本信息。

八、实验检查与考核

1. 实验检查

- 1) 学生课堂考勤;
- 2) 课堂检查学生是否个人独立完成，而不是抄袭他人成果。

2. 实验考核

- 1) 根据撰写实验报告质量进行评分;
- 2) 根据实验程序质量和创新性;
- 3) 实验成绩采用百分制。

## 实验二

一、实验名称

使用 Matplotlib 实现人口数据的绘制

二、实验目的

- 1) 掌握 pyplot 基本语法;
- 2) 掌握子图的绘制方法;
- 3) 掌握散点图、折线图的绘制方法;
- 4) 掌握直方图、饼图、箱线图绘制方法。

### 三、实验要求

- 1) 学生个人独立完成实验;
- 2) 实验后需要撰写实验报告;
- 3) 实验报告应包括原理、步骤、实验结果分析等, 重点在实验结果分析。

### 四、实验环境 (软硬件条件)

1. 每人 1 台 PC 机;
2. 安装 anaconda 3;
3. visual studio code 或 pycharm 2019。

### 五、实验内容 (重难点)

- 1) 使用 NumPy 库读取人口数据;
- 2) 创建画布, 并添加子图;
- 3) 在两个子图上分别绘制散点图和折线图;
- 4) 创建 3 幅画布并添加对应数目的子图;
- 5) 保存和显示图形;
- 6) 根据图形, 分析我国人口结构变化情况以及变化速率的增减情况。

### 六、实验步骤

参考代码:

```
#
# 实训 3.1 分析 1996~2015 年人口各数据特征的分布与分散状况
# p78
# Copyright (c) 2019, Lin LI
# All rights reserved.
#

import numpy as np
import matplotlib.pyplot as plt
plt.rcParams['font.sans-serif'] = 'SimHei' ## 设置中文显示
plt.rcParams['axes.unicode_minus'] = False
```

```

#1) 使用 NumPy 库读取人口数据;
data = np.load("populations.npz")['data']
feature_names = np.load("populations.npz")['feature_names']
#print(feature_names)
#2) 创建画布, 并添加子图;
p = plt.figure(figsize=(12,12)) ##设置画布
plt.title('实训 3.2 分析 1996~2015 年人口各数据特征的分布与分散状况')
#3) 创建 3 幅画布并添加对应数目的子图;
for i in range(0,19,2):
    #此处, 请写入代码完成特征绘图
    ...
    #结束
plt.show()
#4) 保存和显示图形;
plt.savefig('./results/3.1.2.png')

```

## 七、实验结果要求和分析

- 1) 绘制国民经济数据的各个指标
- 2) 分析各个指标对经济影响趋势

## 八、实验检查与考核

### 1. 实验检查

- 1) 学生课堂考勤;
- 2) 课堂检查学生是否个人独立完成, 而不是抄袭他人成果。

### 2. 实验考核

- 1) 根据撰写实验报告质量进行评分;
- 2) 根据实验程序质量和创新性;
- 3) 实验成绩采用百分制。

# 实验三

## 一、实验名称

panda 数据预处理实验

## 二、实验目的

- 1) 掌握缺失值识别方法;
- 2) 掌握对缺失值数据处理的方法;
- 3) 掌握主键合并的几种方法;

4) 掌握多个键值的主键合并。

### 三、实验要求

- 1) 学生个人独立完成实验；
- 2) 实验后需要撰写实验报告；
- 3) 实验报告应包括原理、步骤、实验结果分析等，重点在实验结果分析。

### 四、实验环境（软硬件条件）

1. 每人 1 台 PC 机；
2. 安装 anaconda 3；
3. visual studio code 或 pycharm 2019。

### 五、实验内容（重难点）

- 1) 读取 missing\_data.csv；
- 2) 查询缺失值所在位置；
- 3) 使用 SciPy 中 interpolate 中的 lagrange 对数据进行拉格朗日插值；
- 4) 查看数据中是否存在缺失值，若不存在则说明插值成功。
- 5) 读取 ele\_loss.csv 和 alarm.csv 表
- 6) 查看两表的形状；
- 7) 以 ID 和 date 为两个键值作为主键进行内连接。
- 8) 查看合并后的数据

### 六、实验步骤

参考代码：

按照试验内容编写一下代码实现功能：

```
import pandas as pd
import numpy as np

detail = pd.read_csv('./data/model.csv', sep = ',', encoding = 'gbk')
print('detail 的形状为: ', detail.shape)
print('detail 的数据为: \n', detail)

## 自定义离差标准化函数
def MinMaxScale(data):
    data=(data-data.min())/(data.max()-data.min())
    return data
#此处，请输入代码用 MinMaxScale 实现数据标准化，并横向合并数据。
...
#结束
print('离差标准化之前销量和售价数据为: \n',
      detail[['电量趋势下降指标', '告警类指标']].head())
print('离差标准化之后销量和售价数据为: \n', data3.head())
```

```

#自定义标准差标准化函数
def StandardScaler(data):
    data=(data-data.mean())/data.std()
    return data
##对电量数据做标准化
#此处，请输入代码用 StandardScaler 实现数据标准化，并横向合并数据。
...
#结束
print('标准差标准化之前销量和售价数据为：\n',
      detail[['电量趋势下降指标','告警类指标']].head())
print('标准差标准化之后销量和售价数据为：\n',data6.head())

##自定义小数定标差标准化函数
def DecimalScaler(data):
    data=data/10**np.ceil(np.log10(data.abs().max()))
    return data
##对菜品订单表售价和销量做标准化
#此处，请输入代码用 DecimalScaler 实现数据标准化，并横向合并数据。
...
#结束

print('小数定标标准化之前的销量和售价数据：\n',
      detail[['电量趋势下降指标','告警类指标']].head())
print('小数定标标准化之后的销量和售价数据：\n',data9.head())

```

## 七、实验结果要求和分析

检查输出结果并进行适当分析。

## 八、实验检查与考核

### 1. 实验检查

- 1) 学生课堂考勤；
- 2) 课堂检查学生是否个人独立完成，而不是抄袭他人成果。

### 2. 实验考核

- 1) 根据撰写实验报告质量进行评分；
- 2) 根据实验程序质量和创新性；
- 3) 实验成绩采用百分制。

# 实验四

## 一、实验名称

基于 wine 数据集的数据分析

## 二、实验目的



- 1) 掌握 sklearn 转换器的用法;
- 2) 掌握训练集、测试集划分方法;
- 3) 掌握使用 sklearn 进行 PCA 降维的方法;
- 4) 掌握使用 sklearn 进行聚类和分析结果

### 三、实验要求

- 1) 学生个人独立完成实验;
- 2) 实验后需要撰写实验报告;
- 3) 实验报告应包括原理、步骤、实验结果分析等, 重点在实验结果分析。

### 四、实验环境 (软硬件条件)

1. 每人 1 台 PC 机;
2. 安装 anaconda 3;
3. visual studio code 或 pycharm 2019。

### 五、实验内容 (重难点)

- 1) 使用 pandas 库分别读取 wine 数据集和 wine\_quality 数据集;
- 2) 将 wine 数据集和 wine\_quality 数据集的数据和标签拆分开;
- 3) 将 wine\_quality 数据划分为训练集和测试集;
- 4) 标准化 wine 数据集和 wine\_quality 数据集;
- 5) 对 wine 数据集和 wine\_quality 数据集进行 PCA 降维。

### 六、实验步骤

参考代码:

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split

#1) 使用 pandas 库分别读取 wine 数据集和 winequality 数据集
wine = pd.read_csv('./data/wine.csv', sep = ',', encoding = 'gbk')
winequality = pd.read_csv('./data/winequality.csv', sep = ';', encoding = 'gbk')

print('wine 的形状为: ', wine.shape)
print('winequality 的形状为: ', winequality.shape)

#2) 将数据集分成训练和测试两部分
wine_label = wine.iloc[:,0]
```

```

wine_data = wine.iloc[:,1:]

winequality_data = winequality.iloc[:, :-2]
winequality_lable = winequality.iloc[:, -1]
#此处, 请输入代码, 实现数据集拆分。
...
#结束
## 标准化
#计算训练集的平均值和标准差, 以便测试数据集使用相同的变换官方文档
from sklearn.preprocessing import StandardScaler
#此处, 请用 StandardScaler 实现数据标准化
...
#结束
##PCA 降维
#此处, 请输入实现 PCA 对 wine 数据降维
from sklearn.decomposition import PCA
...
#结束
print('PCA 降维前训练集数据的形状为: ', wine_std_train.shape)
print('PCA 降维后训练集数据的形状为: ', wine_trainPca.shape)
print('PCA 降维前测试集数据的形状为: ', wine_std_test.shape)
print('PCA 降维后测试集数据的形状为: ', wine_testPca.shape)

##根据实训 1 的 wine 数据集处理的结果, 构建聚类数目为 3 的 K-Means 模型

##构建并训练模型
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters = 3, random_state=42).fit(wine_trainPca)
print('构建的 KM-eans 模型为: \n', kmeans)
print('wine_target_train\n', wine_target_train,)
print('kmeans.labels_\n', kmeans.labels_ + 1,)

from sklearn.metrics import fowlkes_mallows_score
for i in range(2,7):
    ##构建并训练模型
    #此处, 请输入代码, 实现聚类和 FMI 评价
    ...
    print('iris 数据聚%d 类 FMI 评价分值为: %f' %(i, score))
    #结束
from sklearn.metrics import silhouette_score
import matplotlib.pyplot as plt
silhouettteScore = []
for i in range(2,15):
    ##构建并训练模型

```

```

    #此处，请输入代码，实现聚类和 FMI 评价
    ...
    print('iris 数据聚%d 类 silhouette_score 评价分值为: %f' %(i,score))
    #结束
plt.figure(figsize=(10,6))
plt.plot(range(2,15),silhouettteScore,linewidth=1.5, linestyle="-")
plt.show()

##
from sklearn.metrics import calinski_harabasz_score
for i in range(2,10):
    ##构建并训练模型
    kmeans = KMeans(n_clusters = i,random_state=12).fit(wine_trainPca)
    score = calinski_harabasz_score(wine_trainPca,kmeans.labels_)
    print('seeds 数据聚%d 类 calinski_harabaz 指数为: %f'%(i,score))

```

## 七、实验结果要求和分析

记录试验结果，并做分析。

## 八、实验检查与考核

### 1. 实验检查

- 1) 学生课堂考勤；
- 2) 课堂检查学生是否个人独立完成，而不是抄袭他人成果。

### 2. 实验考核

- 1) 根据撰写实验报告质量进行评分；
- 2) 根据实验程序质量和创新性；
- 3) 实验成绩采用百分制。