R语言文本分析

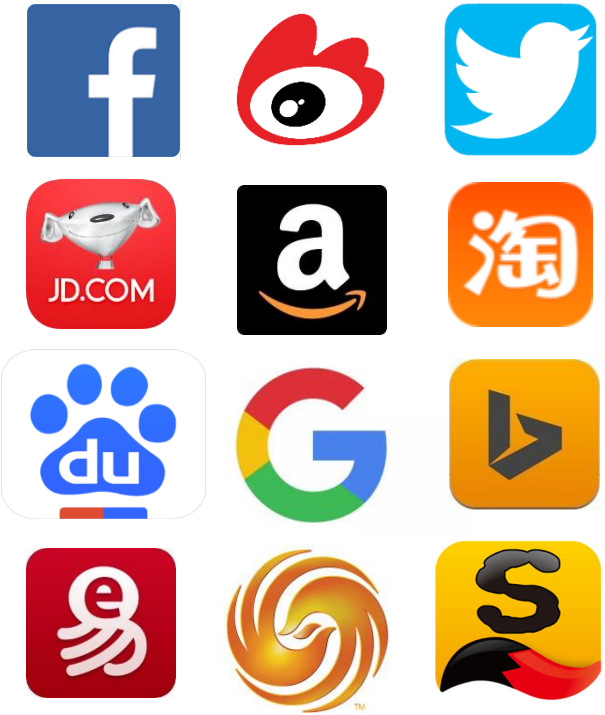
中国人民大学 统计学院
周静
2020.11.23

目录

1

文本分析的价值

无处不在的文本


JD.COM
淘
du


日参院通过TPP批准案 安倍:TPP对日本至关重要
新闻 tpp 安倍 奥巴马 | 3小时前 186
冈比亚现任总统拒绝接受败选 美谴责其违背民意
国际新闻 冈比亚 总统 败选 | 3小时前 1623
裸贷女生谈＂被坑＂经历:还清欠款仍被要求＂肉偿＂
新闻 借贷宝 趣分期 分期平台 | 3小时前
江苏一街道干部挪用公款上亿元 被检察院查处
新闻 挪用公款 检察院 贪污 | 3小时前 661

文本分析在生产中

- 生产厂家——改善产品
  - 从用户评论中挖掘消费者对产品的关注点

外观？

屏幕？

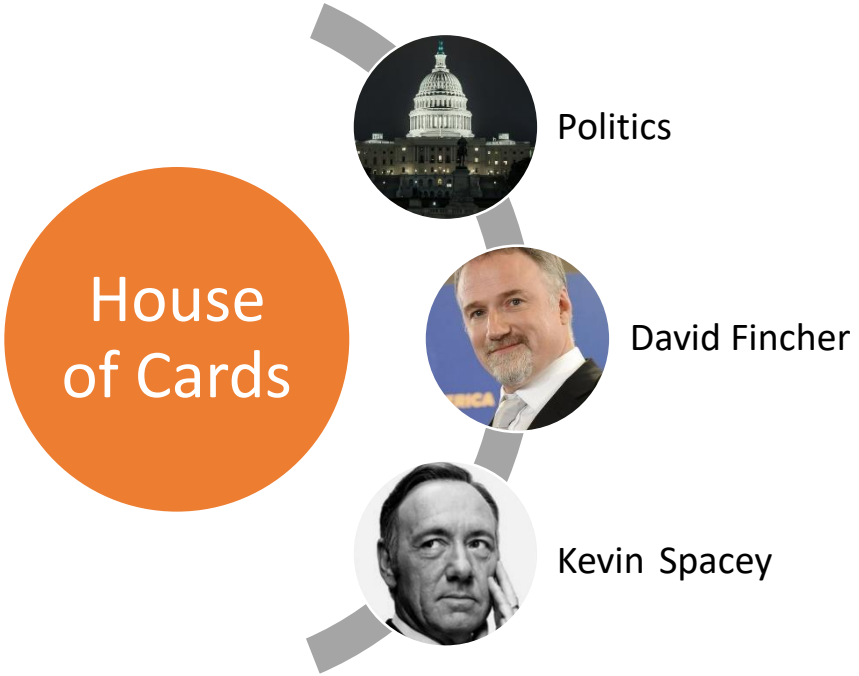
拍照？

系统？

电池？

文本分析在生产中

- 生产厂家——创造产品
  - Netflix → 纸牌屋
  - 根据2900万名订阅用户的收看习惯和偏好



文本分析在金融中

- 金融行业
  - 从上市公司的公告、年报、新闻中探究公司的发展状况
  - 从网民在股吧、论坛上的动态来判断大众对股票的评价和喜好程度

沪指涨0.07%央企改革活跃 目标位

尽管股指短期企稳，但整体做多氛围并未恢复。[千股
要问股][资金动向][LIVE 沪指见顶了吗]

- [大盘]两市低开低走探底回升 市场开启震荡修复
- [分析]多空将迎来大较量 抄底绝佳机会 哪些板块
- [主力]主力洗盘曝一大阴谋 探底翘尾激活反弹 抄
- [研究]机构称调整行情将持续 公募:短期不宜抄底
- [推荐]任泽平:超跌带来买入机会 短期以调整行情
- [博客]反弹还需等待 止跌信号还未出现 机会正在
- [理财师]大赛:逃顶高手牛股三连板 招募令 爆抢
- [股吧]市场最终企稳还需时日 大盘开启震荡修复
- [名家]wu2198：大盘反弹的生命线 抄底的机会在

保监会紧急会议 项俊波:险资勿做短

网友讨论 | 理财师分析 | 相关博客 | 公司新闻 | 研究报告

点击	回复	标题
1180	0	万 科A价值评估及操作指南(2016年12月13日)[置顶][精华]
1482	0	[资金流向]万 科A主力资金连续净流出2天[置顶][精华]
12616	0	周三走势预测[置顶]
20918	0	止跌后【特大长牛超级大牛市将提前80天启动】[置顶]
21670	0	纳米发电机的应用前景广泛，3股受益（附股）[置顶]
40126	0	万科A：短线出现踩踏事件[置顶]
5140	0	人民日报专访万科总裁：楼市不存在崩盘风险[置顶]
26	0	哈哈……三个跌停板之后再慢慢100股100股的买也不为迟！！

文本分析在生活中

- 文本分析帮助我们从纷繁复杂的内容中找到重点

好评度

96%

屏幕大(1671) 系统流畅(1601) 外观漂亮(1565) 反应快(1239) 功能齐全(1202)

照相不错(1081) 通话质量好(993) 信号稳定(903)

推荐菜 环境 价目表 官方相册 品牌故事

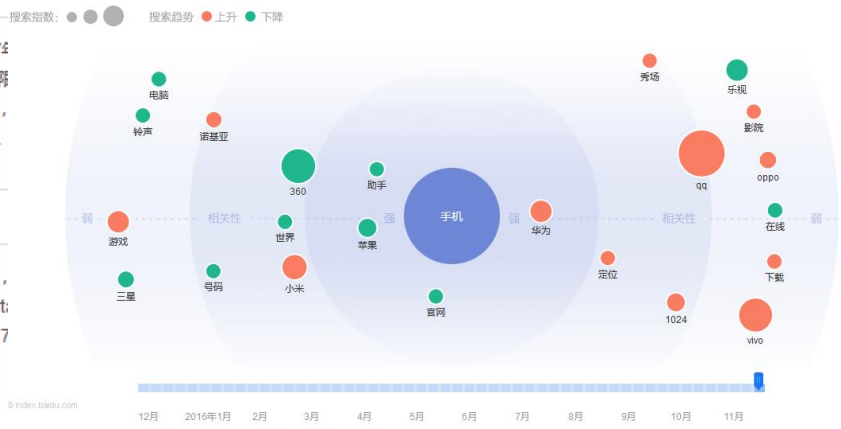
香辣烤鱼 (95) 酱香烤鱼 (63) 豆豉烤鱼 (39) 羊肉串 (33) 香辣牛蛙 (21) 凉面 (13) 口水鸡 (13)

烤巴沙鱼 (12) 果蔬大拌菜 (11) 鸡汁杏鲍菇味烤鱼 (11) 孜然藤椒味 (9) 伤心凉粉 (8) 鱼香味烤鱼 (8)

文本分析在生活中

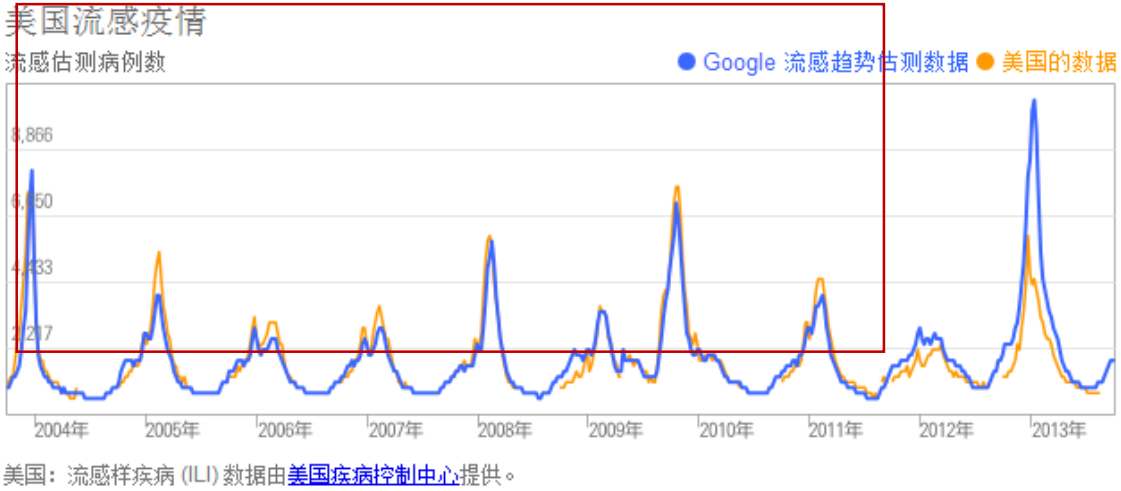
- 各种好玩的榜单

Baidu 指数



文本分析在生活中

- Google——流感预测
  - 监测“温度计”“肌肉疼痛”等一系列和流感相关的关键词在网上的搜索量来追踪分析不同地区的流感趋势，比传统方法快两周



文本分析都包括什么呢?

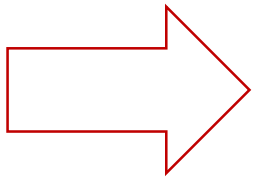
文本分析是什么？

Step 1：从非结构化数据到结构化数据

Step 2：从结构化数据中归纳有意义的指标，用于指导实际生产和生活

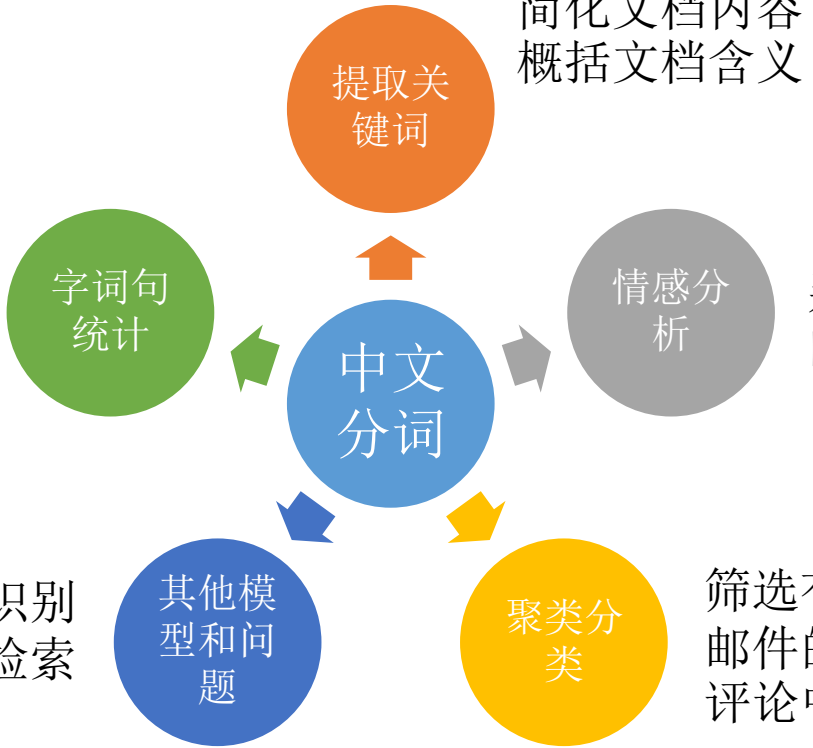
有风	轻送	柳枝	微拂	……	味道
1	1	1	1		1

有风轻送，柳枝微拂，树下的男子一身白衣纤尘不染。携着一把古琴，衣袂飘飘，很有几分潇洒出尘的味道。



字数：48
词数：22
句子数：6
关键词有哪些：出尘，携着、微拂
其他指标……

文本分析包括什么?


中文分词
提取关键词
简化文档内容
概括文档含义
情感分析
判断用户对产品的喜好程度
聚类分类
筛选有用信息（垃圾邮件的识别、在用户评论中过滤广告）
其他模型和问题
命名实体识别
信息检索
字词句统计

2

用R语言进行文本分析

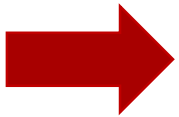
概述

- 使用工具
  - R packages：jiebaR，wordcloud
- 展示数据集
  - 京东自营iPhone 7 & iPhone 7 plus的用户评论数据
- 完成任务

1. 简单的字句统计

- 统计字符数
  - nchar函数
- 统计句子数
  - strsplit函数

东西不错，很好用

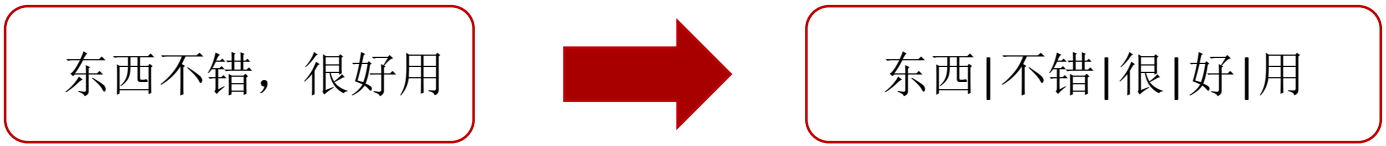


字符数=8
句子数=2

- 更多字符处理函数
- paste()，grep()，gsub()，substr()

2. 中文分词

- R语言中常用的分词包
  - Rwordseg（引用了Ansj包，Ansj是一个开源的java中文分词工具，基于中科院的ictclas中文分词算法，是学术界著名的分词包之一）
  - **Jieba**（基于python写成的一个工业界的分词开源库，具有很好的扩展性）
- 用jieba包进行分词
  - 安装jieba package
  - 调用worker进行分词



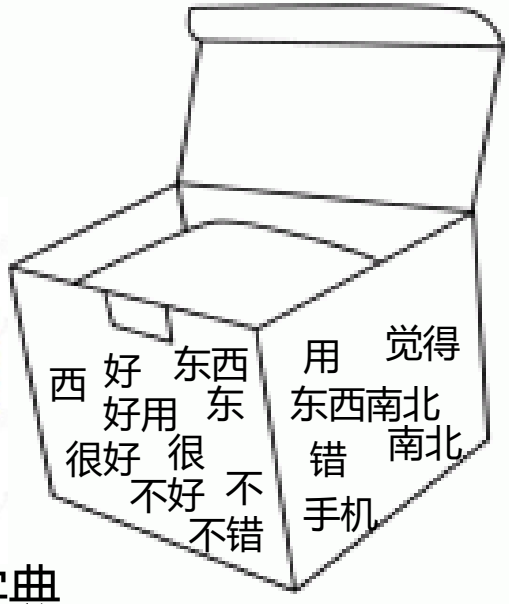
3. 优化字典：去掉停用词

- 停用词大致可以分为两类：
  - 语言中使用的功能词（语气助词、副词、介词、连接词），这些功能词通常极其普遍，并没有什么实际含义，比如：“你、我、他、了、的”
  - 有特定含义，但是应用十分广泛的词，比如“想、做、来、去”等
- 对停用词的总结
  - 很多机构都给出了自己的停用词表，如“百度”，“哈工大”，“四川大学机器自然实验室”等
  - 汇总结果： stopwords.dat
  - 更多参见：https://github.com/dongxiexidian/Chinese
- 如何选择停用词以及是否使用停用词没有统一标准

3. 优化字典：自定义字典

- 绝大多数分词方法都是基于“字典”

东西|不错|很|好|用



- 通过添加用户自定义字典，丰富原始字典

3. 优化字典：选取自定义字典

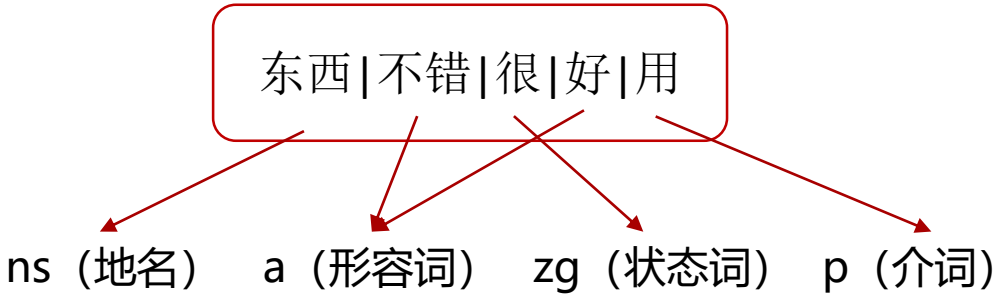
- 从原始分词结果中总结
  - 费时费力，但效果最好
- 有自己领域的行业词汇
- 添加方法
  - copy-paste到默认的用户自定义词典中
  - add_user_words添加少部分词
  - 加载新的用户自定义词典

3. 优化字典：使用搜狗词库

- 下载词库
  - 生活—日常—手机词汇大全
- 将词库转为txt
  - R包：cidian

4. 词性标注

- Jieba分词可以给出每个分词结果的词性
  - Jieba词性标注表.txt



- 对分词结果按词性进行筛选
  - 假设我们只关注“名词、动词、形容词、副词” （tags.txt）

5. 关键词提取

- 提取关键词的方法
  - 按照在全部文本中出现的词频大小
  - TF-IDF（词频-逆向文件频率）
    - TF(Term Frequency)指的是某一个给定的词语在该文件中出现的次数
    - IDF (inverse document frequency) 是一个词语普遍重要性的度量。某一特定词语的IDF，可以由总文件数目除以包含该词语的文件的数目，再将得到的商取对数得到。

5. 关键词提取

[1] 还|不错|吧
[2] 东西|不错|很|好|用
[3] 用|着|还|不错
[4] 物流|很|给力
[5] 质量|不错
[6] 东西|很|好|售后|也|非常|给力|哦

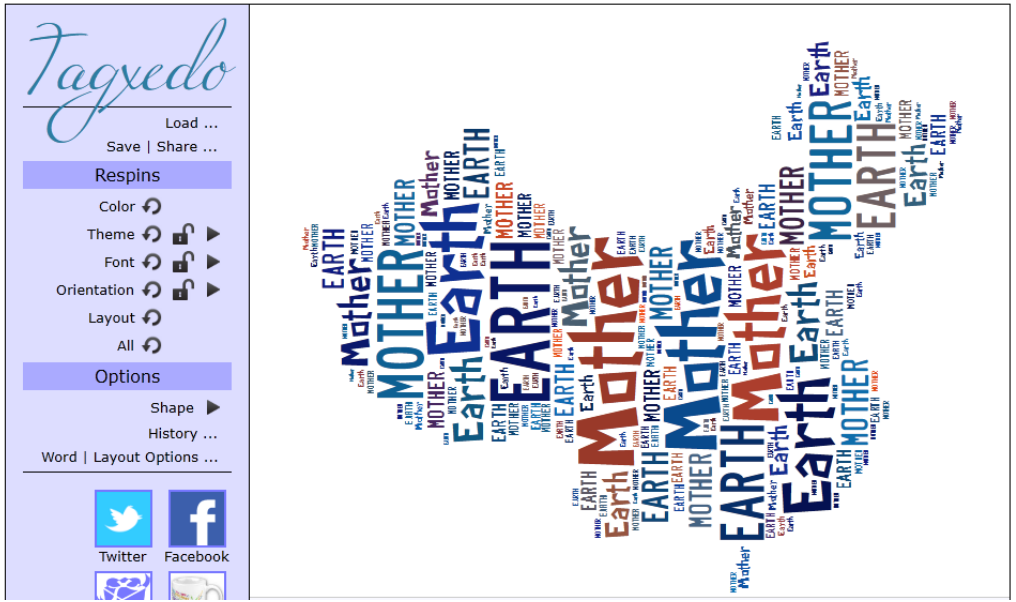
- 6个文件中出现的总词数：3+5+4+3+2+8=25
- “不错”出现的总次数=4
- 词频TF=4/25=0.16
- “不错”出现的文件个数=4
- 逆文件频率IDF=log(6/4)=0.405
- TF-IDF=TF*IDF=0.0648

6. 绘制词云

- 使用工具
  - R包：RColorBrewer, wordcloud
- 使用数据
  - 出现次数最高的前100个词及其频数

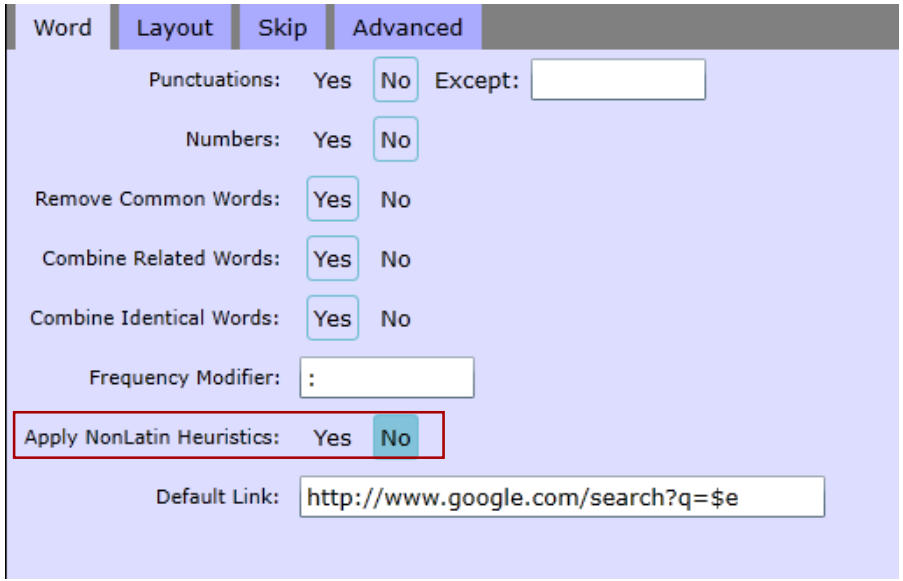
6. 绘制更加优美的词云

- Tagxedo：一个专门绘制词云图的网站
  - http://www.tagxedo.com/app.html
- 常用选项
  - Load:输入要画词云的数据
  - Save|share：保存词云图
  - Color: 随机变化颜色
  - Theme：变化主题色系
  - Font：变化字体
  - Shape：选择词云形状



6. 绘制更加优美的词云

- 中文使用时的特殊处理
  - Word | Layout Options – Word- Apply NoLatin Heuristics - No



7. 衍生指标

- 定义问题
  - 如何从文本内容中生成衍生指标与业务问题密切相关

哪些因素影响消费者对iPhone7的评价？

因变量：用户评分

自变量：用户关注点在哪？

如何衡量各个关注点？

7. 从评论中寻找用户关注点

- 从Top50的词中选出前10个描述手机特点的、中性的的词汇

价格 正品 物流 快递 包装 速度 电流 屏幕 送货 服务

- 衍生指标1：每条评论中是否出现该关注点
- 衍生指标2：每条评论中出现所有关注点的次数

3

挖掘用户评论：
从用户评论看手机销量

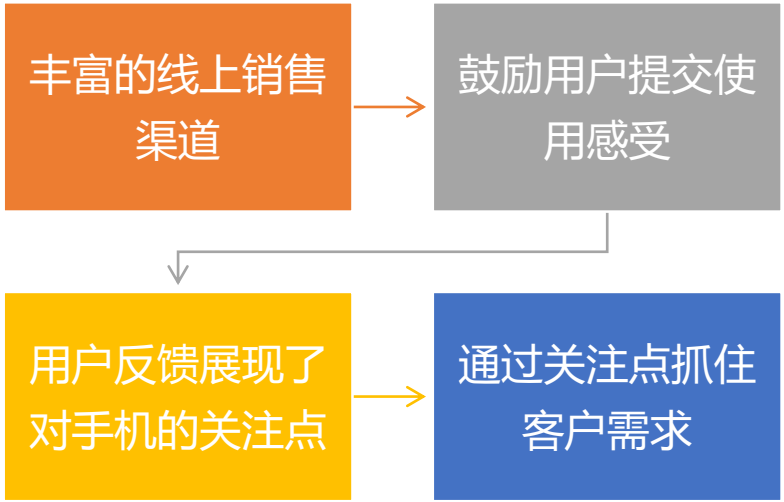
手机行业发展现状

- 手机市场容量大、发展空间广阔
  - 2016年前三个季度，我国手机市场出货量达到4.00亿部，同比增长7.7%。（工信部）
- 市场竞争日益白热化
  - 9月份在市场上销售的机型已达到1600+款
  - 传统手机品牌推陈出新，非传统手机品牌（如格力、360、乐视）觊觎，纷纷入驻
  - 相同价格区间内，同质化竞争严重

手机行业发展现状

- 差异化需求迫切

如何抓住用户需求?



用户评论举例

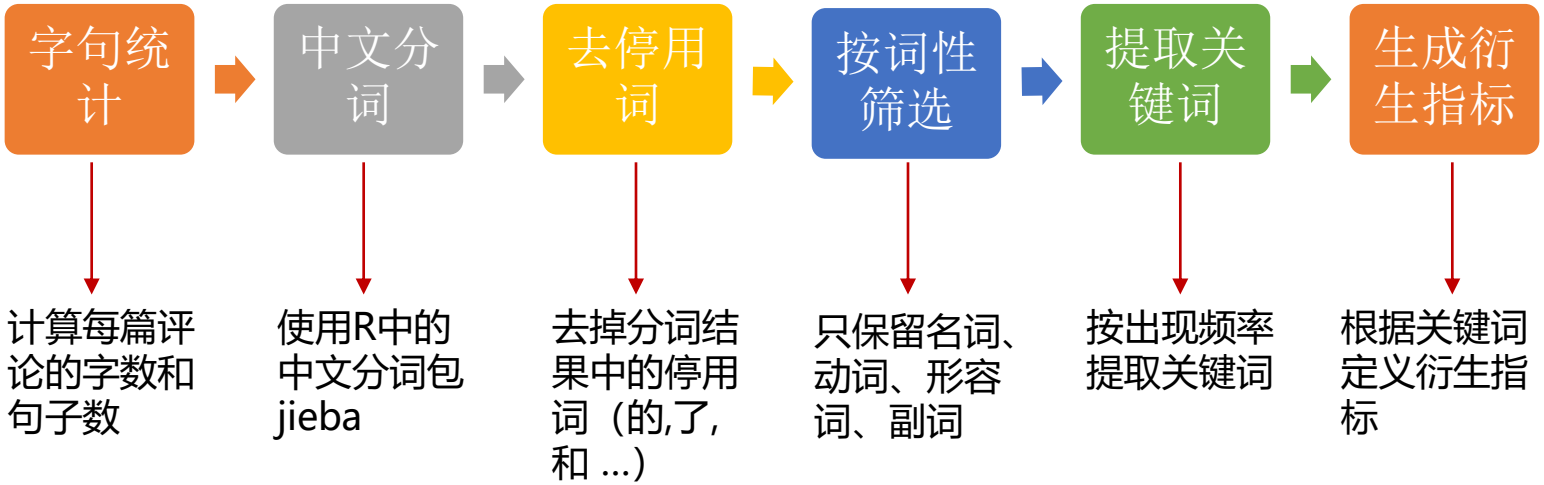
手机已经用了一段时间了，很漂亮，拍照的确不错，物流也给力
偶尔信号消失，没网没信号。电池还是比较耐用的，充电速度也快。
屏幕大小适中，反应比较快，不比国外牌子差，很好用
系统可以深度定制。电池也可以
声音很大，老爸很喜欢
手机不错，好用，手感好，声音也不小

数据概况

- 截止2016年11月31日，某知名电商在其自营平台上销售的手机数据（297部）及能爬到的全部用户评论数据（216754条）。

手机数据	用户评论数据
• 手机的各种参数指标（价格、品牌、屏幕尺寸、前后摄像头像素、SIM卡数量、指纹识别、GPS） • 销售平台有无促销 • 评论总数、好评数、差评数	• 涉及每条评论的评分、购买时间以及具体内容

评论内容处理流程


字句统计
中文分词
去停用词
按词性筛选
提取关键词
生成衍生指标
计算每篇评论的字数和句子数
使用R中的中文分词包jieba
去掉分词结果中的停用词（的,了,和 ...）
只保留名词、动词、形容词、副词
按出现频率提取关键词
根据关键词定义衍生指标

提取关键词

- 手机好在哪？差在哪？
  - 分别在好评（rating>3）和差评（rating<3）中提取出现次数最高的前100个词绘制词云图

提取关键词

- 只看好评！
  - 手机的平均好评率为97.68%
- 好手机的一篮子指标
  - 在所有好评中出现次数最高的前50个词里找出描述手机特点的、中性的所有词

排名	好评词	总频数
1	屏幕	6218
2	电池	6006
3	系统	5743
4	性价比	5509
5	流畅	5104
6	外观	4966
7	手感	4575
8	价格	4476
9	拍照	3600
10	充电	2704
11	声音	2619
12	耐用	2594

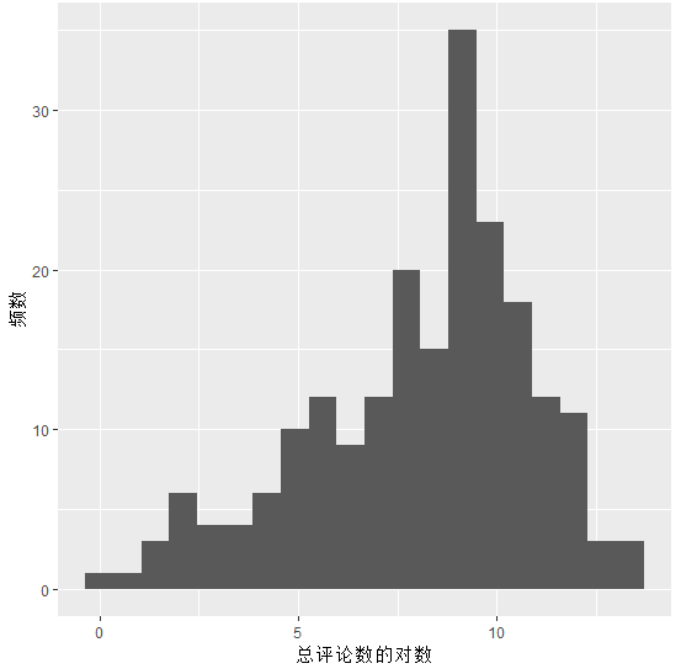
生成衍生指标（以手机为单位）

第 i 部手机

- 该手机所有评论的平均字数和句子数
- 该手机中各个**好评词的指数**（以第k=1个好评词“屏幕”为例）
  1. “屏幕”在该手机所有好评中出现的次数： $n_{i,k}$=16
  2. 所有好评词在该手机中出现的总次数： $\sum_j n_{i,j}$=128
  3. “屏幕”的好评指数： $\frac{n_{i,k}}{\sum_j n_{i,j}}$=16/128=0.125
- 某个词的好评指数越高，说明该手机在这方面表现越突出，做的越好

因变量：总评论数

- 我们将每款手机的总评论数作为手机销量的替代变量。



- 左图为总评论数对数的直方图
- 总评论数最小值：1 （ ZUK Z2 Pro ）
- 总评论数平均值：32860
- 总评论数最大值： 625200 （ 小米 红米 3S ）

回归模型中使用的变量

变量类别		变量名称	说明
因变量		log(总评论数)	
自变量（来自手机）	手机特征	价格	
		品牌	总评论数最多的8大品牌+其他
		屏幕尺寸	
		双卡机类型	单卡、双卡单待、双卡双待、其他
		SIM卡数量	单卡、双卡、其他
		前置摄像头	低（<500）、中（500~1000）、高（>1000）
		后置摄像头	低（<1000）、中（1000~1500）、高（>1500）
		指纹识别	不支持、不支持、其他
		GPS	不支持、不支持、其他
	促销信息	促销信息	有、无
	评论信息	好评率	
自变量（来自评论）	字句统计	平均句子数	为避免多重共线性，平均字数和平均句子数只放其一
	好评指数	12个好评词	为避免多重共线性，去掉“耐用”

回归结果（用AIC进行变量选择）

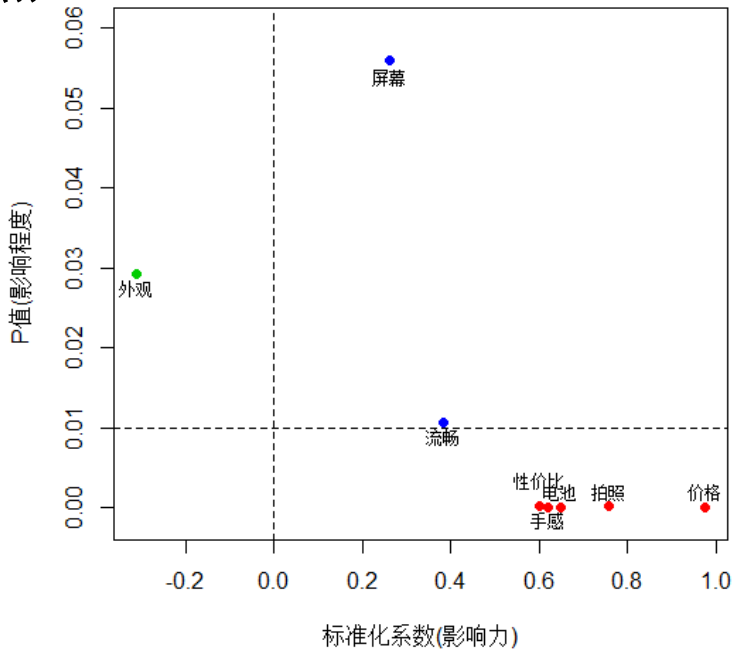
变量	标准化系数	P值	显著性	变量	标准化系数	P值	显著性
华为	2.103	<0.001	***	指纹识别设置=其他	-0.981	0.103	
小米	3.363	<0.001	***	指纹识别设置=支持	0.192	0.655	
乐视	2.630	0.006	**	GPS设置=其他	0.904	0.630	
360	2.247	0.003	**	GPS设置=支持	-0.437	0.803	
三星	2.457	<0.001	***	平均句子数	0.433	0.002	**
OPPO	3.921	<0.001	***	屏幕	0.262	0.056	.
VIVO	3.356	<0.001	***	电池	0.650	<0.001	***
nubia	4.693	<0.001	***	性价比	0.600	<0.001	***
SIM卡数量=其他	-2.153	0.018	*	流畅	0.382	0.011	*
SIM卡数量=双卡	0.710	0.115		外观	-0.310	0.029	*
后置摄像头=高	1.157	0.043	*	手感	0.622	<0.001	***
后置摄像头=中	0.816	0.131		价格	0.976	<0.001	***
前置摄像头=高	-1.852	0.006	**	拍照	0.758	<0.001	***
前置摄像头=中	-0.722	0.083	.	F检验：P值<0.001，调整后R方：63.15%			

模型解读——手机特征

- 品牌效应显著
  - 从品牌上来看，8大品牌，华为、小米、乐视、360、三星、OPPO、VIVO、Nubia的总评论数都显著高于“其他”品牌
- 手机配置并非越高越好
  - 从SIM卡数量来看，单卡和双卡并没有显著差别
  - 手机的后置摄像头像素水平越高，手机的总评论数越多
  - 相反，当手机的前置摄像头处于较高水平时，反而降低手机的总评论数
  - 指纹识别功能的开通并没有显著影响手机的总评论数
  - GPS功能的开通也不显著，但这主要源于数据中“不支持GPS”的样本太少

模型解读——好评指数

- 在12个好评词中，真正影响手机总评论数的指标为：屏幕、电池、性价比、流畅、外观、手感、价格、拍照
  - 左图展示了8个好评词的“影响力”和“影响程度”。
  - 外观的影响力是负的，这说明该自营平台上的消费者普遍比较理性，过分强调手机外观上的特点会削弱其他方面的优势，从而对消费者产生消极影响
  - 价格的影响力和影响程度都很大
  - 拍照、电池、手感、性价比几个特点也有着较大的影响力和和影响程度
  - 手机的流畅性影响力和影响程度一般
  - 屏幕的影响力和影响程度最弱



结论建议

1	手机市场上品牌效应凸显，大品牌应注重维护自身的品牌价值，小品牌应努力标新立异、突出重围。
2	手机并非配置越高，卖的越好。手机厂商应努力开发消费者真正关心的功能。
3	“价格” 依然是消费者关注的重中之重。此外，手机厂商在产品设计以及营销策略上还可以重点突出产品的“性价比”以及“拍照” “续航” “手感”和“流畅度”，从而引起消费者共鸣。

谢谢大家！