# 《5g浪潮下新一代手机在线评论研究》

# ——以京东在线评论为例

### 摘要

生活在这个万物互联的时代，人们时常在网络上发表各式各样的在线评论。而人们在购物后于购物网站留下的在线评论蕴含着大量和用户以及产品相关的有用信息，这些信息无论对于商家还是在线平台管理而言都蕴有重大的商业价值。在线评论——这里专指买家留下的文本形式的评论部分——属于口碑的范畴。众多研究都有力的证明，用户在购买产品时往往参考口碑信息（在线评论），而买家的购买行为以及后续对商品的体验以及评价都受到口碑的影响。因此重视在线评论数据是十分必要的，而人们也已经在文本分析领域提出了许多算法。

2019年，5G产业引来了商用落地之年。5G不仅为全球数十亿人提供高速连接的互联网新形态，更以实时、高速和支持多设备并发的诸多优势特性，成为万物互联的新型关键基础设施。工业互联网、车联网、智能电网、智慧城市、军事网络等都将构架在5G网络上裂变发展。而大众最多和最直接接触到的无疑便是5g手机了。

本文爬取了24种近月来的热门机型，包括了华为、小米、redmi、oppo、vivo、galaxy、魅族、一加八等手机款型的在线评论，在进行词频统计的基础上制作了词云，此外使用LDA主题模型对用户好评和差评进行了情感分析。

##### 关键词： 在线评论；python爬虫；词云；LDA模型

目录

[《5g浪潮下新一代手机在线评论研究》 1](#_Toc44259298)

[——以京东在线评论为例 1](#_Toc44259299)

[摘要 1](#_Toc44259300)

[1 背景介绍 3](#_Toc44259301)

[2 简要分析 5](#_Toc44259302)

[3 实验流程 7](#_Toc44259303)

[4 数据分析 11](#_Toc44259304)

[参考资料 18](#_Toc44259305)

### 1 背景介绍

##### 1.1 在线评论

得益于互联网的飞速发展，各式各样的开放平台兴起——社交媒体大行其道、电子商务几家争霸、论坛博客等数不胜数……于是人们在网络留下了各式各样的痕迹。其中人们在购物后于购物网站留下的在线评论蕴含着大量和用户以及产品相关的有用信息，这些信息无论对于商家还是在线平台管理而言都蕴有重大的商业价值。譬如通过在线评论、平台运营者可以把握产品的质量情况，了解用户对产品的喜好和购买欲望，从而更好地与相似产品进行分析对比，从而规划产品的未来发展战略；此外通过在线评论，商家能够通过文本挖掘解析并统计有关产品属性信息，并从中找到买家更为关注的属性和产品不足的属性，在后续进行优化。

在线评论——这里专指买家留下的文本形式的评论部分——属于口碑的范畴。大量的研究证明（而这也是显而易见的），用户在购买产品时往往参考口碑信息（在线评论），而买家的购买行为以及后续对商品的体验以及评价都受到口碑的影响。

##### 1.2 第五代移动通信技术

5G，即第五代移动通信技术（5th generation mobile networks或5th generation wireless systems、5th-Generation，简称5G或5G技术）是最新一代蜂窝移动通信技术，也是即4G（LTE-A、WiMax）、3G（UMTS、LTE）和2G（GSM）系统之后的延伸。2019年，5G产业引来了商用落地之年。5G不仅为全球数十亿人提供高速连接的互联网新形态，更以实时、高速和支持多设备并发的诸多优势特性，成为万物互联的新型关键基础设施。工业互联网、车联网、智能电网、智慧城市、军事网络等都将构架在5G网络上裂变发展。

众所周知，5G网络将在无人驾驶、远程高清医疗、远程会议、8K直播、VR电影和游戏、云电脑等方面有更多的突破和应用。理论而言，5G的优势主要有高网速，低延迟，多链接：

* 5G理论上其传输速度是4G网络的几百倍，可以达到每秒钟数十GB的传输速度，能够轻松实现应用于全景视频传输，图像更为清晰，画面也更加流畅，如能结合VR终端能带来更好地身临其境的效果；
* 5G网络延迟时间将降低到 1 毫秒以下，在视频通话、工业物联网、无人驾驶、商用无人机等领域将发挥举足轻重的作用；
* 5G网络将人与物，物与物等之间的连接，使用网络技术将城市设施、家居生活、物流状态等融于一体，真正实现“万物互联”。

在ToC领域，在5G时代的前1-2年，5G能给个人带来的最大便利，只有一个：高网速。而这一优势的直接体现，无疑便是5g手机了。

即便受到疫情影响，各手机厂商也早在三四月便推出了大量5g机型，趁着5g尚热大捞其财——五六月各5g手机几乎不是降价而是砍价的行为更是让人大跌眼镜。

本文爬取了24种近月来的热门机型，包括了华为、小米、redmi、oppo、vivo、galaxy、魅族、一加八等手机款型的在线评论，在进行词频统计的基础上制作了词云，此外使用LDA主题模型对用户好评和差评进行了情感分析。

##### 1.3 文本情感分析

文本情感分析，又称意见挖掘(Opinion Mining)，是指对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程[1];情感分析属于计算语言学，情感分析运用到人工智能，NLP，信息检索，机器学习等多个领域的知识。如从淘宝上的商品评论分析某一件商品是好是坏，从豆瓣电影评论分析用户对某部电影的喜爱程度等。

情感分析一般包括几个步骤，即情感信息的抽取、情感信息分类以及情感信息的检索与归纳。

情感信息抽取即抽取文本中能表达情感的信息，带有情感倾向的词语称情感词，比如：喜欢、讨厌、恶心等表达用户情感的词语。目前对情感词的抽取主要有两种方法，一种基于语料库，另一种基于词典。

情感信息的分类包括两种，主客观信息分类和主观信息分类。主客观信息分类即将文本分为主观信息和客观信心两类，其依据是文本中是否包含情感信息。主观信息分类即对情感的分类，通过各类分类器将情感文本分类为褒贬，或者其他更多的类别。

情感信息的检索与归纳，即检索时，不仅检索包含用户主题的文档，也检索包含情感信息的文档。在搜索过程中，搜索所得到的结果不仅依据传统的信息检索模型，并且对文档中进行主客观识别，对文档表达的情感打分，再结合文档相关性来对检索结果排序。

本文使用python扩展库corpora来建立词典和建立情感词库，并通过LDA模型来分析用户情感。

### 2 简要分析

就本文而言，只是对在线评论进行爬取后进行简单的分析，透过这些粗浅分析验证和探究以下几个问题：

* 对于各大手机厂商在新一代5g手机市场方面的表现如何？
* 对于用户的一些行为偏好进行分析：特定群体对特定颜色的偏爱、评分倾向等
* 各机型的用户体验如何？
* 各机型大家关注的点在哪里？
* 各机型的缺点在什么地方？

##### 2.1 平台和机型选择

由于之前从未接触数据库、爬虫的相关知识，因此在网上搜出的第一篇教程便让本人选择了京东。本还想爬淘宝的评论，但是实在力不从心，京东的评论数据很多浅显易做的工作也没有做。

对于机型选择，在进行网上搜索后，在题为《2020年（6月）5G手机盘点和挑选指南》的文章下选择了下面这张图所包含的机型。除了少数京东平台购买量较少评论不足和寥寥几种非常用机型外都做了数据爬取，共爬取了取了24种机型。



图1 机型选择

##### 2.2分析工具选择

Python编程语言十分适合爬取网络数据，此外网络上的教程几乎都是基于python的。除了易于爬取数据之外，python还有大量的扩展库用来分析各式各样的数据，本文使用python来完成实验。

### 3 实验流程

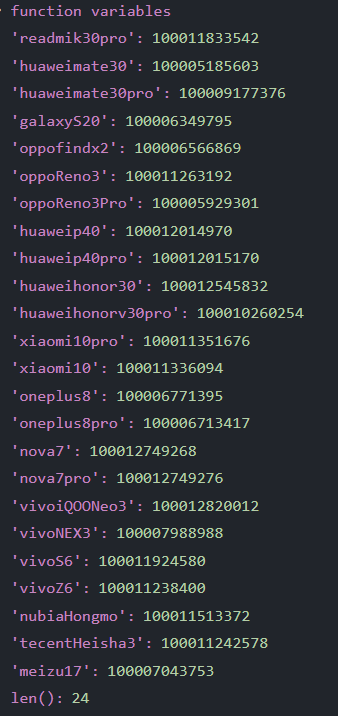
本部分对本文实验的流程进行了介绍，包括数据选择与爬取、数据整体分析、数据可视化、lda情感分析

##### 3.1 数据选择与爬取

如前所述，本文使用python语言编写爬虫，爬取了24种近月来的热门机型，包括了华为、小米、redmi、oppo、vivo、galaxy、魅族、一加八等手机款型的在线评论。

3.1.1 产品id

在选择了24种机型后，从网页找到24种产品的id，如图2所示。



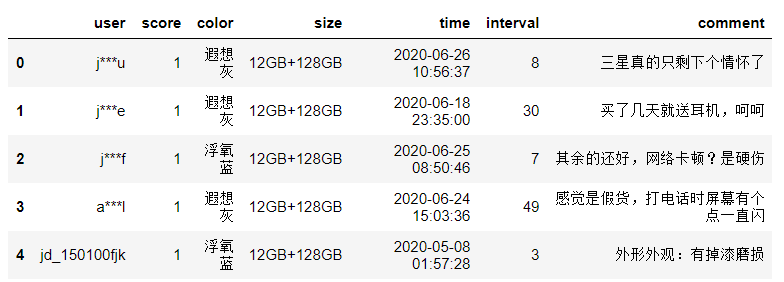
图二：产品id

3.1.2 爬取的内容和方式



图三：爬取的内容和方式

由于京东只能得到前100页的评论，每页包含十个评论，因此选择对评分从1到5分别爬取前100页，数据保存内容选择了'user', 'score', 'color', 'size', 'time', 'interval', 'comment'，其中time为评论的时间，interval为购买到评论以天为单位的时间间隔。数据格式如下：

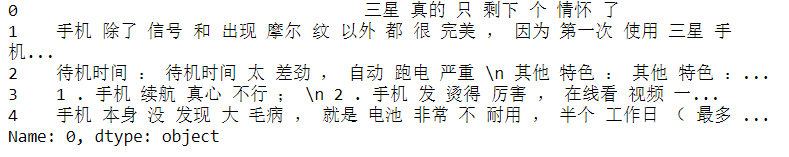


图四：数据demo

##### 3.2 数据预处理

3.2.1 分词

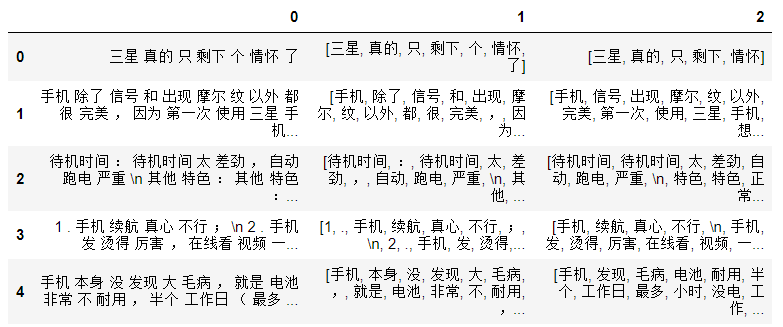
使用python的jieba库对评论内容进行分词，效果如下



图五：分词

3.2.2 去除停用词

停用词(Stop Words) ，词典译为“电脑检索中的虚字、非检索用字”。在SEO中，为节省存储空间和提高搜索效率，搜索引擎在索引页面或处理搜索请求时会自动忽略某些字或词，这些字或词即被称为Stop Words(停用词)。在自然语言处理中，停用词一般不携带有价值的信息，我们选择去除掉这些词。效果如下：



图六： 去除停用词

##### 3.3词云绘制

本文使用了中文分词和词云生成器“微词云”，网址为https://www.weiciyun.com/，该生成器直接对文本进行分割和词频统计，然后绘制词云。效果如下：

. 

图7 词频统计示例 图八 词云示例

##### 3.4 LDA情感模型

3.4.1 数据预处理

这一部分与3.2相同

3.4.2 构建情感词典并抽取情感词

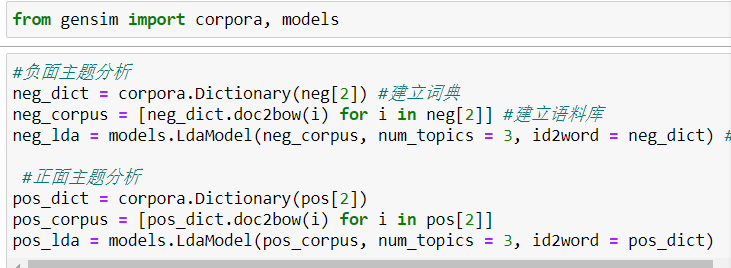


图9 词典和语料库建立以及LDA模型

如图9所示，本文使用python扩展库gensim.corpora来建立词典和情感词语料库。

3.4.3 构建LDA模型

如图9所示，本文使用python扩展库gensim.models种的LDA模型生成情感分析模型，正面负面都选择了三个主题，正面选择了评分为5的评论，负面选择评分为1的评论。

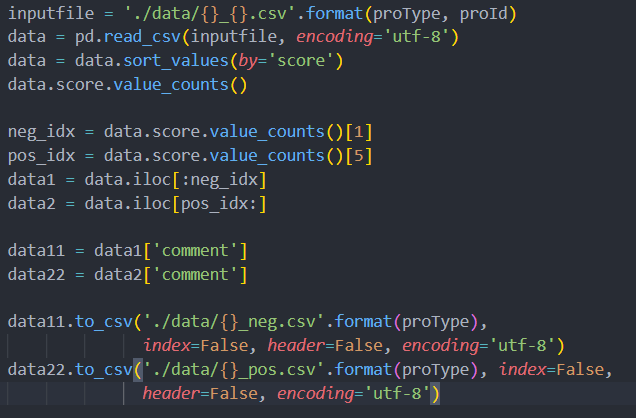


图10 选择正面评论和负面评论

构建LDA模型见图9。

### 4 数据分析

##### 4.1 整体分析

表1 整体分析

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 机型 | 价格 | 评论数 | 爬取数 | 好评率 |
| 1 | huawei p40 | 4488 | 16w | 2200 | 97% |
| 2 | huawei p40pro | 6488 | 21w | 2520 | 97% |
| 3 | huaweimate30 | 4500 | 49w | 3400 | 97% |
| 4 | huaweimate30pro | 6399 | 45w | 3400 | 97% |
| 5 | huawei honorv30pro | 3089 | 19w | 3310 | 96% |
| 6 | huaweinova7 | 2999 | 10w | 1930 | 96% |
| 7 | huaweinova7pro | 4099 | 4w | 1689 | 96% |
| 8 | huawei honor30 | 3189 | 10w | 2450 | 95% |
| 9 | oppoReno3 | 2699 | 3.7w | 1788 | 95% |
| 10 | oppoReno3pro | 3699 | 4.3w | 1930 | 95% |
| 11 | vivo NEX 3 | 4998 | 2.5w | 1742 | 95% |
| 12 | vivo Z6 | 2198 | 7w | 2060 | 95% |
| 13 | 腾讯黑鲨游戏手机 | 3699 | 6.5w | 2651 | 95% |
| 14 | galaxy s20 | 6999 | 1.7w | 1711 | 94% |
| 15 | oneplus8pro | 5999 | 6.4w | 2192 | 94% |
| 16 | oppofindx2 | 4999 | 1w | 1600 | 94% |
| 17 | vivo S6 | 2498 | 3w | 1754 | 94% |
| 18 | oneplus8 | 4599 | 9.5w | 2860 | 93% |
| 19 | vivo iQOO Neo3 | 2998 | 12w | 3220 | 93% |
| 20 | xiaomi10pro | 4999 | 13w | 3342 | 92% |
| 21 | 魅族17 | 3699 | 1.5w | 1687 | 92% |
| 22 | readmik30pro | 2999 | 23w | 3052 | 91% |
| 23 | xiaomi10 | 3999 | 32w | 3400 | 91% |
| 24 | 努比亚红魔 | 3499 | 3.2w | 2430 | 91% |

从上表可以清晰的看出各厂家的一些信息：

* 华为手机在所有价格和机型方面好评率都领先其他机型；
* oppo和vivo好评率几乎都在94%和95%，不温不火；
* 三星手机无人问津，实际上对其他机型本也打算爬取，但购买量太低；
* 小米、魅族和努比亚红魔好评率垫底，而小米在垫底各厂家之中又和其他厂家形成鲜明对比——小米的购买量遥遥领先其他垫底厂家；
* 小米在购买量上仅次于华为，然而好评率却天上地下。可见小米手机成为国内最强任重而道远，触目惊心的用户体验无疑在蚕食小米的用户市场。作为组装厂家的小米就笔者而言也是印象极差。举例来说，自去年redmi和xiaomi分离后，readmi生产了口碑尚可的redmik20系列，在新闻发布会上更是对自家产品赞不绝口，**并扬言readmi极具性价比，就价格而言是降无可降。然而在今年上半年的5g市场中，readmik30和其pro版本却在上市仅仅两三个月后官方降价700元，其降无可降之辞更是沦为笑柄——短短时间直接砍掉五分之一左右的价格更是骇人听闻。**而对其的LDA情感分析更是清晰的展现出了降价的负面影响——这款红米手机的降价影响鹤立鸡群，而保价也出现在了两个负面模型分析中。

**作为对比，整体72个负面模型之中降价出现20次，保价出现9次，而2020年小米和红米的两款主力机型的六个负面模型降价出现5次，保价出现4次。笔者也购买了一款小米手机，应该不会出现下一次的购买了。**

**降价保价如此频繁出现在自家产品，官方却未有做为，至少在智能手机领域，小米似乎已经在蚕食其口碑了，更是和小米手机出道以来主打的性价比背道而驰。**

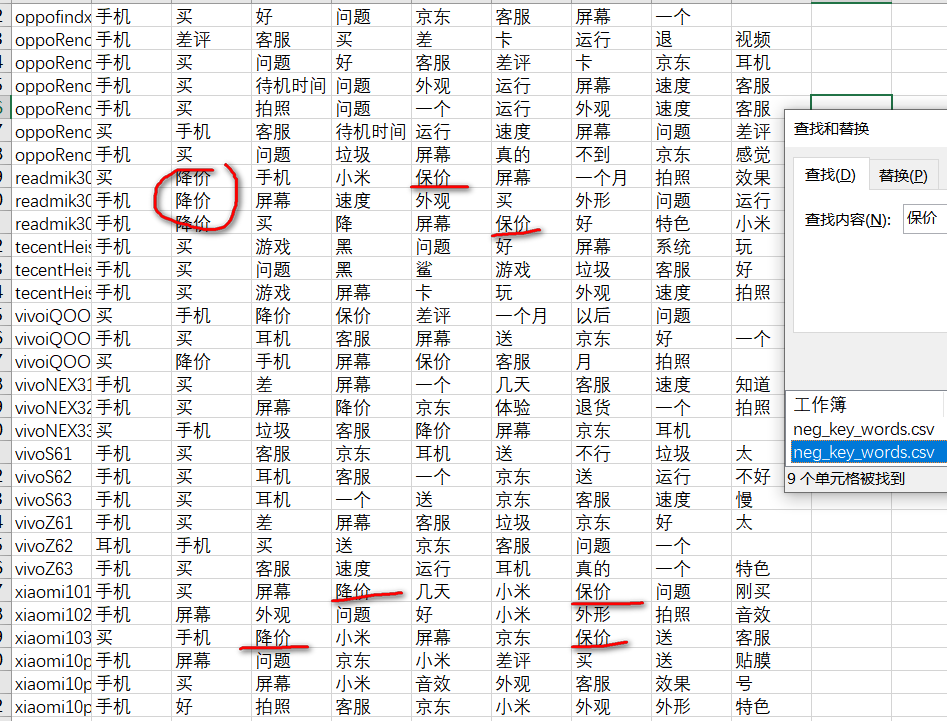


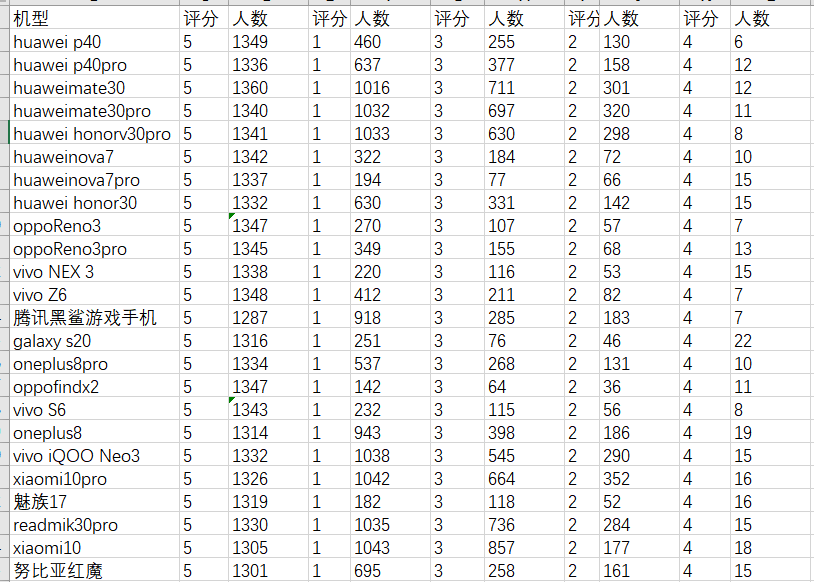
图11 小米手机问题一角

##### 4.2评分分析

在用户评分方面，所有机型的买家都倾向于最高分和最低分，然后是平均分，而2分和4分却少人问津，尤以4分明显——打分为4的用户徘徊在1%左右。可见用户在体验不是很差时都会打出最高分，这可能和商家的促销、好评返现、好评送配件等行为密切相关。

就爬取数据而言，无论是好评比例还是评论人数都有显然的注水行为。清晰可见，诸如华为小米大部分机型都有上千条以上的差评，考量实际购买力，用户体验可想而知。

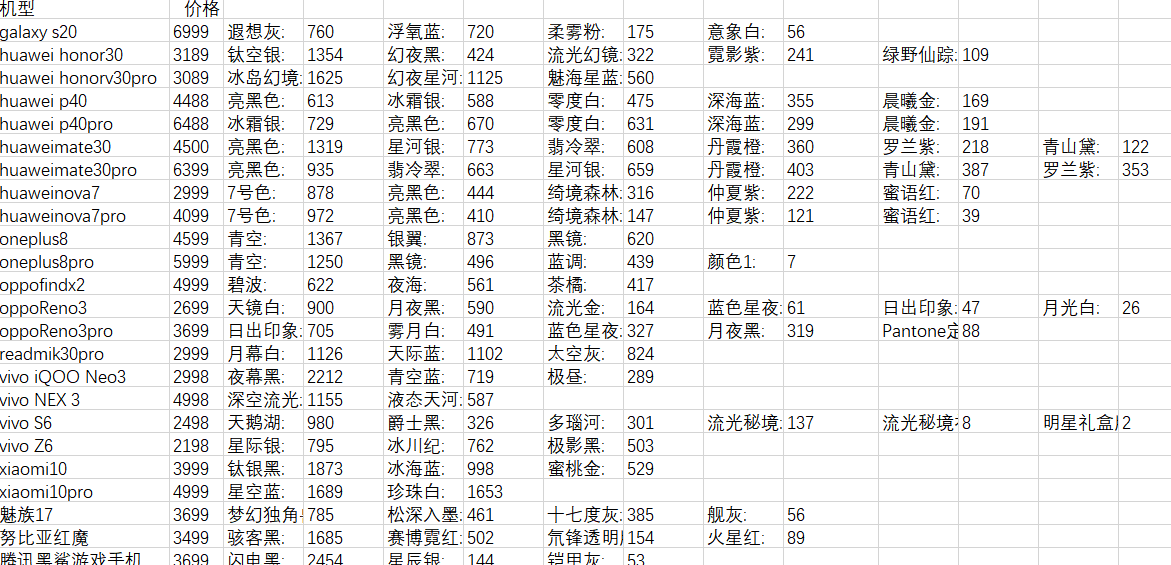
表2 用户评分统计



##### 4.3 用户偏好分析

主要从颜色方面分析验证

表3 用户颜色统计



* 对于主打游戏的黑鲨、红魔、IQOO3来说，都是以黑色系列购买者最多，黑鲨手机更是占据了接近93%的份额。
* 对于主打年轻人市场的各款手机，用户喜好偏向于白色和一些厂家自定义的亮色系。
* 具体而言，用户喜好的颜色和用户群体具有密切联系。

##### 4.4 各机型词云和词频

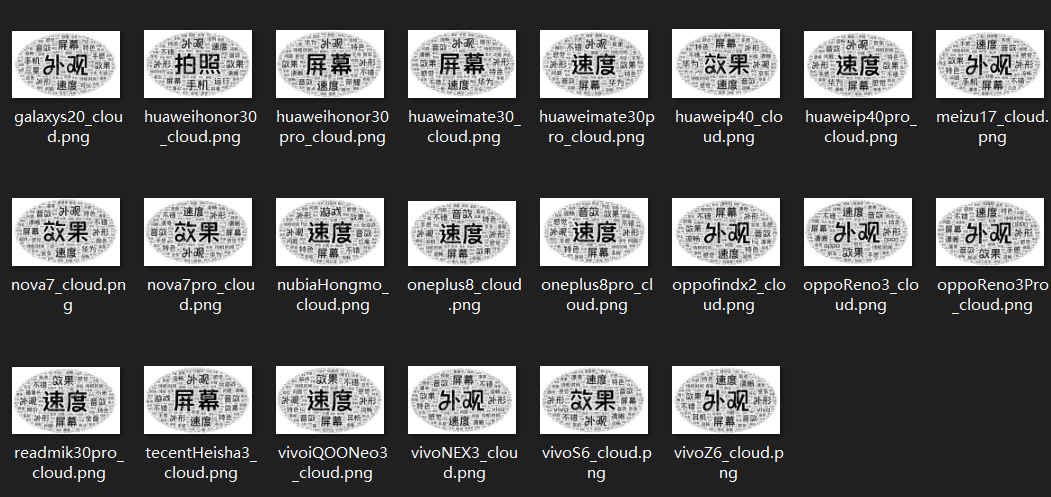


图12 词云一览

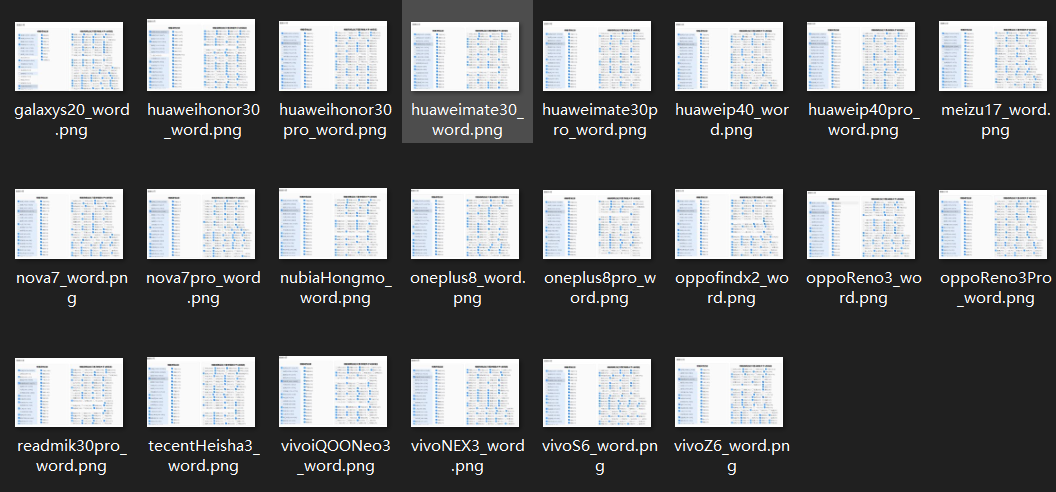


图13 词频统计一览

各机型的词频出现最多的与机型十分相关。由于时间关系难以一一分析，词云图位置为‘data/data\_fig/word\_cloud/’下，相关的词频统计位于‘data/data\_fig/word\_frequency/’文件夹下。

整体而言，“屏幕“、”速度“、”效果“、”外观“以及诸如”游戏“、”音效“等在每种机型都词频很高。

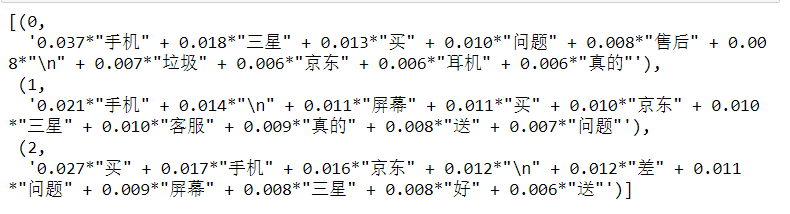
此外通过词云和词频统计也能发现一些机型整体的问题，如redk30pro这一机型的词云出现了降价、断流、保价等，则可认为该机型在这些方面存在问题，购买时需要注意



图14 redmik30pro词云

##### 4.5 基于LDA情感模型结合评分统计分析用户体验

LDA模型是一种常用的文本情感分析模型，本文简单使用该模型来分析用户体验等。

图15 lda模型示例

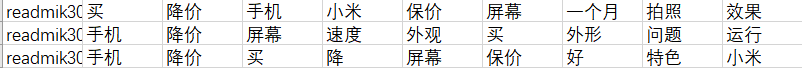


图16负面情感词提取示例

在构建lda模型后，将关键词提取为表格形式，图16即redmik30pro的负面情感词提取，从中可以清晰的看出该机型的降价、保价问题十分严重。

4.5.1差评数分析

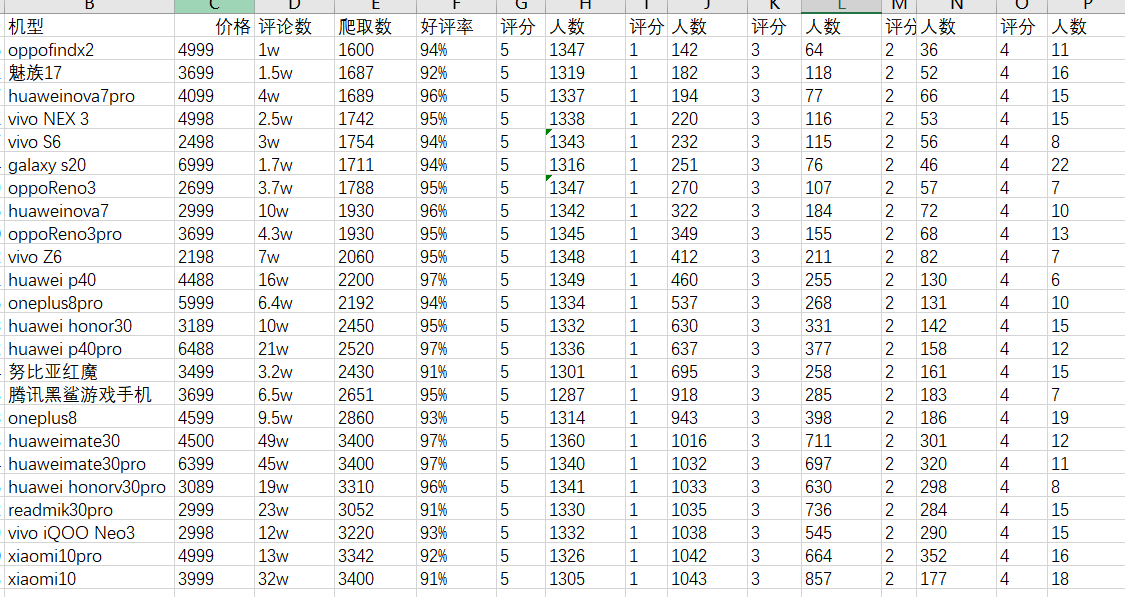


图17 差评升序排列

就上图而言，结合评论人数，笔者认为：

* 华为nova7、华为Nova7pro、华为P40、华为P40pro表现良好；
* oppo系列和vivo系列差评量不大，值得尝试；
* 对于几款主打游戏的手机而言，京东给出的总评论数据体量不大，但几款手机的差评人数却不可小视；
* 小米的两款手机评论总数体量不及华为主力机型，然而爬到最多差评，表现较差。

4.5.2 正面情感词提取分析

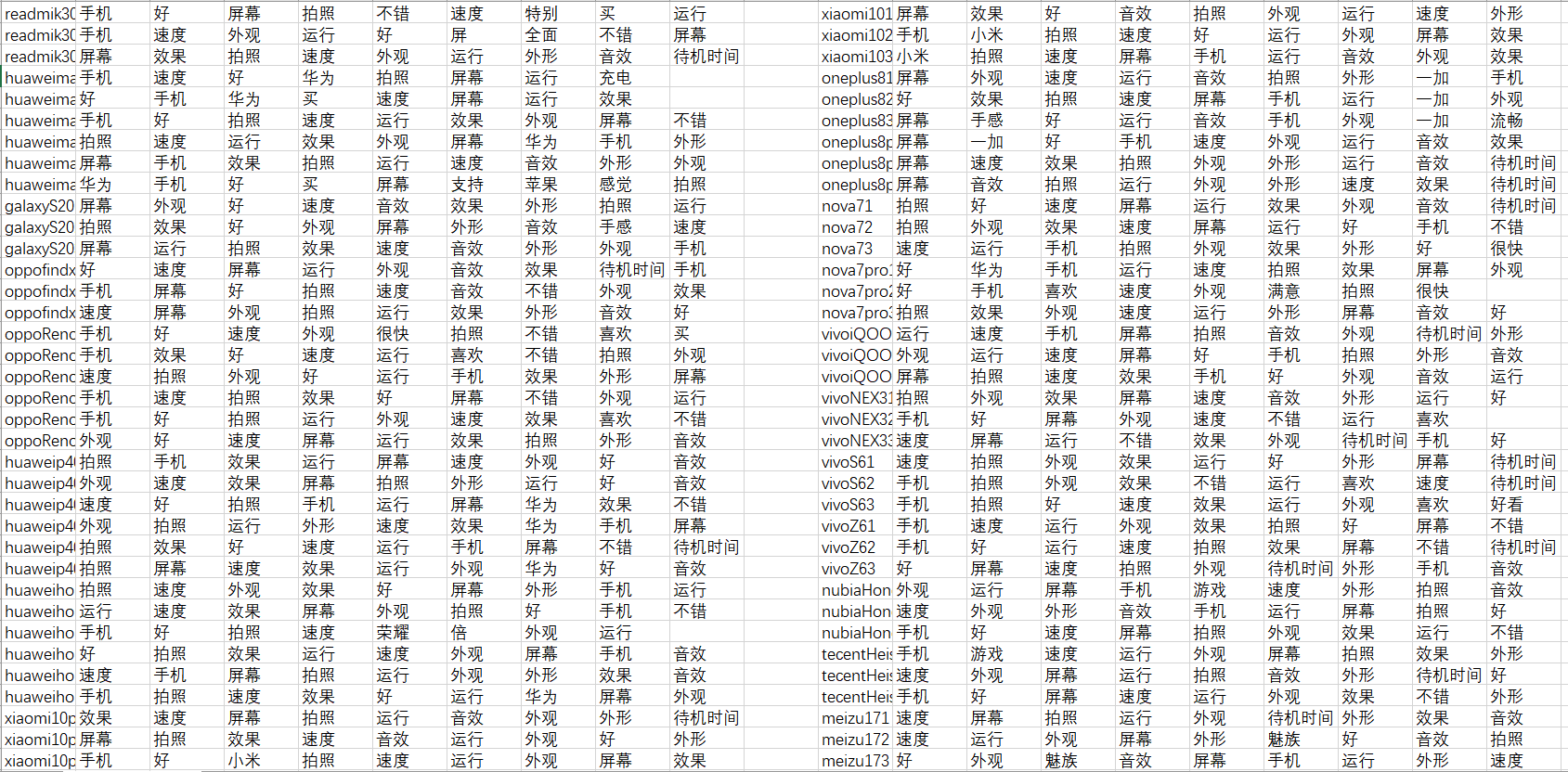


图18 正面情感词提取结果

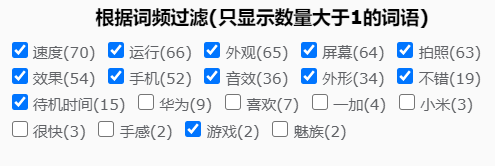


图19 正面情感词统计结果



图20 正面情感词词云

对新一代5g手机而言，lda模型提取出的正面情感词集种在速度、运行、外观、屏幕、拍照等词汇，并且几乎每种机型都出现了这些词汇，可见对于5g手机，这些方面是优势和卖点。

4.5.3 负面情感词提取分析

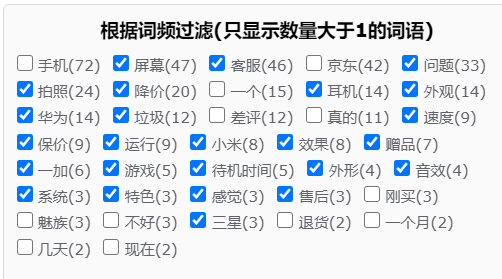
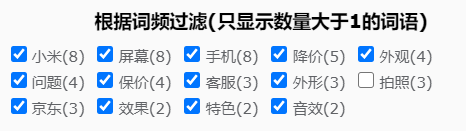
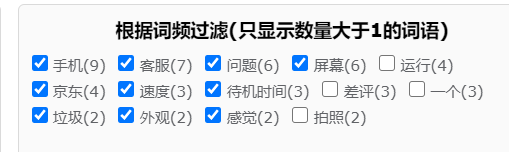
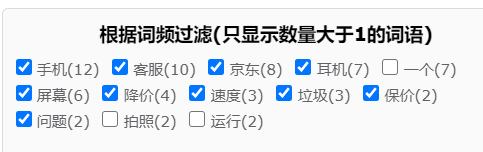
 

图21 负面情感词统计结果及词云

整体而言，屏幕问题和京东自身问题较为严重，而降价保价也出现多次，印证了新一代5g手机价格浮动的事实。

 华为 小米

oppo vivo

图22 华为、小米、oppo、vivo负面情感词统计

就华为、小米、oppo和vivo四家厂商而言，分别生成了18，9，9，12个负面情感模型。

* “客服“、”京东“多次出现，可见客服态度存在一定问题；
* Redmik30pro、xiaomi10、huaweihonor30、huaweihonorv30pro、huaweimate30、huaweimate30pro、vivoiQOONeo3、vivoNEX3等机型频频出现降价保价关键词，说明这几款手机在发行后出现了价格跳水；
* 与上一条相对应huaweip40、huaweip40pro、vivoS6、vivoZ6以及oppo系列降价保价问题较小。
* 就本文负面词频统计而言，“屏幕“是人们最关注的问题，其次为拍照、耳机、外观等。

### 参考资料

[1] https://zhuanlan.zhihu.com/p/104536974