**湖南女子学院**

**课程报告**

­­­­­­­­­­­­­­­



|  |  |
| --- | --- |
| **课程名称** | 大数据技术原理与应用 |
| **题 目** | 招聘网站大数据分析 |
| **系 部** | 信息科学与工程学院 |
| **专业班级** | 18级计算机电子商务一班 |
| **学 号** |  |
| **学生姓名** |  |
| **指导教师** | 刘 广 |
| **职 称** | 讲 师 |

**2021 年10 月**

# 大数据技术原理与应用课程报告

## 大数据

大数据的概念：

大数据是由数量巨大、结构复杂、类型众多的数据结构的数据集合，在合理时间内，通过对该该数据集合的管理、处理、并整理成为能帮助政府机构和企业进行管理、决策的讯息。

大数据的发展历程：

1、上世纪末，是大数据的萌芽期，处于数据挖掘技术阶段。随着数据挖掘理论和数据库技术的成熟，一些商业智能工具和知识管理技术开始被应用。

2、2003年-2006年是大数据发展的突破期，社交网络的流行导致大量非结构化数据出现，传统处理方法难以应对，数据处理系统、数据库架构开始重新思考。

3、2006年-2009年，大数据形成并行计算和分布式系统，为大数据发展的成熟期。

4、2010年以来，随着智能手机应用，数据碎片化、分布式、流媒体特征更加明显，移动数据急剧增长。

5、2011年麦肯锡全球研究院发布《大数据：下一个创新、竞争和生产力的前沿》，2012年维克托·舍恩伯格《大数据时代：生活、工作与思维的大变革》宣传推广，大数据概念开始风靡全球。

6、2013年5月，麦肯锡全球研究所发布了一份名为《颠覆性技术：技术改进生活、商业和全球经济》的研究报告，报告确认了未来12种新兴技术， 而大数据是这其中需求技术的基石。

7、2014年5月，美国白宫发布了2014年全球“大数据”白皮书的研究报告《大数据：抓住机遇，守护价值》。报告鼓励使用数据推动社会进步。

大数据的作用和技术：

1、在大数据采集与预处理方向。这个方向最常见的问题是数据的多源和多样性，导致数据的质量存在差异，影响到数据的可用性。

2、大数据存储与管理方向。这个方向最常见的挑战是存储规模大，存储管理复杂，需要兼顾结构化、非结构化和半结构化的数据。分布式文件系统和分布式数据库相关技术的发展正在有效的解决这些问题。其中大数据索引和查询技术、实时及流式大数据存储与处理的发展。

3、大数据计算模式方向。如今出现了多种典型的计算模式，包括大数据查询分析计算、批处理计算，流式计算、迭代计算、图计算、内存计算。

4、大数据分析与挖掘方向。在数据类迅速膨胀的同时，还要进行深度的数据分析和挖掘，因此越来越多的大数据分析工具和产品应运而生。

5、大数据可视化分析方向。通过可视化方式来帮助人们探索和解释复杂的数据，有利于决策者挖掘数据的商业价值，进而有助于大数据的发展。

6、大数据安全方向。当我们在用大数据分析和数据挖掘获取商业价值的时候，黑客很可能在向我们攻击，收集有用的信息。通过文件访问控制来限制呈现对数据的操作、基础设备加密、匿名化保护技术和加密保护等技术正在最大程度的保护数据安全。

1. **爬虫实训**

爬虫的概念：

简单来讲，爬虫就是一个探测机器，它的基本操作就是模拟人的行为去各个网站溜达，点点按钮，查查数据，或者把看到的信息背回来。就像一只虫子在一幢楼里不知疲倦地爬来爬去。

爬虫的具体应用：

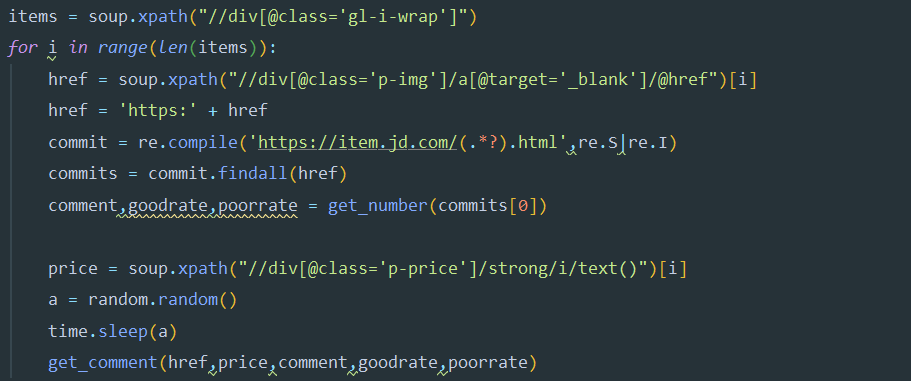
首先我们的目标网站是爬取京东的一些基本信息来根据这些重要信息来进行数据分析，数据可视化得出合理的建议

在这次目标中我们主要的用到的是如下：

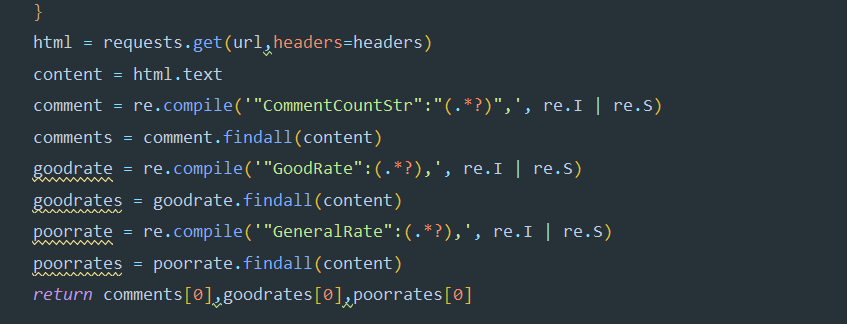


其中最主要的是requests库和lxml库，requests是第三方请求库，用于请求该页面的信息获取对应的内容，lxml库对应的就是把获取到的信息进行格式化，方便我们用xpath语法来进行信息的定位，从而获取我们需要的内容

数据的获取：

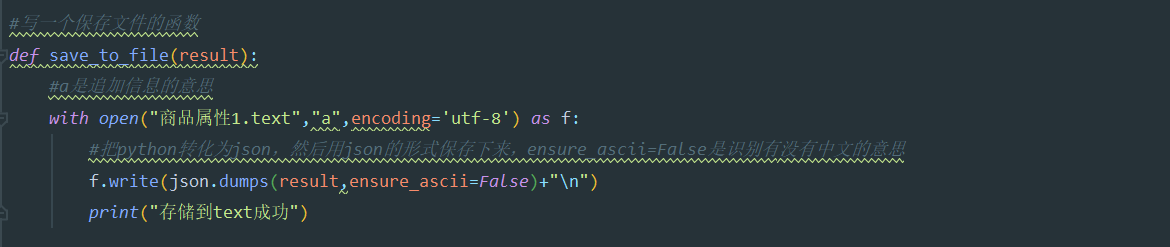


我们通过xpath语法的定位来获取对应页面的连接，去获取基本信息和价格，然后再用正则表达式去定位好评和差评的数量

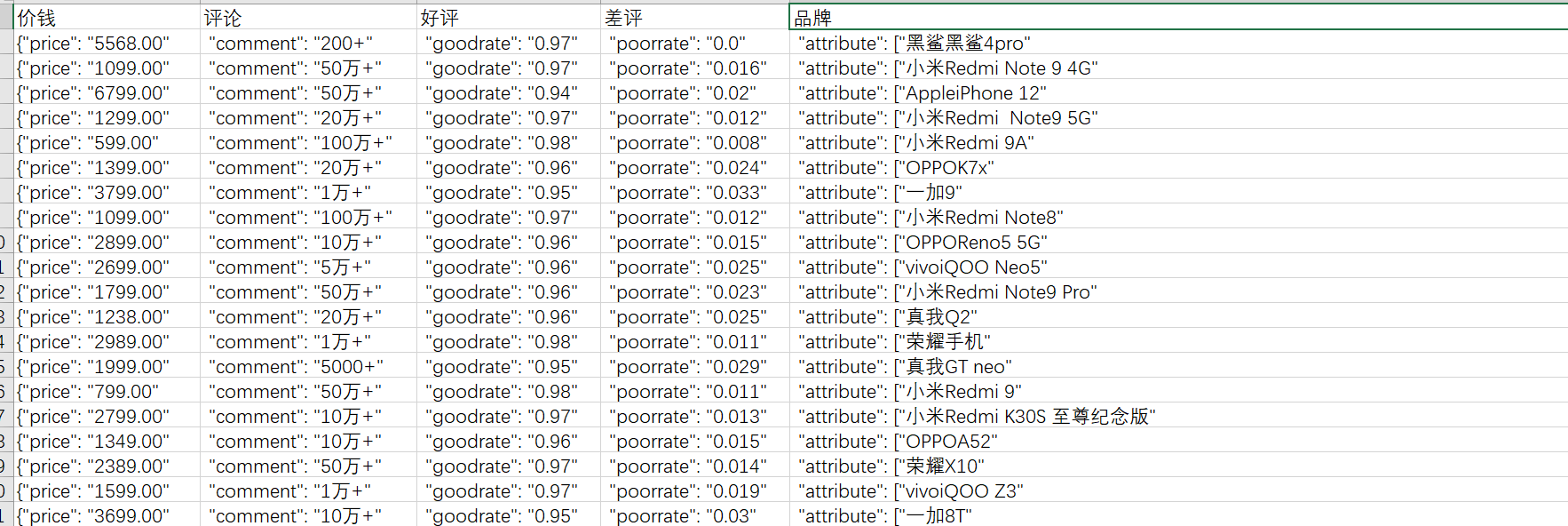


数据保存：

当获取完上面的数据之后，我们对数据保存，我们首先先用txt文件把数据保存下来



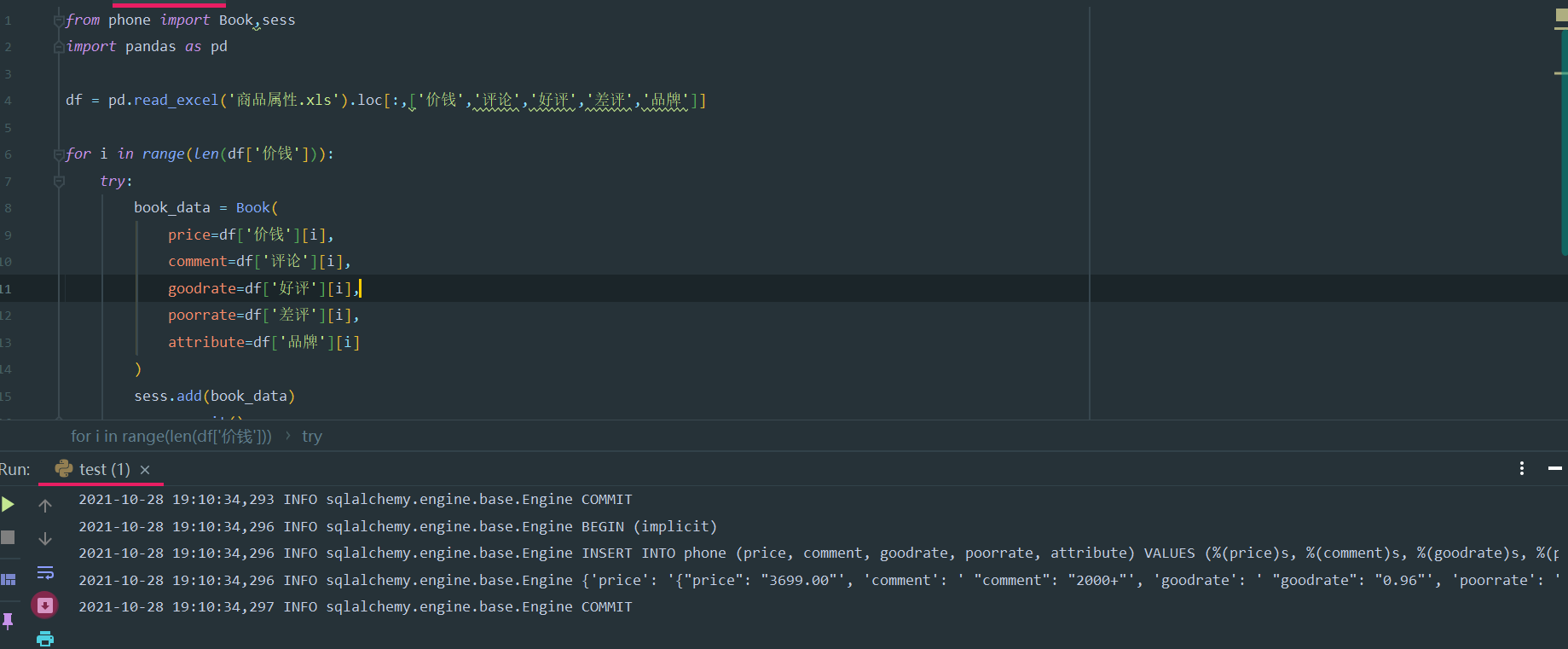
然后再对数据进行转换成CSV文件



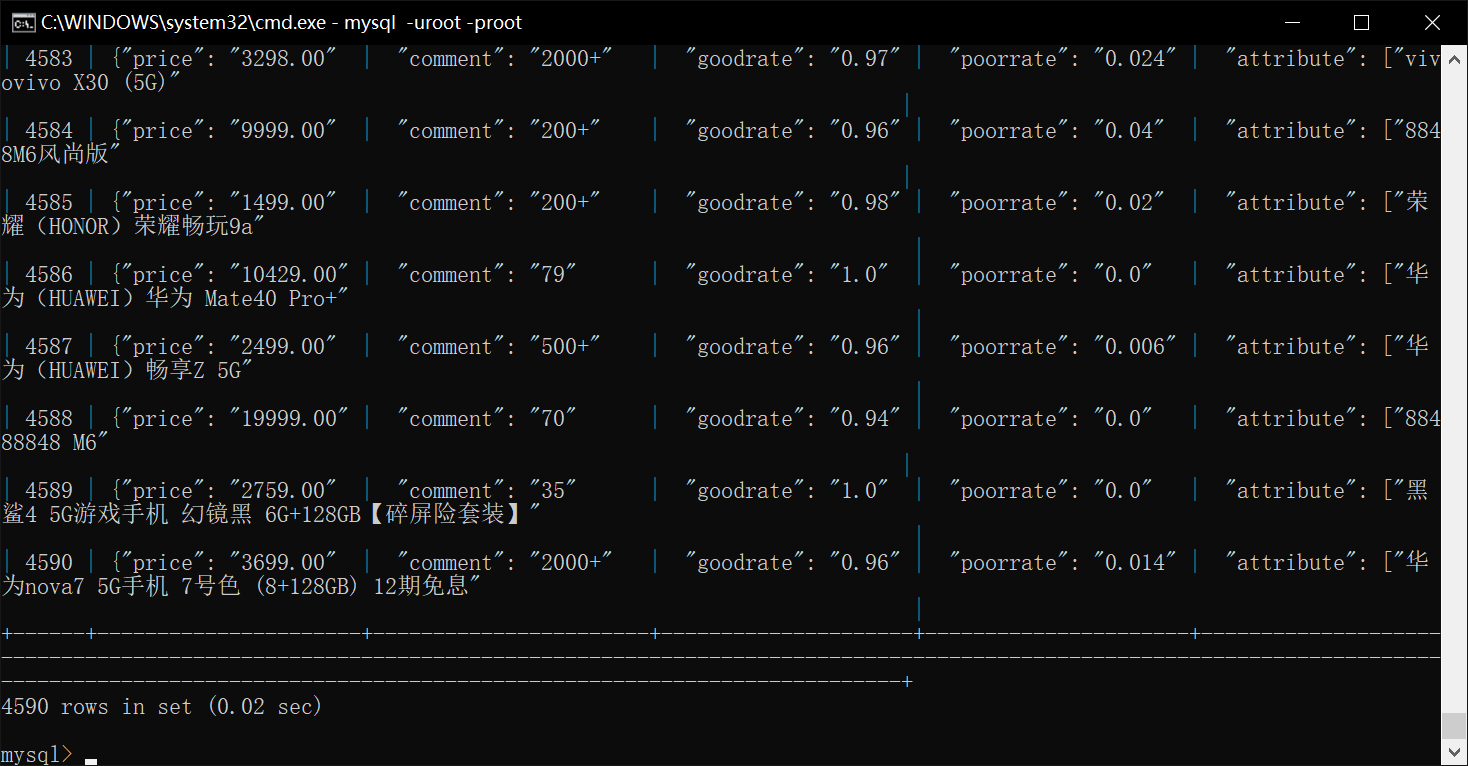
然后再创建一个数据库



把信息用数据库保存下来，防止数据丢失问题，再把数据导入进数据库



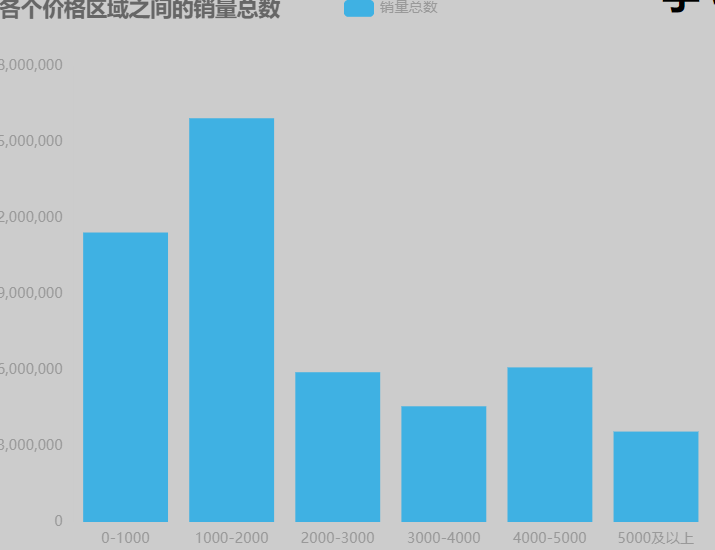
然后再去我们的数据库进行查看



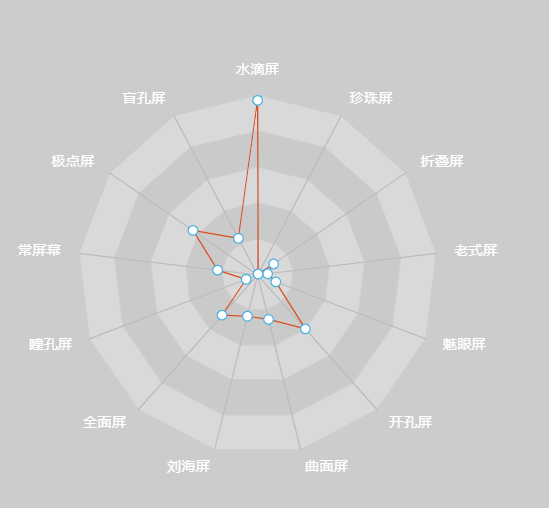
信息很好的被保存下来了

## 数据分析、数据可视化

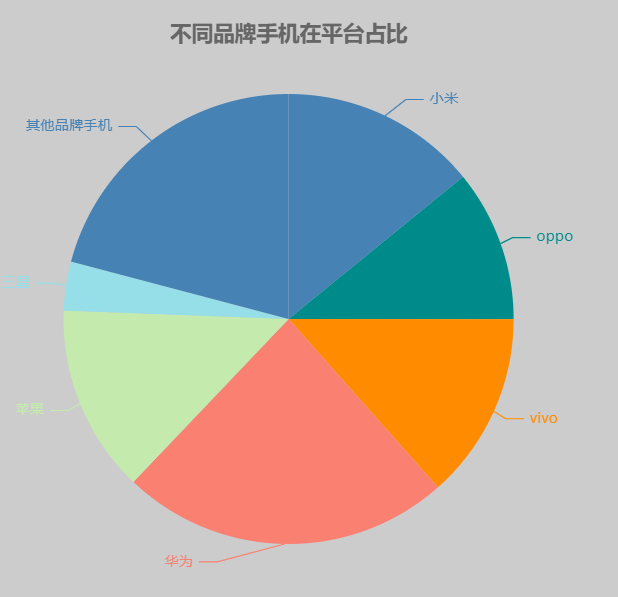
接着我们再对我们的数据进行可视化和数据分析处理



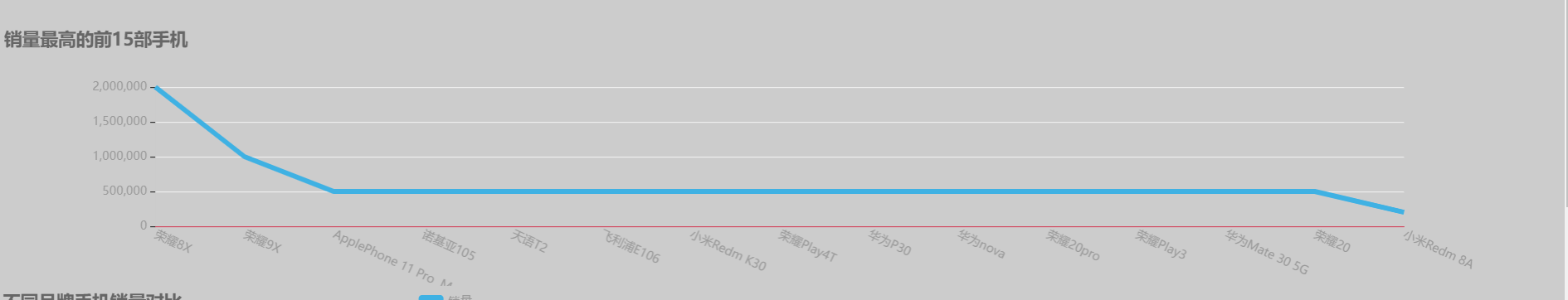
由图可知，手机销量最好的是在1000-2000之间，说明日常人们使用手机，主要还是追求性价比高一点没有盲目追求高科技高消费的手机



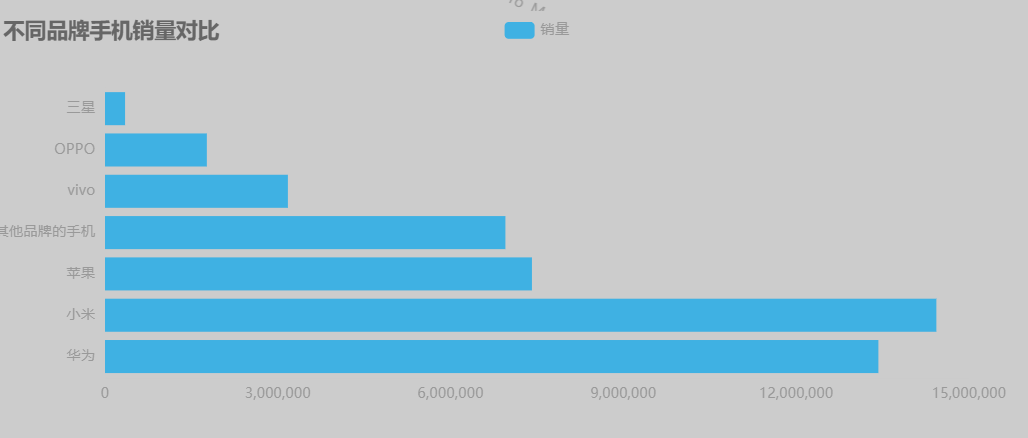
由这个雷达图可知，目前我国的国人大多数喜欢水滴屏的手机为主



在众多手机品牌中，我国华为还是市场占有份额还是最大的



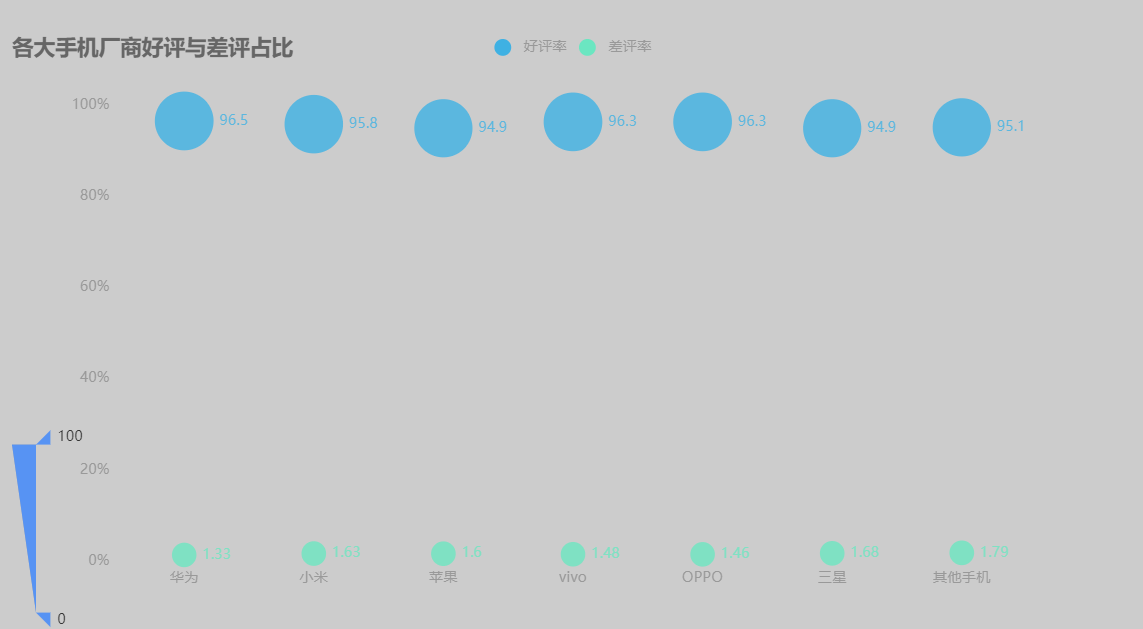
由销量最高的前15部手机也可以看出，华为和荣耀占有的市场份额最大，前15名，华为荣耀就占了9名



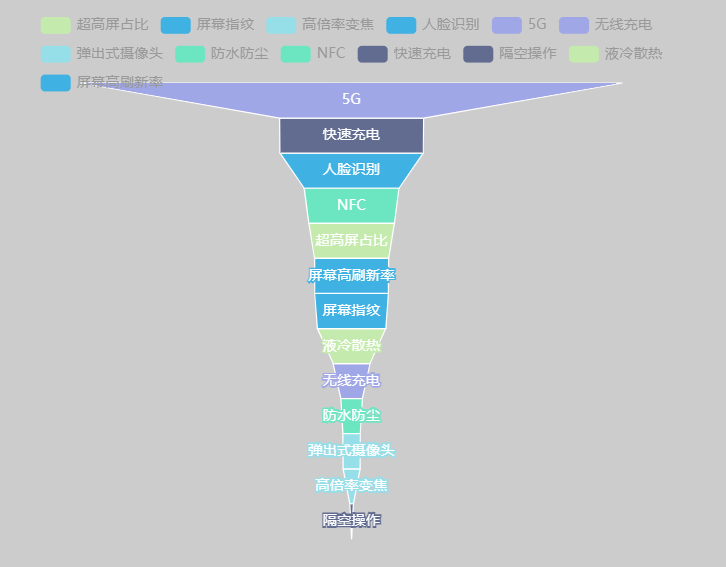
在不同品牌销量中，小米位居第一，当然这也是和它的价格有关，小米在低端手机的市场占有份额较大



而根据词云图来说，目前人们最关心的还是这几个点，像素，5G,网络这些都和我们日常使用手机功能息息相关，说明手机在人们日常生活中扮演着越来越重要的角色

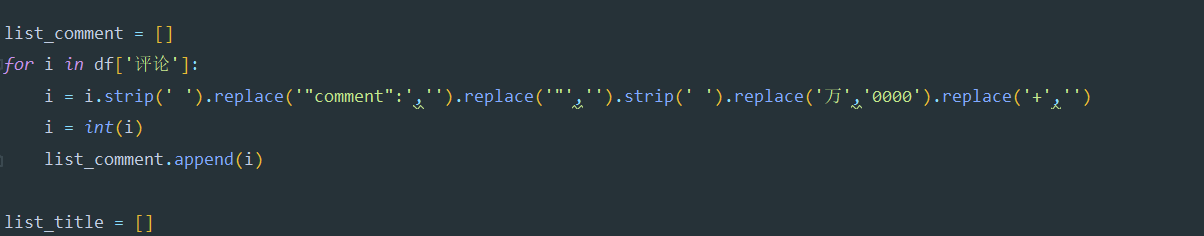


而手机各大厂商的好评率，华为还是老大哥，好评位居第一，不过小米也不差和华为只是差距0.7



最后在众多卖点中，还是5G影响力最大

在数据清洗中，有众多的标点符号需要我们一个个去清空



## 四、课程结论、心得体会

在此次实训中，我们运用大数据技术对京东电子商务网站进行大数据的分析，熟悉了大数据技术原理、爬虫技术、数据的清洗、数据分析、数据可视化等内容，掌握了一些有关大数据分析的方法和技能，为我们以后在工作事业中运用IT技术打下了良好的基础。

在实训中，我们不仅掌握了基础技术知识，也学会了许多解决问题的方法，当遇到问题，冷静，想办法一点一点的排除障碍，到最后获取成功，一种自信心就由然而生，这应该就是工作的乐趣。我们还学会了在工作中与人的合作与交流，同乐同累，合作互助，这是团体的精神，也是我们必须要学习的东西，也了解到大数据这一方面的很多相关知识和具体案例通过这次实训不仅提升自己了代码水平而且也进一步扩展了自己对大数据这一方向的兴趣