**SnowNLP｜2｜爬取评论情感分析**

1. **基本环境**

|  |  |
| --- | --- |
| 程序名称 | getComment.py  getModel.py  pieChart.py  snowanalysis.py  iphone\_wordcloud.py |
| 硬件环境 | Win |
| 软件环境 | Python3, Anaconda |

1. **核心问题：**爬取京东某商品评论，并用开源中文情感分析库SnowNLP进行情感分析
2. **解决过程：**
   1. **爬虫**

利用 python requests + re 爬取京东iphone8 plus商品下的共计500页，以及其iphone8plus 的型号（颜色，存储大小）1000组，存入txt。

对评论数据进行了简单的清洗，去除了带有图片的评论中带有的html代码：

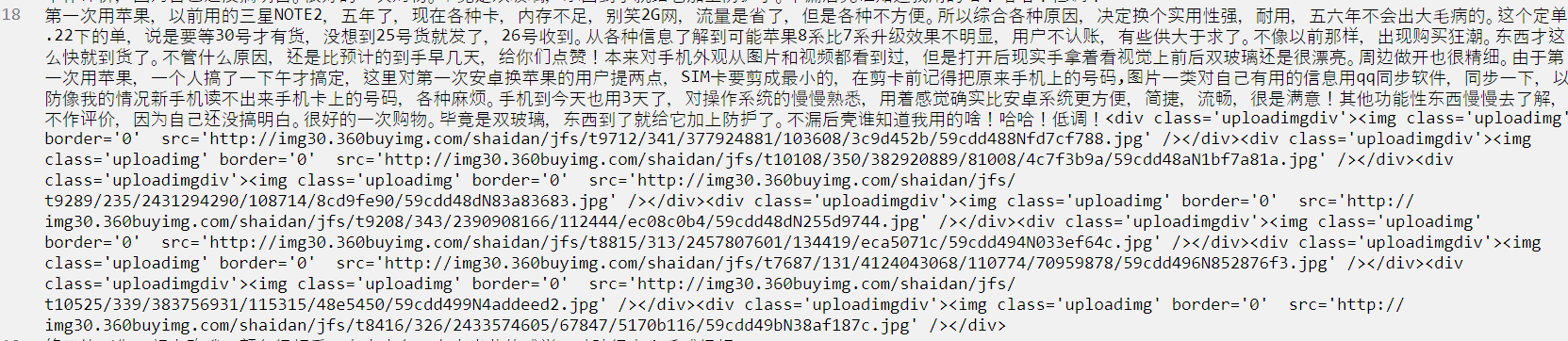


图 1 原图

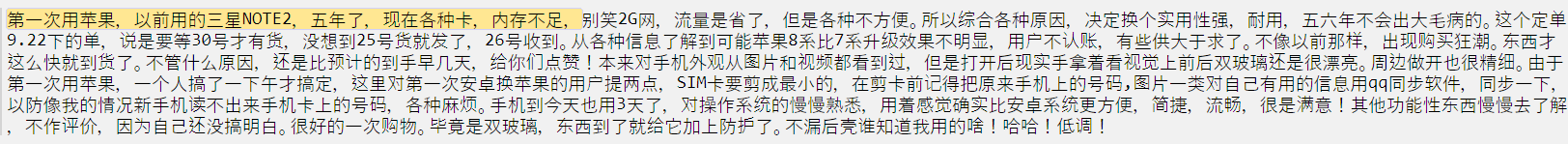
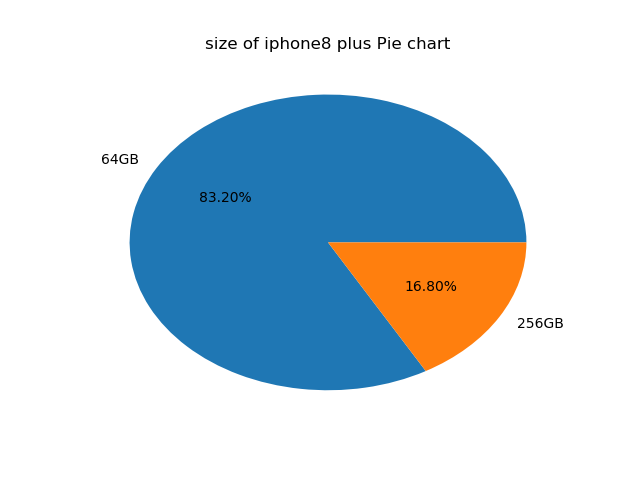
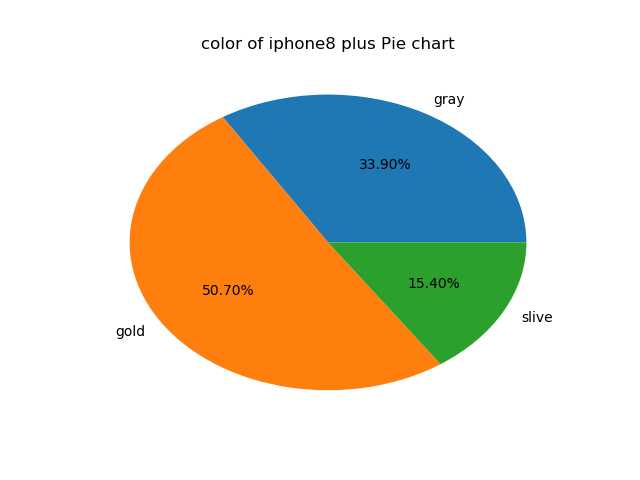


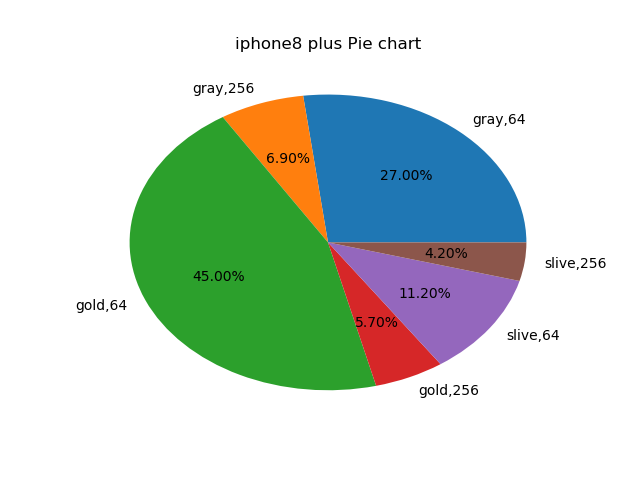
图 2 清洗后数据

* 1. **iphone8 plus各个颜色、型号的受欢迎程度**

 利用matplotlib 画饼图进行分析：

可以看出金色是最热销的颜色，占到销量的一半以上，而银色仅占到整体的15.4%。

从存储容量来看，64GB版本是80%以上消费者的选择，而只有16.8%的消费者选择显得容量过大的256GB版本。



综合各个型号的饼图，金色64GB版本是最热销的版本的iphone8plus。

* 1. **对评论制作词云和进行情感分析**
     1. **情感分析**

利用开源中文情感分析库snowNLP(<https://github.com/>

isnowfy/snownlp)对评论进行情感分析。

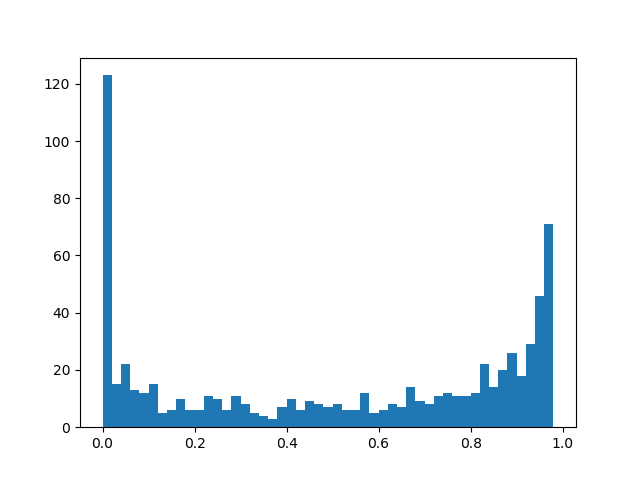
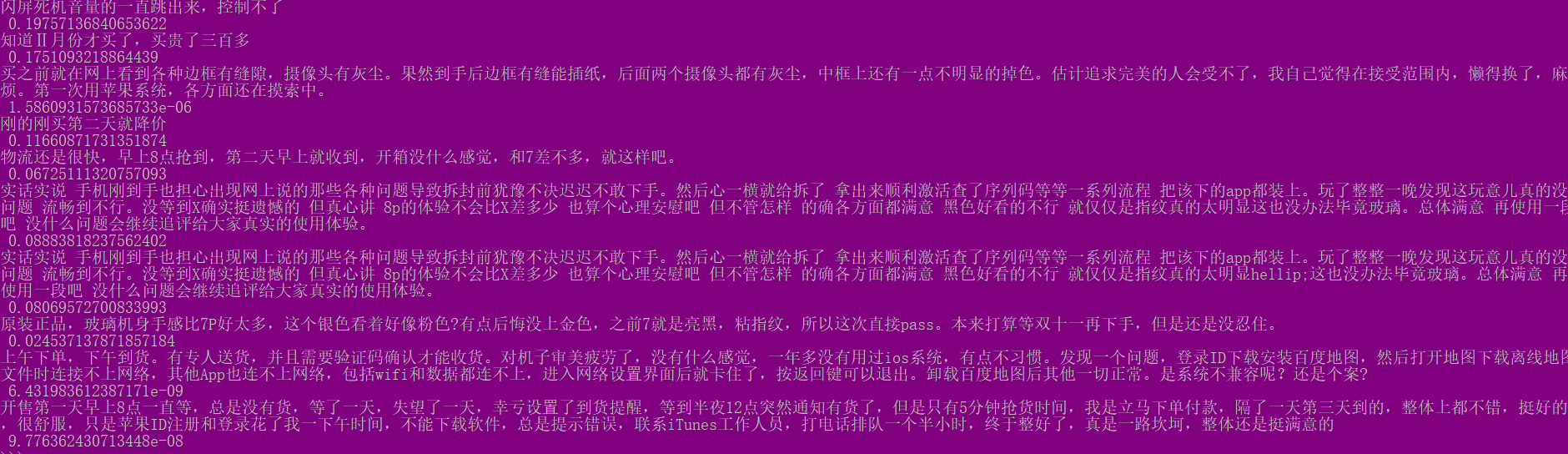


图 3情感分析结果

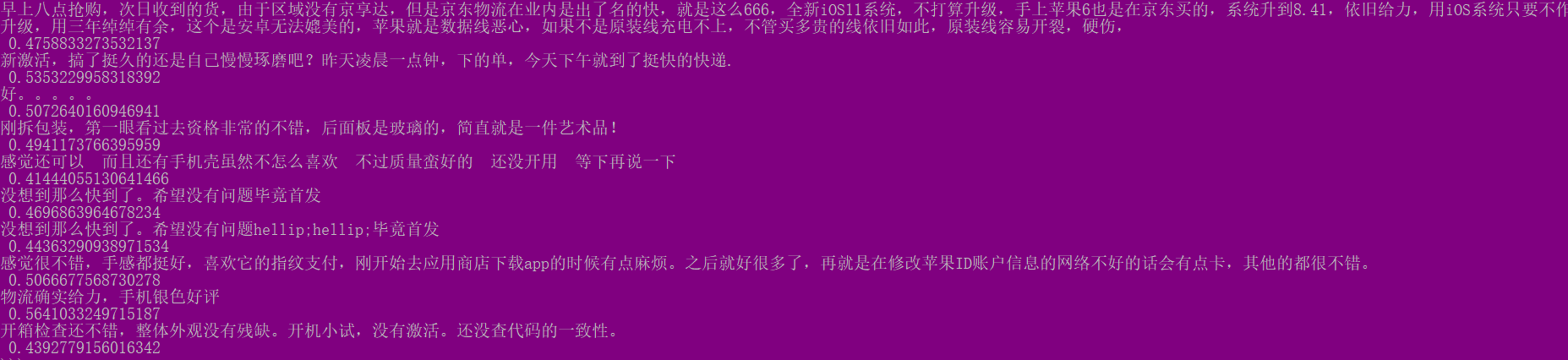
注：0-1的区间，越靠近1越积极，靠近0为消极。

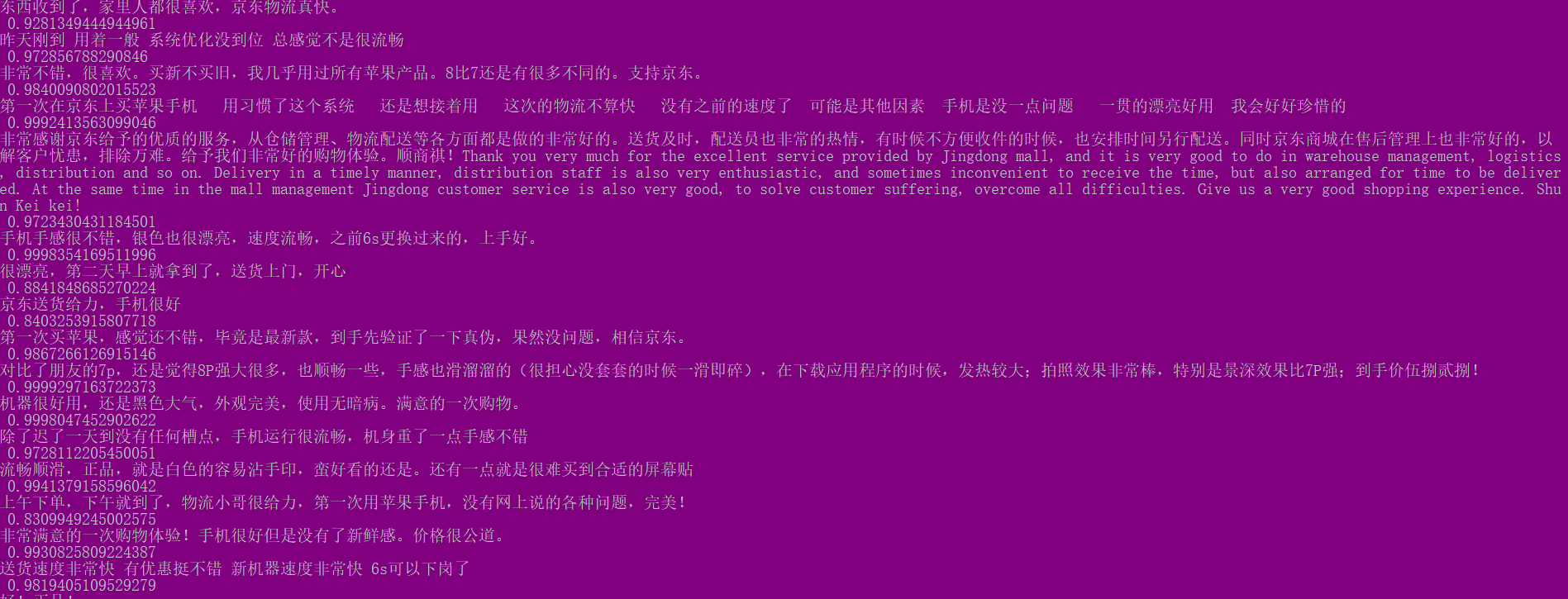
图中出现较为明显的两极分化的评论，选取部分区间评论查看他们的评价。

[0-0.2]展示部分：



[0.4-0.6]:



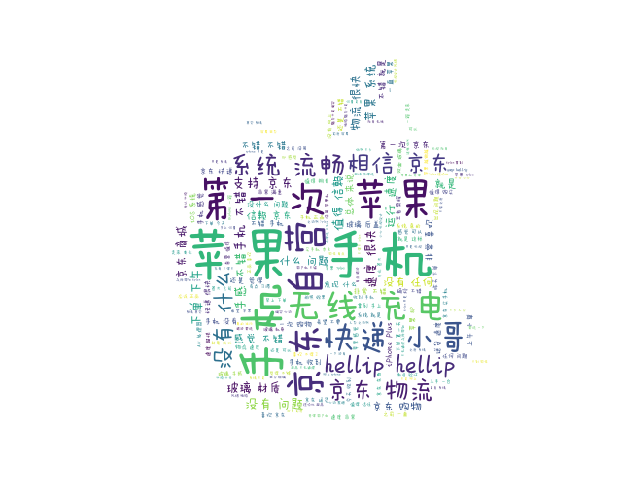
[0.8-1.0]:

从部分实例看出， 在[0.8-1.0]及[0.4-0.6]及中立及积极区间内的评价较为准确，但在消极[0-0.2]区间内的评论有比较严重的失准情况出现。在对手机某些特征如手感、外观、材质等情感分析准确度较低。

查看snowNLP的原始训练集发现，两端数据集中没有关于手机等消费电子产品的数据，可能是上述情况产生的原因。

* + 1. **词云**

使用 jieba分词 和 wordcloud建立词云：



“京东”、“自营”、“快递小哥”出现得最多， 但这些词汇与iphone8Plus 无关， 而是对购物平台京东的评价。原因可能是在数据处理时没有处理反水军评论，无法确定这些好评中的水分。但这个现象还是印证了生活中大家对于京东的评价甚高，如自营的正品保障以及快递速度高等等。

寻找关于iphone8plus的词汇可以看到“系统流畅”、“玻璃材质”、“无线充电”等，这些都是iphone8plus的重要卖点。

当然字体最大的苹果手机和苹果可以看出，消费者对于苹果这个品牌的认同感。