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 20大数据技术与应用(5)班

学生姓名 姚玲

学 号 202003170533

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：豆瓣电影

爬取内容：豆瓣电影TOP250

网页地址为：https://movie.douban.com/top250

网址规律：

第一页网址为：https://movie.douban.com/top250

第二页的网址为：https://movie.douban.com/top250?start=25&filter=

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+25

|  |
| --- |
| 源码如下 |
| C:\Users\姚玲\Desktop\姚玲  python期末 豆瓣top250\截图\001.PNG001  C:\Users\姚玲\Desktop\姚玲  python期末 豆瓣top250\截图\003.PNG003 |

**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：

“排名”，“电影链接”，“海报链接”，“电影名”，“上映国家”，“上映年份”，“电影类型”，“导演”，“评分”，“评分人数”。

|  |
| --- |
| 源码如下 |
| C:\Users\姚玲\Desktop\报告\001.PNG001  C:\Users\姚玲\Desktop\姚玲  python期末 豆瓣top250\截图\002.PNG002 |

**任务3：【数据存储】**

将所爬取的相关数据存入到csv

|  |
| --- |
| 源码如下 |
| C:\Users\姚玲\Desktop\姚玲  python期末 豆瓣top250\截图\2.PNG2 |

**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 |
| C:\Users\姚玲\Desktop\姚玲  python期末 豆瓣top250\截图\005.PNG005  C:\Users\姚玲\Desktop\报告\002.PNG002 |

**任务5：【数据分析与可视化展示】**

**可视化（1）:**

|  |
| --- |
| 总的来看，从1931年到1983年，TOP250的影片基本都是1到2部，1931年之后开始总体呈现大部以上甚至4部以上，可以看出1931年之后国际电影事业得到很大的发展。源码及图形如下 |
| **C:\Users\姚玲\Desktop\姚玲  python期末 豆瓣top250\截图\008.PNG008**  **C:\Users\姚玲\Desktop\姚玲  python期末 豆瓣top250\截图\015.PNG015** |

**可视化（2）:**

|  |
| --- |
| 宫崎骏（8部）、克里斯托弗·诺兰（7部）、史蒂文·斯皮尔伯格（6部）、李安（5部）、王家卫（5部）导演的电影进入top250的数量排名前三。源码及图形如下 |
| **C:\Users\姚玲\Desktop\姚玲  python期末 豆瓣top250\截图\010.PNG010**  **C:\Users\姚玲\Desktop\姚玲  python期末 豆瓣top250\截图\011.PNG011** |

**可视化（3）:**

美国、日本、英国、中国产出高质量电影数量靠前。其中，美国上榜的电影数量最大占主导地位。

|  |
| --- |
| 源码及图形如下 |
| **C:\Users\姚玲\Desktop\姚玲  python期末 豆瓣top250\截图\012.PNG012**  **C:\Users\姚玲\Desktop\姚玲  python期末 豆瓣top250\截图\013.PNG013** |

**可视化（4）:**

|  |
| --- |
| 豆瓣电影top250都是评分在8.3及以上的电影，评分很高的电影往往伴随着很高的讨论度，这些电影往往带有人生的思考，积极的立意，精良的制作。源码及图形如下 |
| **C:\Users\姚玲\Desktop\姚玲  python期末 豆瓣top250\截图\020.PNG020**  **C:\Users\姚玲\Desktop\姚玲  python期末 豆瓣top250\截图\021.PNG021** |

**可视化（5）:**

制片国家是美国、日本、英国和中国产出高质量电影数量占比较大。可以看出中国近年来电影事业的发展迅速，但是美国的电影优质电影的占比最大，因此一些特点和技术的学习是有必要的。

|  |
| --- |
| 源码及图形如下 |
| **C:\Users\姚玲\Desktop\姚玲  python期末 豆瓣top250\截图\014.PNG014**  **C:\Users\姚玲\Desktop\姚玲  python期末 豆瓣top250\截图\022.PNG022** |