

# 观点口碑分析

谢恬

厦门大学，数学科学学院

2017 年 4 月 28 日



# 背景介绍

- 同质商品的网购基于已有消费者评论
- 已有文献研究
  - 基于关联规则挖掘评价词与评价对象
  - LDA主题模型在短文本上效果不佳

# 背景介绍

- 天猫, 京东等电商平台
- 部分商品已提取标签
- 标签信息无分级, 量化
- 无开源数据集

The screenshot shows a Taobao product page for a USB drive. The page includes a header with '商品详情' (Product Details) and '累计评价 157568' (Total Reviews 157568). Below the header, there are several review snippets and a list of extracted tags. Red boxes and arrows highlight specific elements for data extraction:

- 商品提取标签 (Product Extraction Tags):** A red box highlights a set of tags: '快递不错(2491)', '质量很好(2094)', '是正品(790)', '态度不错(787)', '性价比高(43)', '快递服务好(10)', and '质量一般(119)'. An arrow points from the label to this box.
- 具体对应文本 (Specific Corresponding Text):** A red box highlights the text '性价比很高' in a review snippet. An arrow points from the label to this box.
- 具体对应文本 (Specific Corresponding Text):** A red box highlights the text '性价比很高。好评' in another review snippet. An arrow points from the label to this box.
- 具体对应文本 (Specific Corresponding Text):** A red box highlights the text '买过性价比非常高, 而' in a third review snippet. An arrow points from the label to this box.

The review snippets shown are:

- U盘质量很好, 是正品, 美观大方, 性价比很高, 值得拥有。全五星好评! 服务态度很好, 我很满意。 03.26
- 很好性价比很高 04.16
- 物流超快, 宝贝读卡速度快。 性价比很高。好评 04.11
- 非常好用, 抢购了好几家, 后来选择了这一家, 朋友1层的, 且价格也不是很贵, 非常喜欢, 给商家100赞。 03.19

# 团队组建

- 谢恬：数学科学学院大二生，曾参与财报文本情感分析项目
- 周韵丰：管理学院大三生，参与数据挖掘项目，了解数据库知识
- 岳忠信：经济学院大一生，精通爬虫，为研究生等开设爬虫课程
- 李裕洋：经济学院大一生，熟练使用pandas，熟悉数据清洗流程



# 设计思路



- 数据获取
- 数据预处理
- 数据库构建
- 情感分数计算
- 结果可视化

# 设计思路

- 数据获取
  - 记录标签信息
  - 爬取已提取标签与未提取标签的评论

```
id: 302597086045
pics: ["/img.alicdn.com/bao/uploaded/i1/1421005018761582664/TB2m3DGhh1mpuFjSZPfXXc9iXXa_!!0-r
picsSmall: ""
position: "220-11-11,20;420-11-21,34;420-11-0,5;"
rateContent: "物流还行吧,不快不慢。包装的倒是挺严实的,就是运送途中再小心点就好了,这明显是被摔过的。"
rateDate: "2017-03-01 22:08:22"
reply: ""
sellerId: 1855527029
serviceRateContent: ""
structuredRateList: []
```

220, 420等编号对应物流不错,包装严实

# 设计思路

- 数据预处理

- 去除无效及冗余, 自动回复等评论

初次评价:  
2016.12.12

此用户没有填写评论!

- 过滤停词

# 设计思路

## ➤ 数据库构建

- 依据标签, word2vec聚类, 人为判断将评论分割, 存入不同维度

价格	物流	质量	外观
性价比很高	物流很快	东西很不错	无论是外观还是质感都是一样好
实惠漂亮性价比不错	物流服务很贴心	质量还不错	最主要是小巧玲珑很可爱很好看
比实体店便宜	物流服务态度及速度都不错	东西很棒啦	东西很小巧可爱



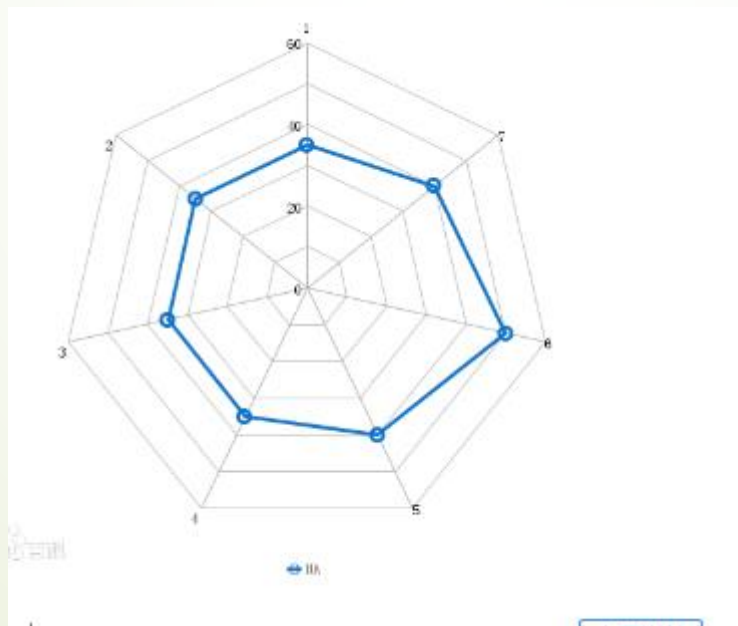
# 设计思路

- 情感分数计算
  - 导入程度副词词典
  - 将评论分割，存入不同维度
  - 对相应评论简化处理
  - 字典与评论数据库匹配

价格		
分数	评判标准	示例
1	“很”等强烈程度副词	比xx店贵很多
2	挺贵、略贵等轻微不满	挺贵的
3	不贵，还行，尚可等词	也不算贵
4	挺等程度副词	挺划算的xxx
5	超、超级、最很，非常等	赶上活动超级划算

# 设计思路

- 结果可视化
- 以雷达图方式呈现





## 新颖之处

- 口碑多维度分析，切实满足顾客购物需求
- 在数据预处理部分简化评论提高匹配速度
- 量化评论情感倾向，不仅仅是正负面判断




# 发展前景



- 优化细节(如标签提取)，减轻流程人力成本
- 构建各类产品的口碑量化模型
- 建立基于口碑的网站/小程序，为消费者决策提供更可靠的商品质量评估

# 经费使用计划

资金	用途
2000元	订阅相关文献资料及材料打印；
2000元	购买相关优化数据获取的工具，获取数据及相关参考源码等；
1000元	参与相关学科竞赛或学术会议；
3000元	相关学术论文发表；
3000元	项目出成果后请相关开发人员协助开发口碑分析小程序或网页等；



总结

谢谢大家！

