**数据科学家的第一个 Project**

****

**学号：201600301079**

**姓名：崔玉峰**

**班级：2016级**

1. **实验概述：**
2. **实验目标：**

作为数据科学第一个工程，并没有固定的实验题目，只给了Demo进行参考，结合自己的爱好以及能力，我选择一个关于游戏统计的项目进行实验：

**分析某游戏众评网站上热门游戏的评分，标签热度等信息，进行统计分析。**

1. **实验环境：**

硬件环境：PC

软件环境：PyCharm，JupyterNotebook

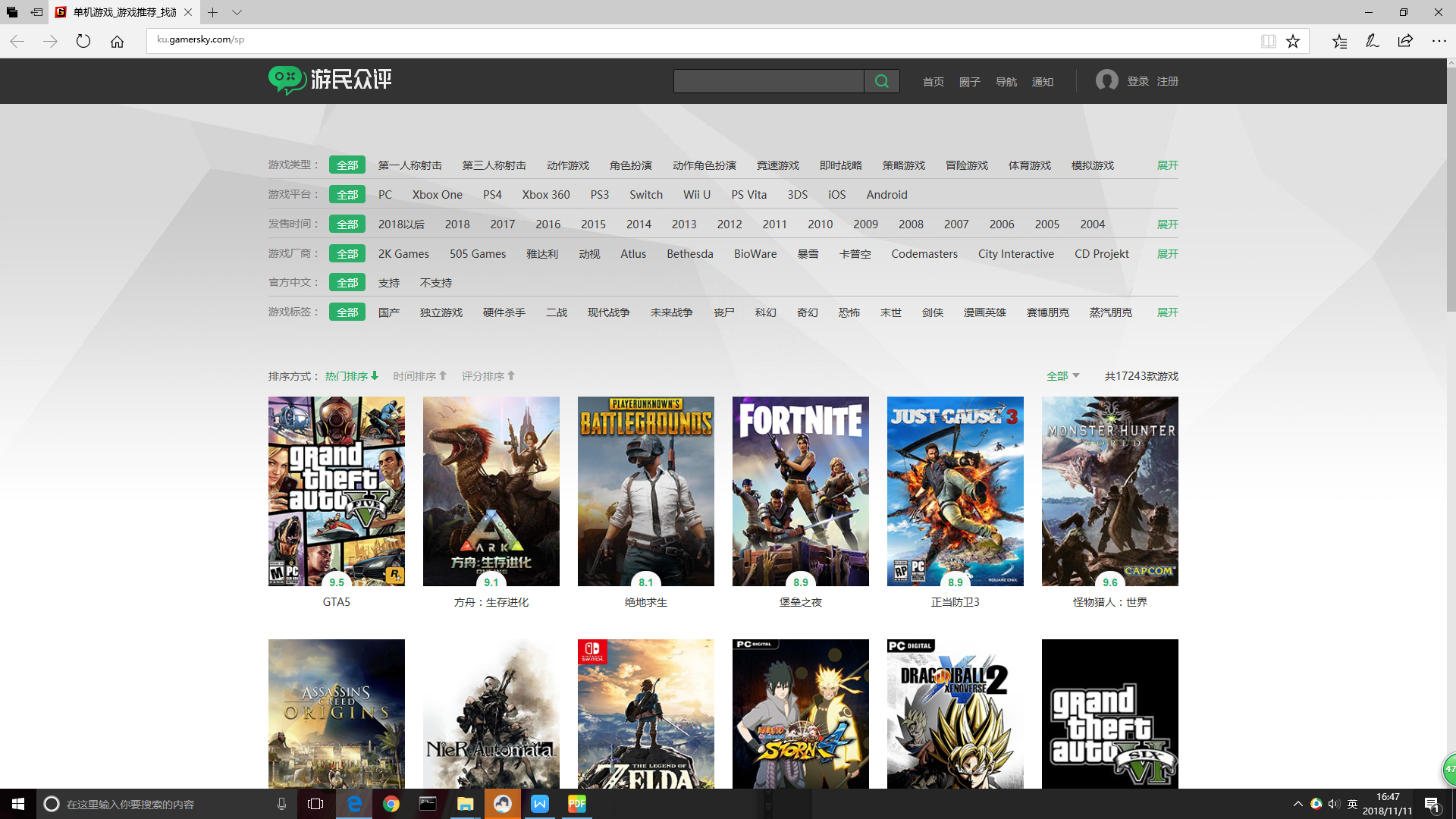
实验语言：Python

1. **数据获取：**

Python作为一个优秀语言，拥有很多的开源库可以使用，所以用Python写爬虫是一个不错的想法，但是因为以前从未写过python爬虫程序所以走了很多弯路

1. **通过Scrapy库进行爬取网站信息：**

网站样式：



scrapy是现在最火的爬虫库，适合构建大型爬虫项目，但是因为对于这个库的了解不够，通过代码编写爬虫后发现，并不能获得完整的数据，因为这个网页是一个通过JS动态生成的动态网页，很多数据并不能通过爬虫直接得到所以只能换其他方法进行数据获取。

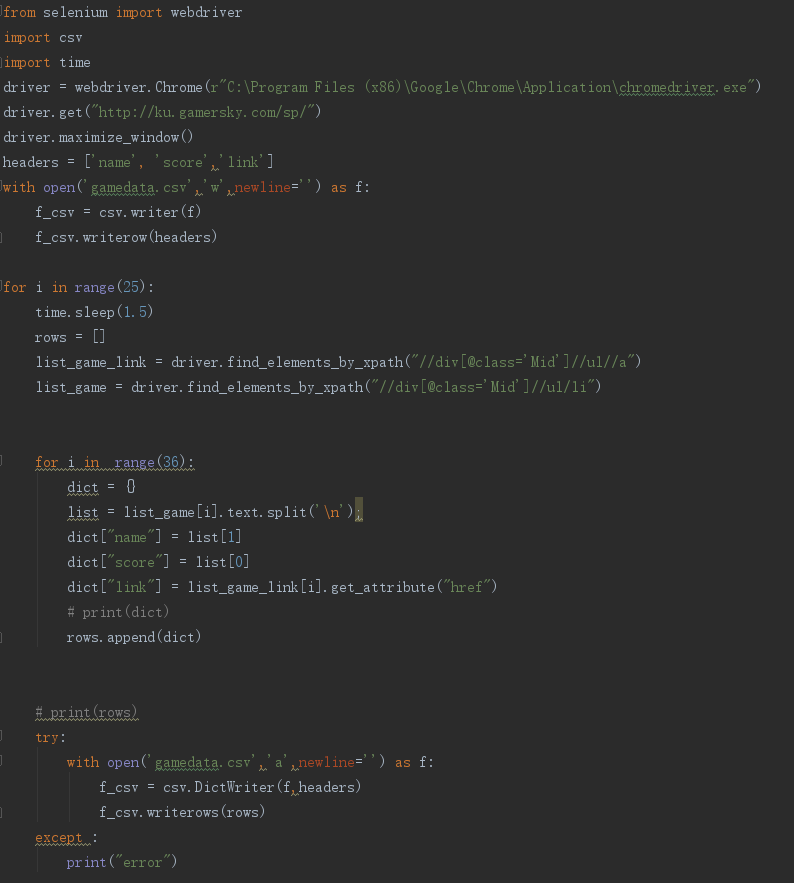
1. **Selenium+ChormeDriver自动化网页测试工具**

通过网上搜集方法，发现通过**Selenium+ChormeDriver**的方式可以自动化网页测试，可以获取动态数据。

1. **首先测试爬取了首页游戏名，游戏评分和游戏url数据并将数据存入csv文件：**

主要是通过工具读出每一页的网页数据，用xpath语法进行数据的筛选，然后存入gamedata.csv文件中，然后会自动翻页读取下一页网页的数据，

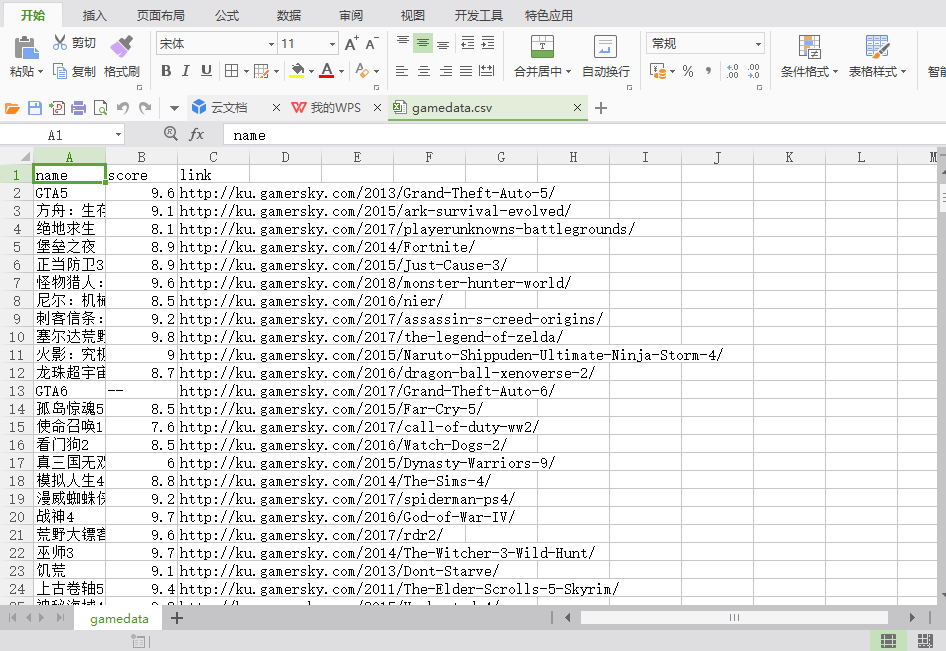
**具体代码实现：**



**数据爬取结果：**

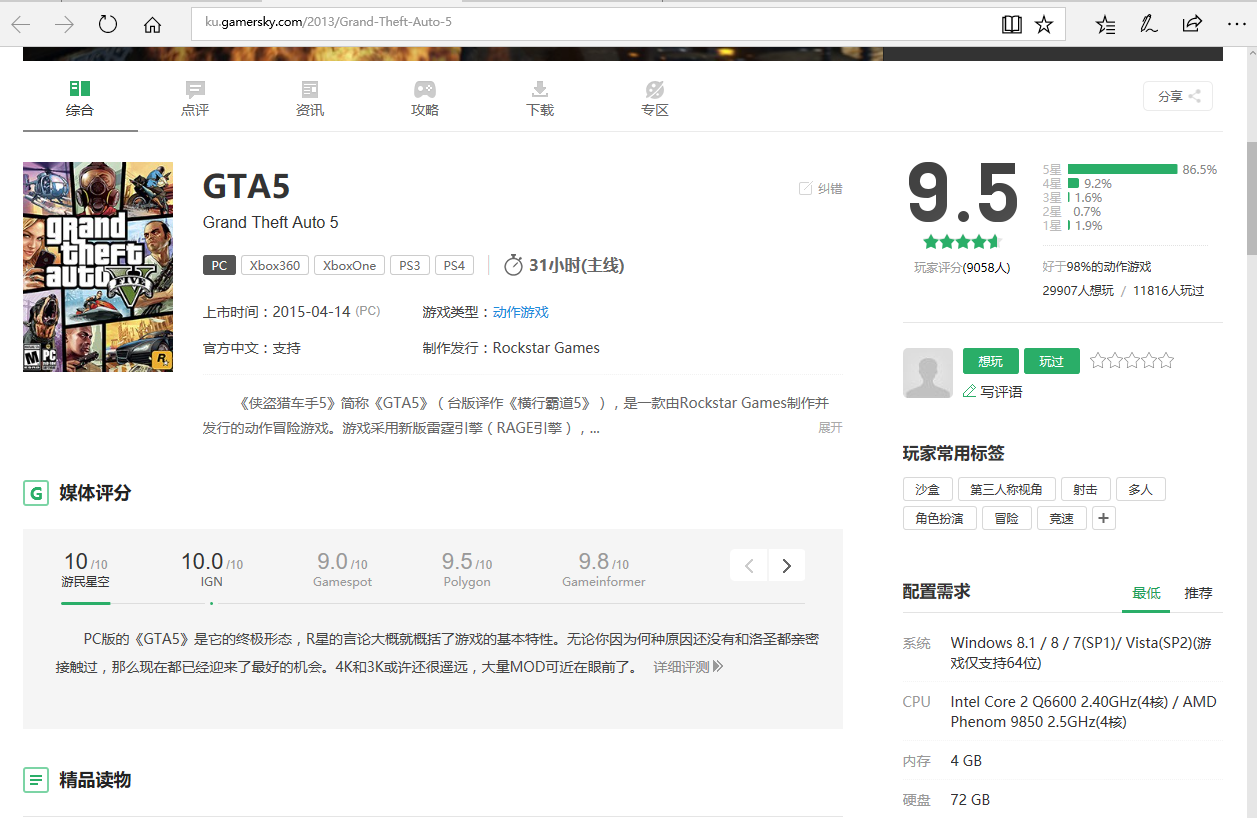
共901条数据 包括游戏名，游戏评分，游戏url,

排序序为网站根据游戏的热度进行排序的，与评分无关



1. **爬取每个游戏更多的具体数据：**

仅得到游戏的名字，和评分数据是远远不够的，还需要得到更多游戏类型，评价等数据才能进一步分析数据。所以通过游戏url进入每个游戏具体的界面进行游戏的爬取：



通过这个界面可以得到更多的关于游戏的数据，包括游戏的类型，制作方，游戏的标签，游戏的热度数据。

爬取数据的方式与之前一样，但需要借助第一次爬取数据的游戏url,通过进入每个URL，爬取页面的数据并存入gamedata\_detail.csv文件中：

**具体代码实现：**



**数据爬取结果：**

共791条数据包括游戏名，游戏类型，游戏制作方，游戏期待人数，游戏玩过人数，以及玩家对该游戏打的标签：



1. **数据清洗：**

对于本实验研究的游戏均是已经发行的热门游戏，对于一些尚未发行的游戏和一些评分人数不足的冷门游戏应该进行清除。

具体实现代码：



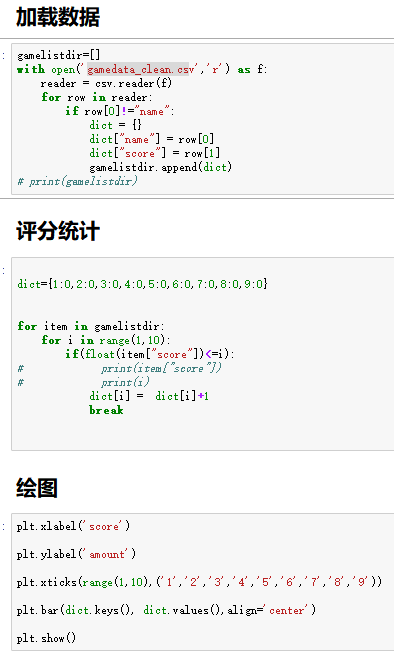
1. **数据统计分析：**

因为对于很多数据统计算法不是很熟悉，所以本次实验对于数据的统计更多是对于词频，出现次数，并通过可视化的方式(词云，图表)呈现出来。统计分析的代码通过ipynb文件书写，可以直接运行查看效果：

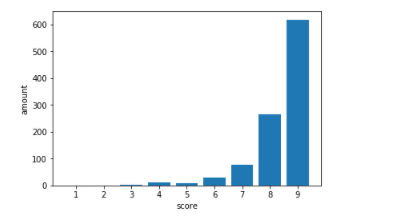
1. **评分区间统计**

加载gamedata\_clean.csv数据，统计评分的分布：

**具体代码：**



统计结果：



可以看出大部分的热门游戏评分都分布在，8分以上，毕竟一个游戏的质量才能决定他的热门程度

1. **游戏标签统计**

每个游戏有多个标签，通过对每个标签出现次数的统计，可以看出玩家对游戏的什么方面最感兴趣

**具体实现代码:**



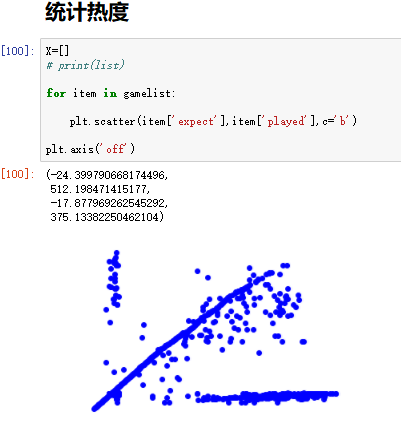
**输出结果：**



通过统计结果，发现玩家最感兴趣的标签竟然是**奇幻，**我以为统计结果会是**动作**，或是**冒险,**通过最终的数据可以看出玩家更期待在游戏中找到**奇幻的单人冒险**，他们希望从游戏中获得一段不一样的经历，而不仅仅是媒体等报道中所说追求**暴力。**只是我之前没有想到过的结论。

1. **统计期待人数和玩过人数的关系：**

将**期待人数**和**玩过人数绘制在图上可以看出他们之间的关系**



游戏的热度情况大致分三类，热门游戏，热门新游戏，小众热门游戏

可以看出大部分的**热门游戏**的期待人数和玩过的人数呈线性关系，；但是还是有部分期待人数明显大于玩过的人数的**热门新游戏**；以及玩过的人数大于期待人数的**小众热门游戏**

1. **其他数据的统计情况**



