

大规模商品挖掘计算

何小锋

京东商城/基础架构部

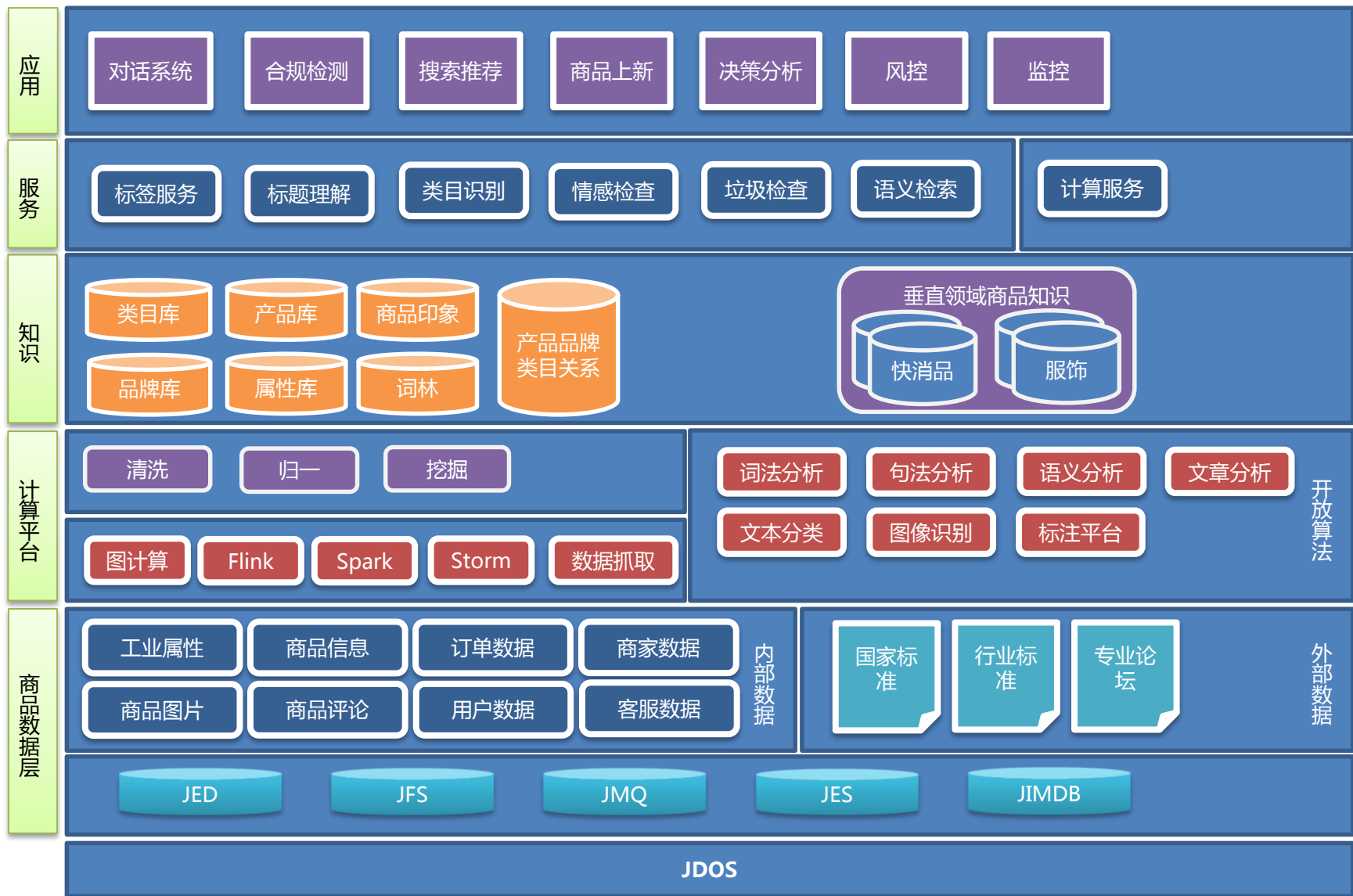
TEL : 13910526009

EMAIL : hexiaofeng@jd.com



JD.京东
.COM





现状：

- 商品数据录入难以有效监管，数据质量参差不齐
- 用户反馈数据没有得到有效使用



目标：

- 对于商家录入的商品数据进行清洗，提升数据准确率
- 对于原先没有得到有效利用的数据，整合抽取
- 为商家生态提供算法支持，从源头把控商品数据质量

商品标题堆砌、展示不全



iPhone7/7plus手机壳/保护套 苹果7plus
超薄全包硅胶透明电镀软壳 5.5英寸 炫亮黑☆炫亮电镀



运动鞋男鞋夏季男士网面跑步鞋女休闲板鞋
透气韩版潮鞋情侣旅游大码网布鞋子 男 黑色 40

iPhone7/7plus手机壳/保护套 苹果7plus 超薄全包硅胶透明电镀软壳 5.5英寸 炫亮黑☆炫亮电镀

分词

iPhone7/7plus 手机壳 / 保护套 苹果7plus 超薄 全包 硅胶 透明 电镀 软壳 5.5英寸 炫亮黑☆炫亮 电镀

命名实体识别

产品词

产品词

产品词

产品词

风格

风格

材质

风格

工艺

款式

尺寸

颜色

工艺

短文本理解：依存分析

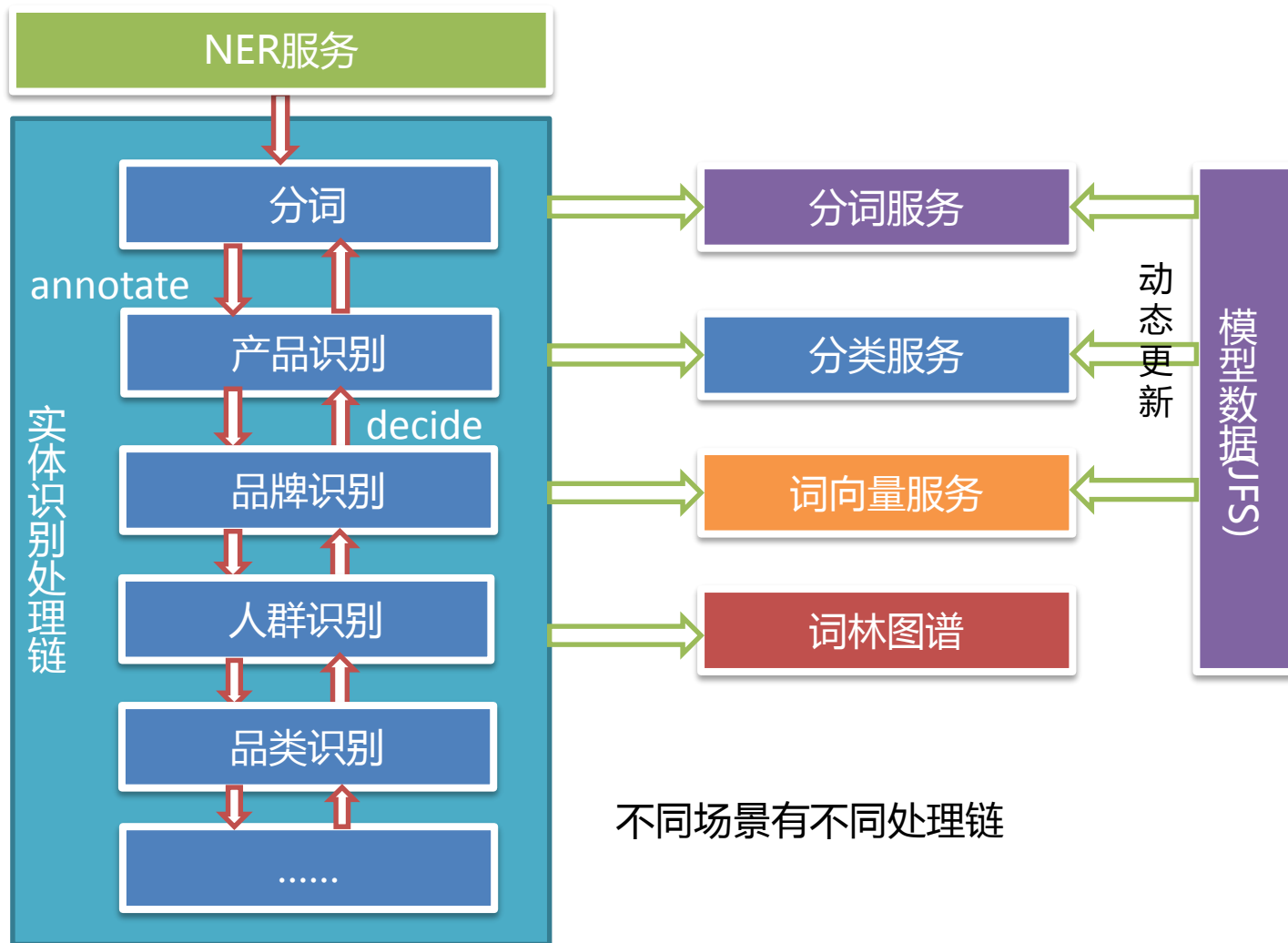
Head：手机壳

pure modifier：超薄、透明、硅胶、炫亮黑

constraint：iphone7

应用：标题重组

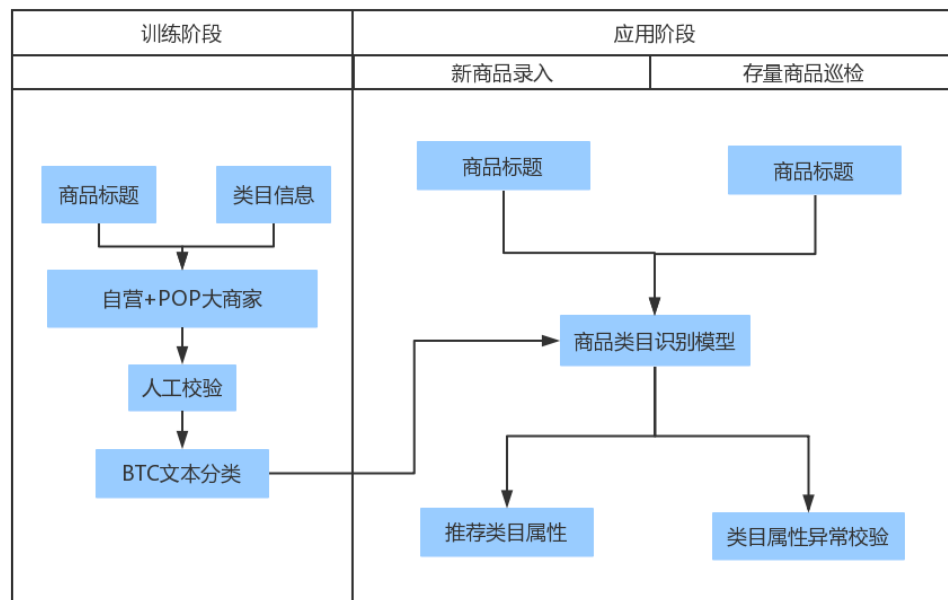
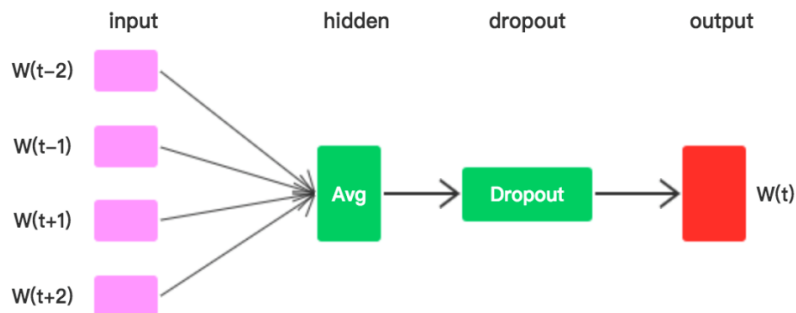
Iphone7 超薄电镀硅胶手机壳 炫亮黑





类目错绑严重

- 商品录入量大，难以管控：大型店铺sku数十万条
- 商品类目数多，精准录入难：三级分类近4000条
- 主观理解商品类目划分错误：部分商品类目有重叠，难界定

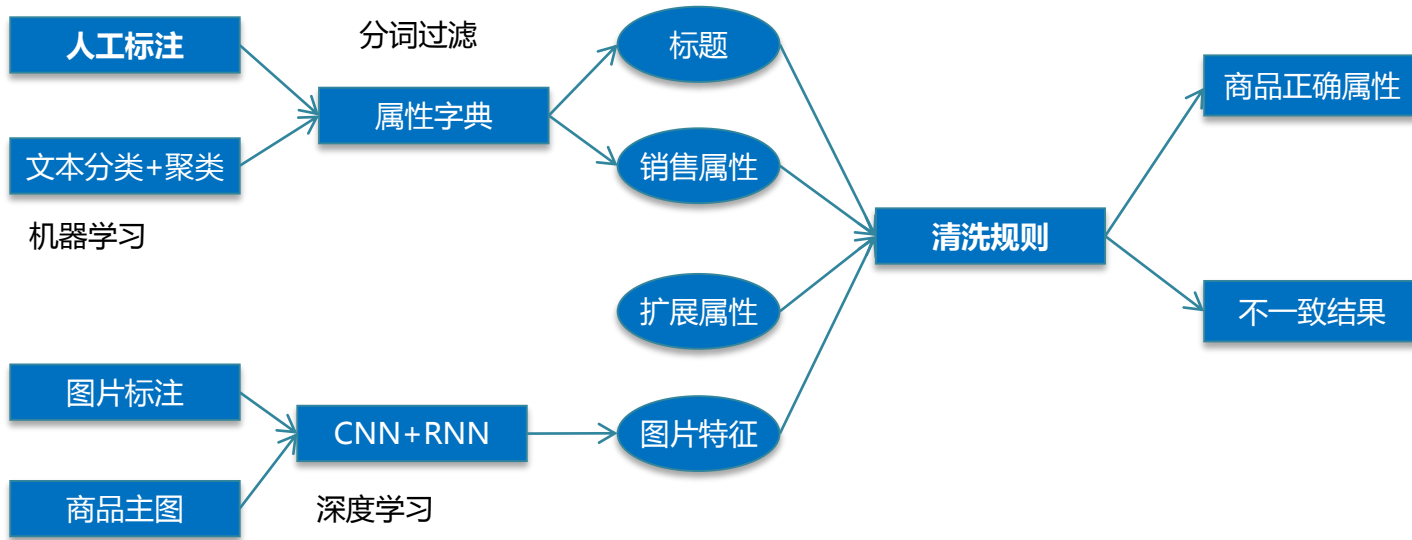
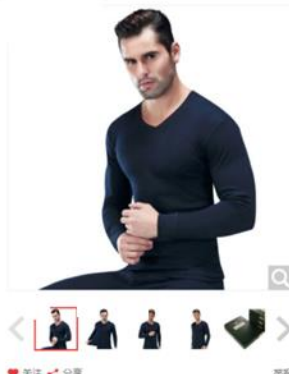


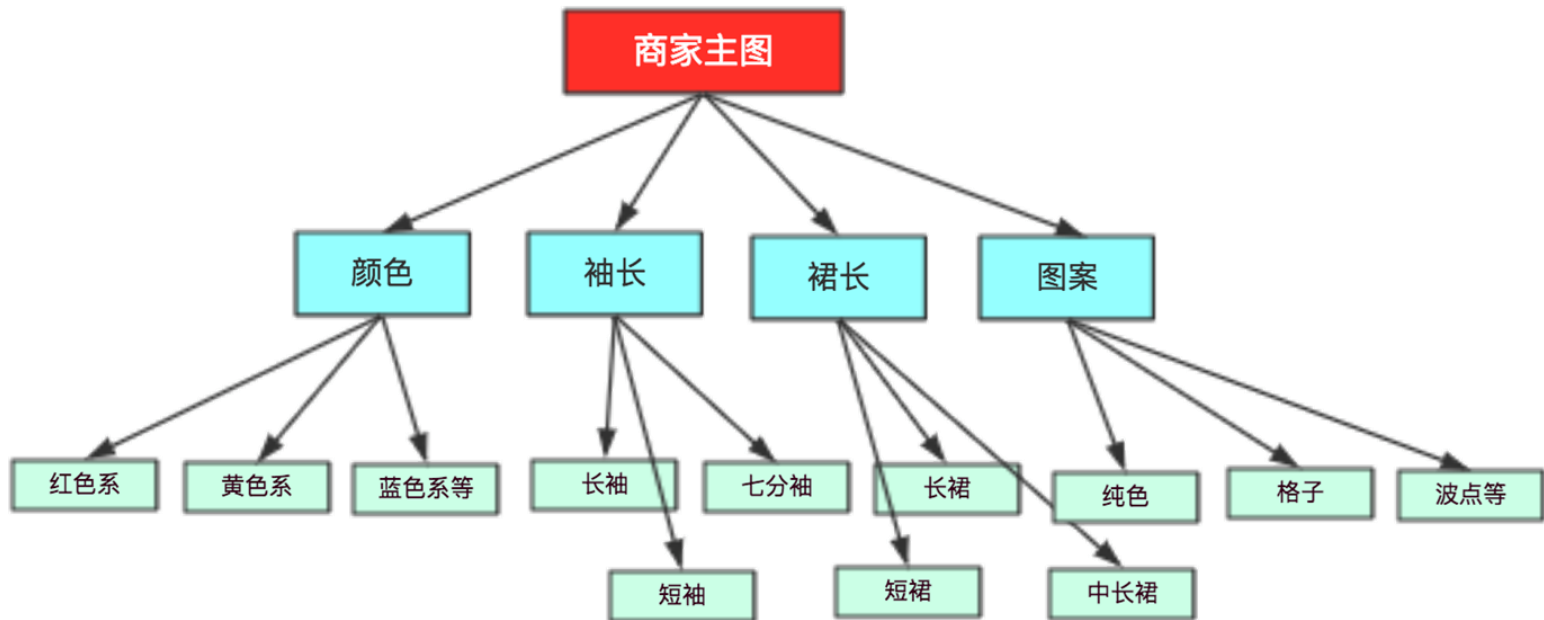
商品自动分类

- 在多个试点一级类目实现分类准确率99%
- 修正上千万SKU类目错绑属性

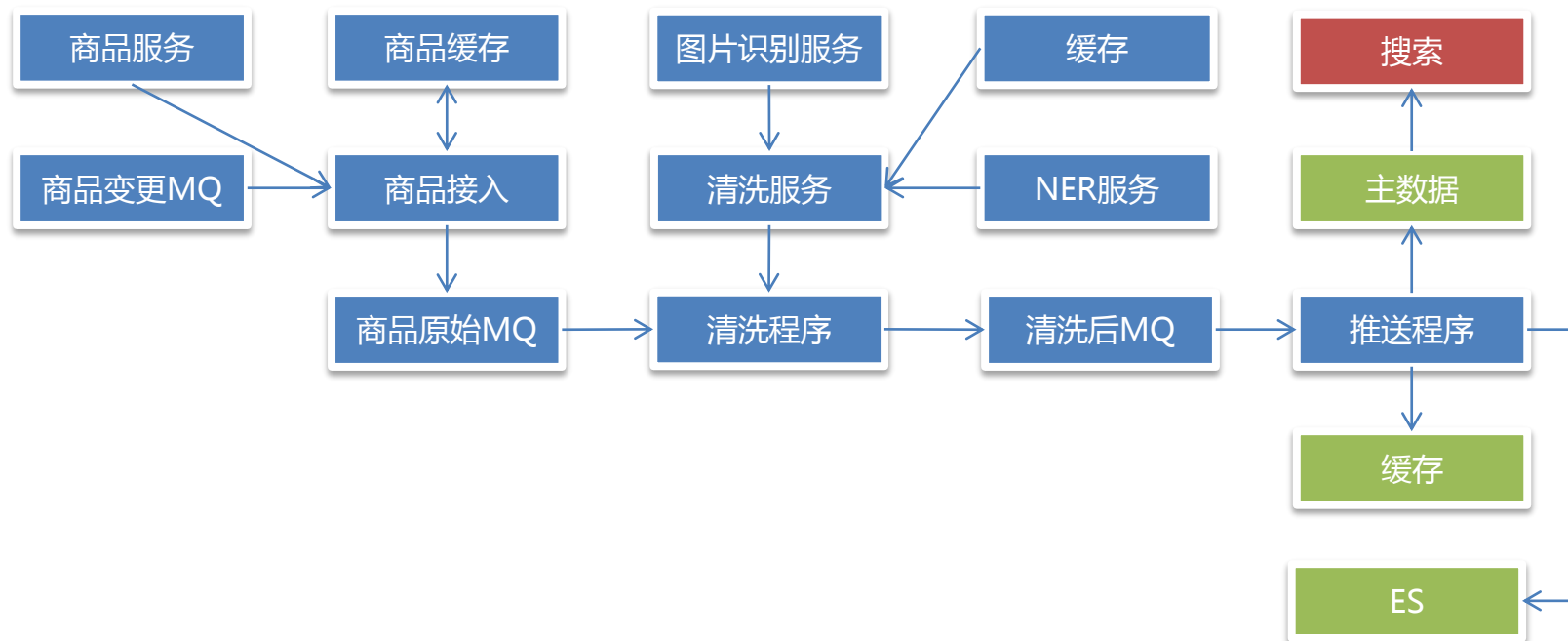
商品属性间不一致校验

JD.COM 京东





- 目前覆盖四个一级类目：运动户外、服饰内衣、鞋靴、礼品箱包
- 颜色识别覆盖61个三级品类，准确率95.65%



****B
京京国24625

★★★★★

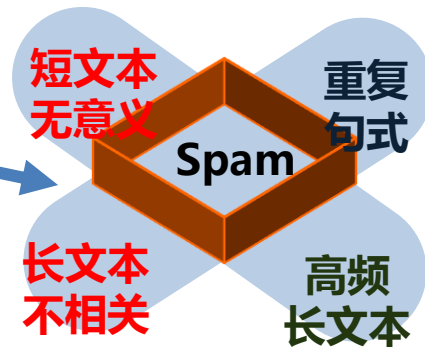
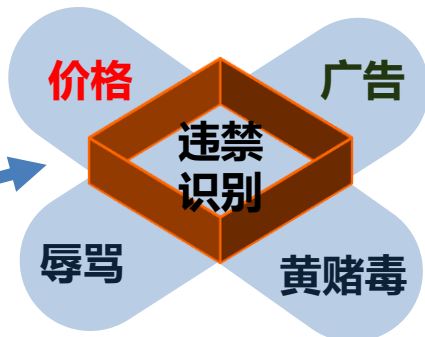
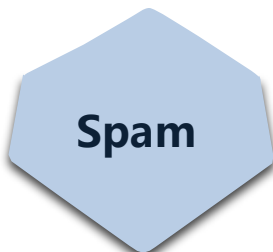
我不是果粉，因为某种功能才买的苹果，等了好久没有抢到优惠券，等不及了！6599买的，开始也担心质量问题，下单第二天收到货，马上在官网查了序列号，显示未激活，心放下一大半，外观没有问题，然后进行了激活了，查看保修期整整一年，完美新机，其他的电流声，漏光的问题，没有出现，用了几天，第一个不同就是快，指纹识别做的非常精准，关于电量，如果一直玩的话，一天肯定撑不下来的，手机没有问题，品质和制作工艺很完美很精致，大家可以放心购买，买手机一直在京东，大家不用打理那些黑京东的，不激活前现查序列号，连包装都不拆的，查出来如果有问题直接退货，就可以了，没有什么好担心的，最后赞下快递，京东快递服务真好，双手递交给我的，就一点足以说明服务质量，赞！！！！

刚使用，后期再来追评评论。



黑色 128GB 2017-07-12 08:51

举报 5 3



月***钧
PLUS会员

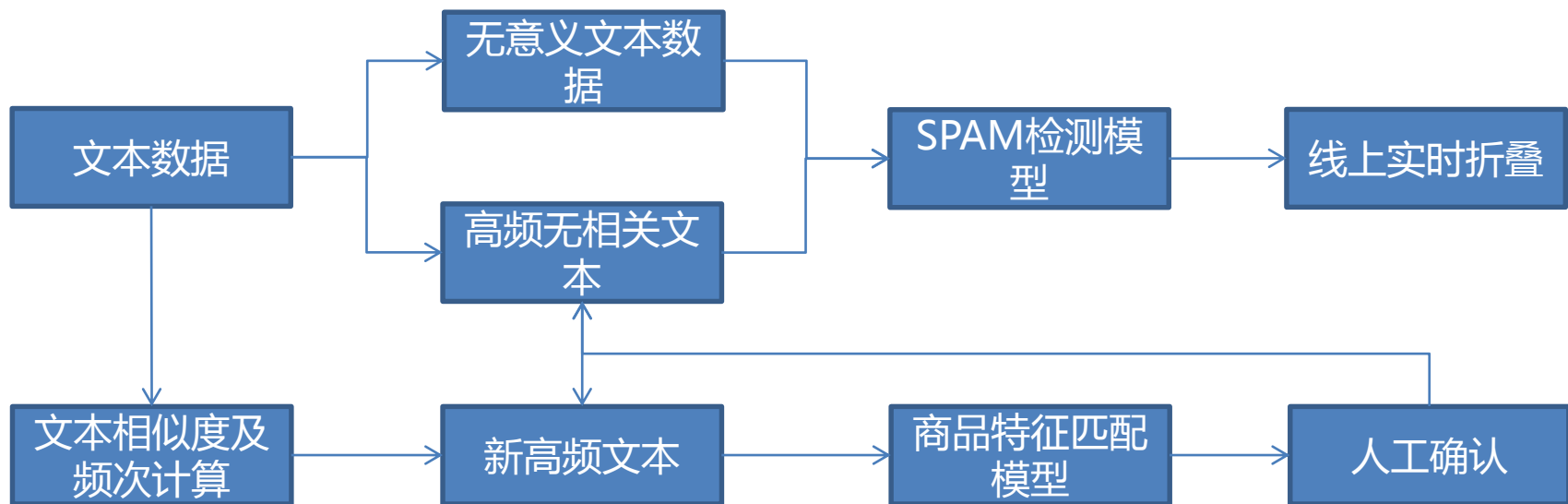
★★★★★

此评价和商品无关，商品质量很好，非常满意。tian猫店铺诈骗1200万跑路，阿里监管形同虚设 tian猫店铺北极绒楚翔时代专卖店和楚翔时代手表店利用tian猫15天售后规则篇走消费者1200万资金，受pian者在19日赶去湖南当地，pian子人去楼空。经过去工商局查询，注册地址是空的，电话号码是空号，这样的店铺在添猫是过了年审的。因为店铺是tian猫的五年和八年店铺，再加上tian猫有个售后规则是15天内可以申请售后，好多人就拍了对方店铺的睡衣和手表。结果发现出事之后，第一时间拨打tian猫售后电话或者咨询阿里小蜜，得到的回复都是钱没有问题，一定会给您退回来的。并且有电话录音和聊天截图。结果到了退款时间tian猫不仅不给退款还让客服介入，强行关闭的订单。在拍单的过程当中，pian子说过他的店铺tian猫从来不查，给pian子跑路的时间。异常订单存在两年之久，tian猫都无视，Ma云只关心金融，不管平台，任由pianzi猖獗活动，店铺异常不去监管，一出事就关了店铺处理，堂堂tianmao店铺玩空手套，根本没有实力，马云不作为只赚钱，唯利是图，没有责任和道义，星星之火，个人能力微不足道。转发让更多人看到，不要再上当了。

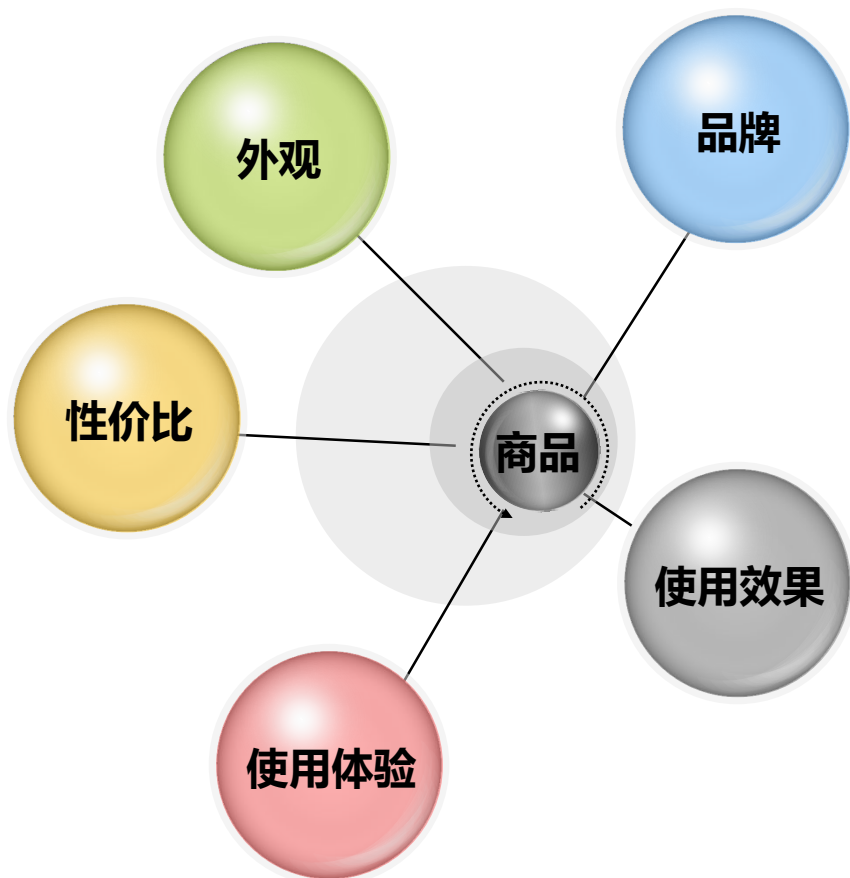
通话质量好

黑色 128GB 2017-07-08 22:39

举报 0 0

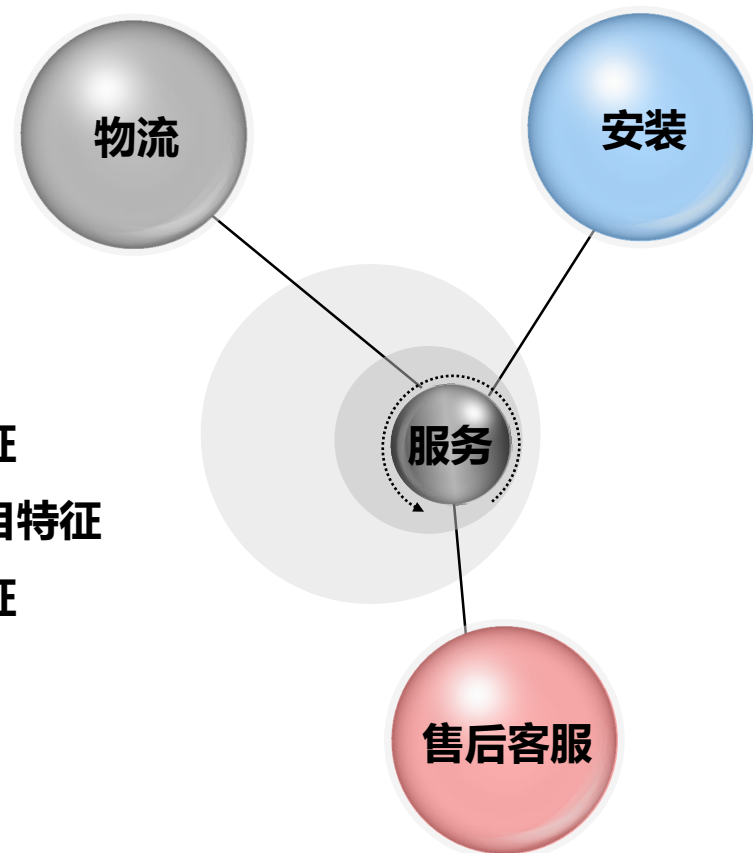


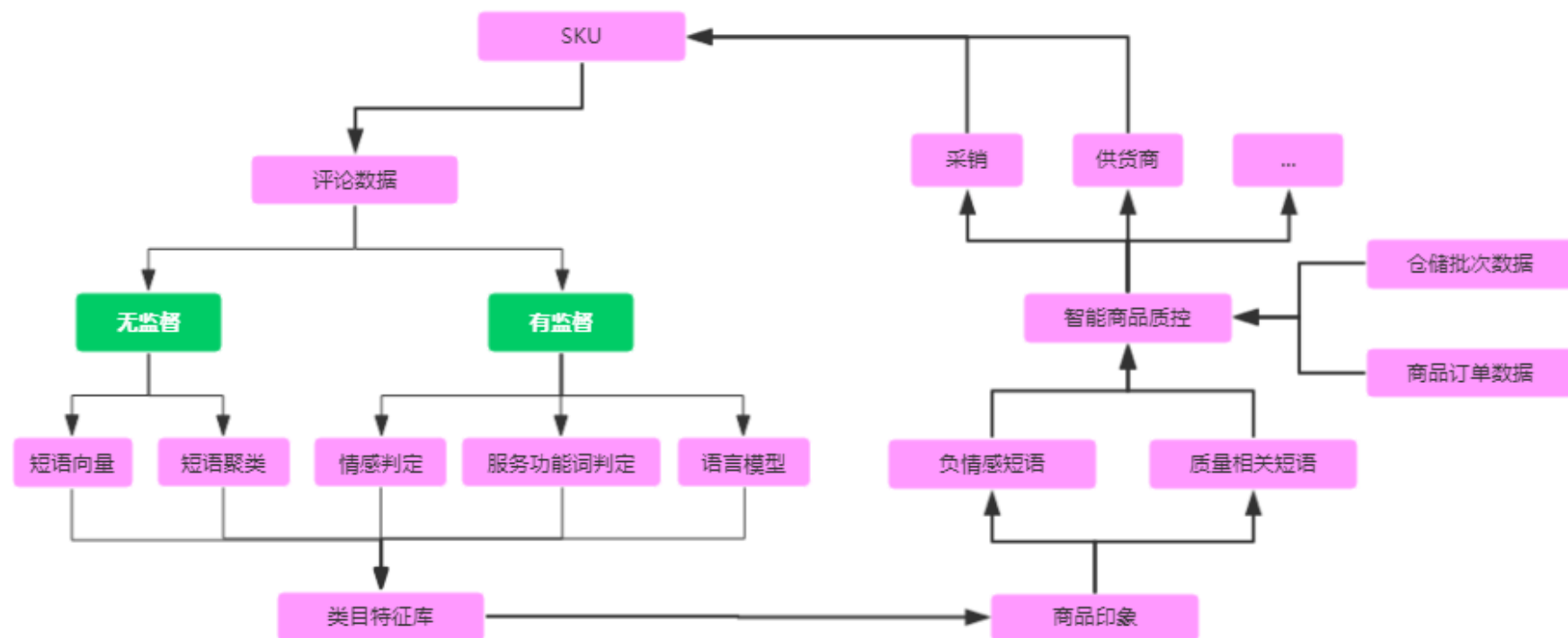
Spam高频长文本自动发现机制



多维度特征

- 服务特征
- 商品类目特征
- 品牌特征





- 评论的情感分析
- 商品与评论相关性分析
- 评论的关键词抽取与聚类
- 商品评论标签

商品评价

好评度

95%

是真品(15982)

就是快(11882)

物流很快(10169)

货

挺不错(2619)

品质值得信赖(1788)

服务态度好(1507)

全部评价(27万+)


晒图(500)

追评(4400+)

好评(25万+)

中评(5300+)

差


x***4

PLUS会员

★★★★★

手机用了快一个月，有些小毛病，但不影响使用，所以

金色 2017-07-15 10:34

j***6

京享值9706

★★★★★

上网查了，是正品，还是很好用的

- 商品评价标签

自然语言处理

深度学习

Spark

ES

机器学习

Flink

Storm

THANKS

