# 评论数据介绍

## 目标

抓取京东海尔旗舰店 家电商品评论数据（包含其所有洗衣机、冰箱、），详细大类列表[[1]](#endnote-1)见附录。

将抓取的评论数据采用分词和关键字过滤转换为结构化指标（71个字段），详细列表[[2]](#endnote-2)见附录。

## 爬虫环境介绍：

Python2.7：基础环境

Scrapy1.3.2：爬虫框架，用于爬取代理服务器列表；爬取京东评论信息

Redis3.2.100：内存数据库，使用其set结构，实现数据持续化，任务队列，解决任务去重等问题

其他包依赖：

requests: 校验代理服务器可用性

python-redis: 用于操作redis

scrapy\_proxies: 用于切换代理

## 主要工作介绍

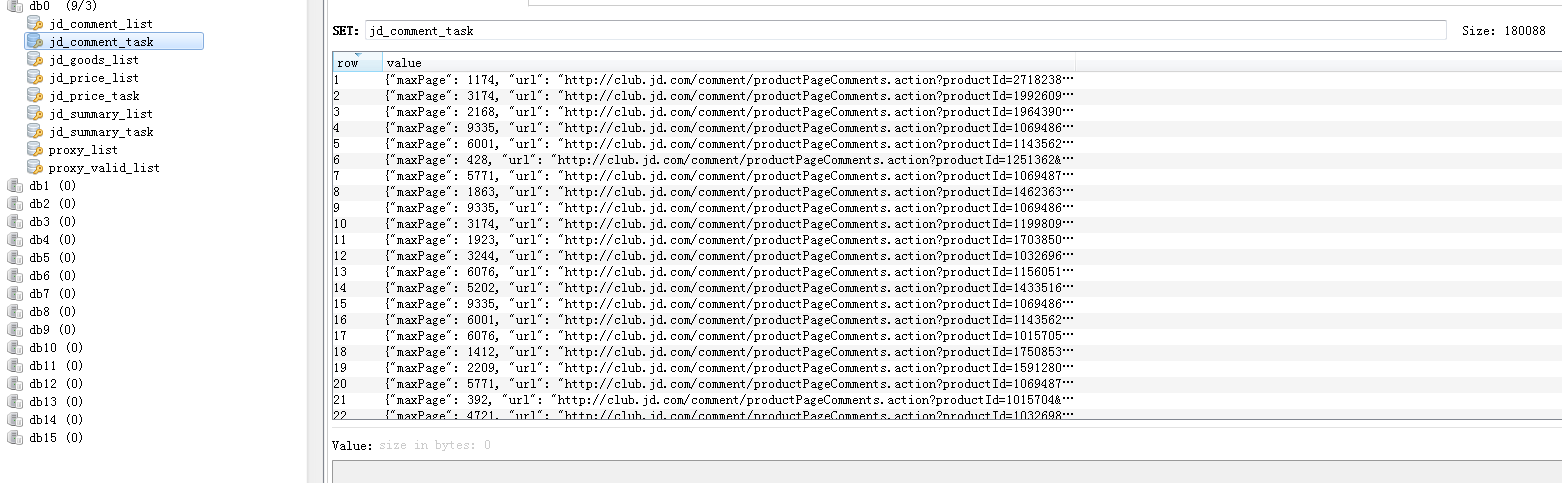
## 爬取并且检验代理服务器列表

通过scrapy爬取一些代理服务器列表(usproxy, proxydb)，通过requests库访问club.jd.com对代理服务器进行可用性校验。

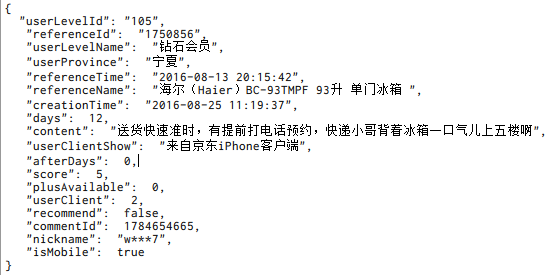
## 爬取京东评论数据

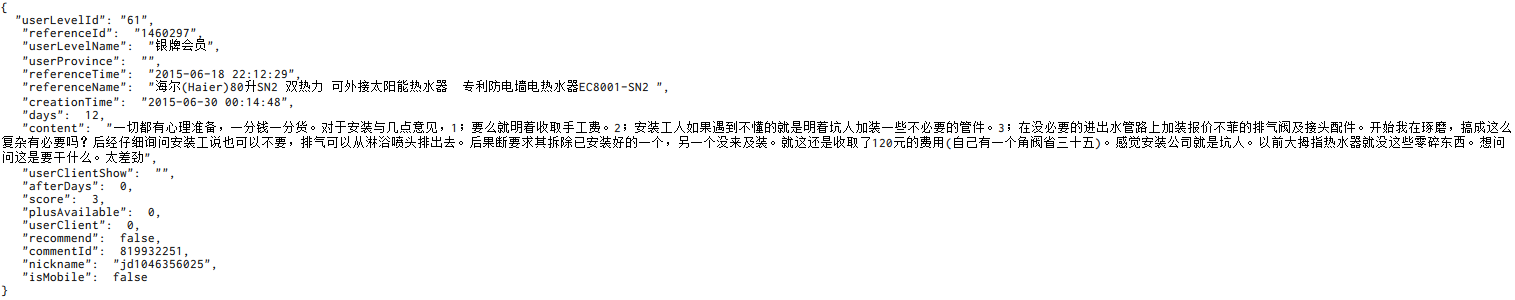
通过cat\_list获取所有商品数据(存储为redis set类型，名称jd\_goods\_list)，大概232个商品，同时生成jd\_price\_task和jd\_summary\_task；

分别根据jd\_price\_task和jd\_summary\_task获取jd\_price\_list和jd\_summary\_list，并且根据jd\_summary\_list生成jd\_comment\_task（大概18W条，每条对应10条评论数据）；



最后根据jd\_comment\_task不断抓取，得到jd\_comment\_list，抓取到如下所示的评论数据：





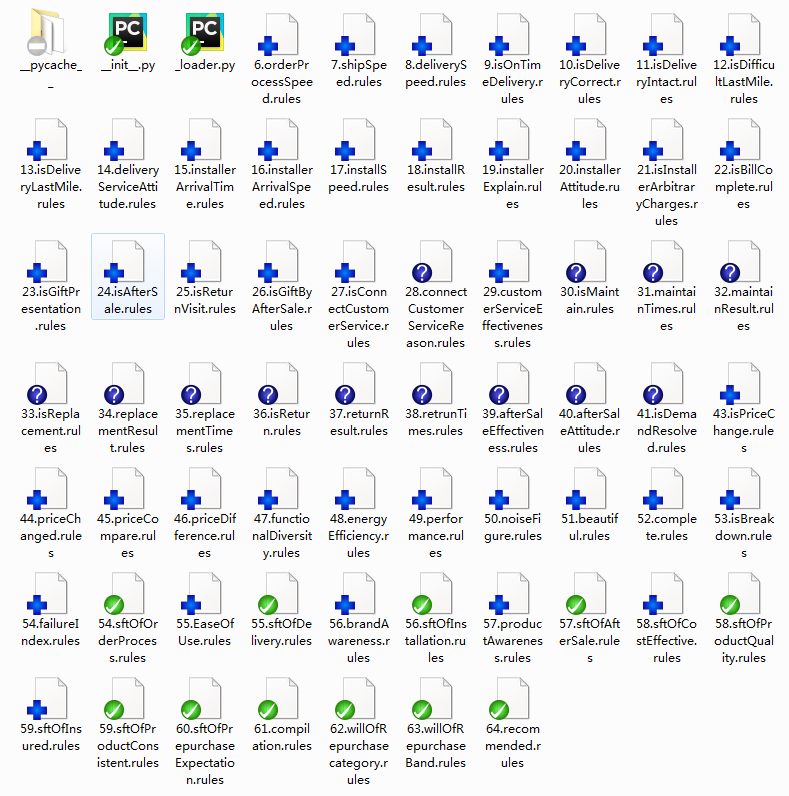
## 给予关键字进行指标转换

通过爬虫获取的评论数据，评论字段是非结构化数据，难以用于数据分析操作。因此需要对数据进行结构化处理，实际上这已经属于NLP范畴。

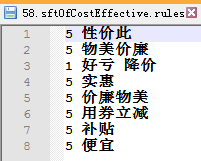
通过分析发现，评论内容往往观点比较突出，而且重复度相对较高，因此决定采用通过关键字进行过滤的方式对数据进行结构化处理。

使用开源分词库jieba进行分词处理，对于部分字段（抱怨成都），采用在线的api分析接口进行情感分析处理，列表[[3]](#endnote-3)见附录

关键字规则如下：



每个字段的规则结构如下：



## 难点

最大的难点数据的结构化处理，因为涉及到NLP，涉及到语义分析，虽然现有非常棒的分词系统（例如哈工大的语言技术平台），但是受专业和时间所限，无法进行句型和语义上的分析。并且鉴于评论数据的特殊性，采用关键字过滤方式对每个字段进行处理。

京东相关的评论获取接口存在反爬虫（不返回错误信息，只是不返回数据），因此一方面需要使用HTTP代理，另一方面受代理数量和质量限制，并且数据量比较大无法一次性抓取完成，需要对数据进行持久化。

真正抓取的时候，由于对京东反爬虫机制细节数据（频率，策略等）不清楚，凭经验是对IP的访问次数作限制，并且到达限制后如果持续抓取无效页面会延长限制时间。

因此必须要使用HTTP代理才能加快爬取速度。同时，redis可以解决数据持久化和任务队列的功能，但是无法解决

## 数据规模

原始京东评论数据>=100万条

## 附录

1. 商品种类列表：

   737,794,870

   737,794,13701

   737,794,880

   737,794,878

   737,794,12392

   737,794,12401

   737,13297,13298

   737,13297,1300

   737,13297,13117

   737,13297,13690

   737,13297,13691 [↑](#endnote-ref-1)
2. userProvince,

   userLevelName,

   price,

   userClientShow,

   createTime,

   orderProcessSpeed,

   shipSpeed,

   deliverySpeed,

   isOnTimeDelivery,

   isDeliveryCorrect,

   isDeliveryIntact,

   isDifficultLastMile,

   isDeliveryLastMile,

   deliveryServiceAttitude,

   installerArrivalTime,

   installerArrivalSpeed,

   installSpeed,

   installResult,

   installerExplain,

   installerAttitude,

   isInstallerArbitraryCharges,

   isBillComplete,

   isGiftPresentation,

   isAfterSale,

   isReturnVisit,

   isGiftByAfterSale,

   isConnectCustomerService,

   connectCustomerServiceReason,

   customerServiceEffectiveness,

   isMaintain,

   maintainTimes,

   maintainResult,

   isReplacement,

   replacementResult,

   replacementTimes,

   isReturn,

   returnResult,

   retrunTimes,

   afterSaleEffectiveness,

   afterSaleAttitude,

   isDemandResolved,

   equipmentModel,

   priceChanged,

   priceCompare,

   priceDifference,

   functionalDiversity,

   energyEfficiency,

   performance,

   noiseFigure,

   beautiful,

   complete,

   isBreakdown,

   failureIndex,

   EaseOfUse,

   brandAwareness,

   productAwareness,

   sftOfCost-effective,

   sftOfInsured,

   sftOfOrderProcess,

   sftOfDelivery,

   sftOfInstallation,

   sftOfAfterSale,

   sftOfProductQuality,

   sftOfProductConsistent,

   sftOfPrepurchaseExpectation,

   score,

   sftLevel,

   compilation,

   willOfRepurchasecategory,

   willOfRepurchaseBand,

   recommended, [↑](#endnote-ref-2)
3. http://ictclas.nlpir.org/nlpir/ [↑](#endnote-ref-3)