黔南民族职业技术学院

**实 验 报 告**

实验课程 数据挖掘技术

实验项目 Python数据挖掘与分析

实验地点 408实训室

指导教师 姜永成

班 级 19大数据技术与应用(2)班

学生姓名 陈露忆

学 号 201903170236

教师评分

日 期

**期末实验 Python数据挖掘与分析**

**一、实验目的**

1. 掌握数据挖掘的基础知识，能够深刻理解并熟练运用Python进行数据处理的过程。

2. 掌握数据获取与收集方法。

3. 学会数据清洗和整理的方法。

4. 了解数据统计方法。

5. 能够利用可视化工具进行数据的展示。

1. **实训任务**

**任务1：【分析网页url】**

爬取的网站名称为：中商情报网

爬取内容：全国A股上市公司数据

网页地址为：<https://s.askci.com/stock/a/>

网址规律：

第一页网址为：<https://s.askci.com/stock/a/>/0-0?reportTime=2021-03-31&pageNum=1#QueryCondition

第二页的网址为：<https://s.askci.com/stock/a/>/0-0?reportTime=2021-03-31&pageNum=2#QueryCondition

以此类推可以发现该网址分为三部分，每翻一页网址的变化都为&pageNum+1

|  |
| --- |
| 源码如下 |
| 1  2 |

**任务2：【利用scrapy框架方法爬取中商情报网全国A股上市公司数据】**

从网页的分析来看，所需爬取的数据为：

“公司名称”，“省份”，“城市”，“营业收入”，“利润”，“员工人数”，“上市日期”，“行业分类”，“产品类型”

|  |
| --- |
| 源码如下 |
| 3  4 |

**任务3：【数据存储】**

将所爬取的相关数据存入到csv

|  |
| --- |
| 源码如下 |
| 5 |

**任务4：【数据清洗和导入MYSQL】**

将所爬取的相关数据导入到mysql

|  |
| --- |
| 源码如下 |
| 6  7  8 |

**任务5：【数据分析与可视化展示】**

**可视化（1）:**

使用中国地图可看出A股公司在全国的地域分布,可以发现超过300家的上市公司分布于北京，上海，浙江较为多。

|  |
| --- |
| 源码及图形如下 |
| **09**  **C:\Users\JKL\Desktop\期末（2）\tu\1.png** |

**可视化（2）:**

利用雷达图表达上市公司里行业分类的情况，可以准确的了解到汽车零件行业分类最为突出。

|  |
| --- |
| 源码及图形如下 |
|  |

**可视化（3）:**

利用三个水球图可看出在省份前十中北京，上海，广东所占百分比。

|  |
| --- |
| 源码及图形如下 |
| **x** |

**可视化（4）:**

用散点图分析部分公司所占人数分布。

|  |
| --- |
| 源码及图形如下 |
|  |

**可视化（5）:**

由折线图可看出近三年来A股上市公司的比较，可以准确的发现近三年来2019年上市公司较突出。

|  |
| --- |
| 源码及图形如下 |
| **m** |

**可视化（6）:**

使用柱状图来分析A股公司上市在前十的城市。

|  |
| --- |
| 源码及图形如下 |
| **88**  **5** |

**可视化（7）:**

利用饼图来分析公司的营业收入，可知A股上市公司的营业收入都平均处在与14亿左右。

|  |
| --- |
| 源码及图形如下 |
| pie1  pie |