**一、实验介绍**

**1. 实验目的**

本实验的主要目的是通过实际操作，综合运用所学的大数据处理、数据可视化、数据存储与图形化展示等相关知识，提升学生的综合实践能力和问题解决能力。具体目标包括：

• 掌握运用requests库和正则表达式进行网页数据抓取的基本方法，通过实际操作双色球网站信息的抓取，深入理解网页数据解析与提取的原理。

• 通过对抓取的双色球开奖信息进行数据处理与汇总，提高学生利用pandas库进行数据处理的能力，包括数据清洗、转换和格式化等。

• 掌握MySQL的数据存储方式，学习将处理后的双色球开奖信息保存到MySQL数据库中，实现数据的持久化存储与管理。

**2. 实验内容**

本实验使用requests网页解析库和正则表达式实现双色球网站信息的抓取（http://kaijiang.500.com/shtml/ssq/），抓取各期双色球开奖信息，并对数据进行处理和存储，对数据进行可视化的展示，通过实验，让学生更深入的了解requests库和正则表达式原理及使用，掌握pandas数据处理、MySQL数据存储等知识。

**实验内容主要包括以下步骤：**

**• 网页数据抓取：**使用requests库发送HTTP请求，获取双色球开奖信息的网页内容。

•**数据提取：**编写正则表达式，从网页内容中提取各期双色球开奖号码、时间等相关信息。

**• 数据处理：**利用pandas库对提取的数据进行清洗、整理，形成结构化数据集。W3school

**• 数据存储：**将处理后的数据保存为CSV文件，并学习使用MySQL数据库进行数据存储。

**3. 实验知识点**

• requests网页解析库（爬虫库）

• 正则表达式

• Pandas数据处理

• Python连接MySQL数据库