# 京东20W条数据统计清洗分析



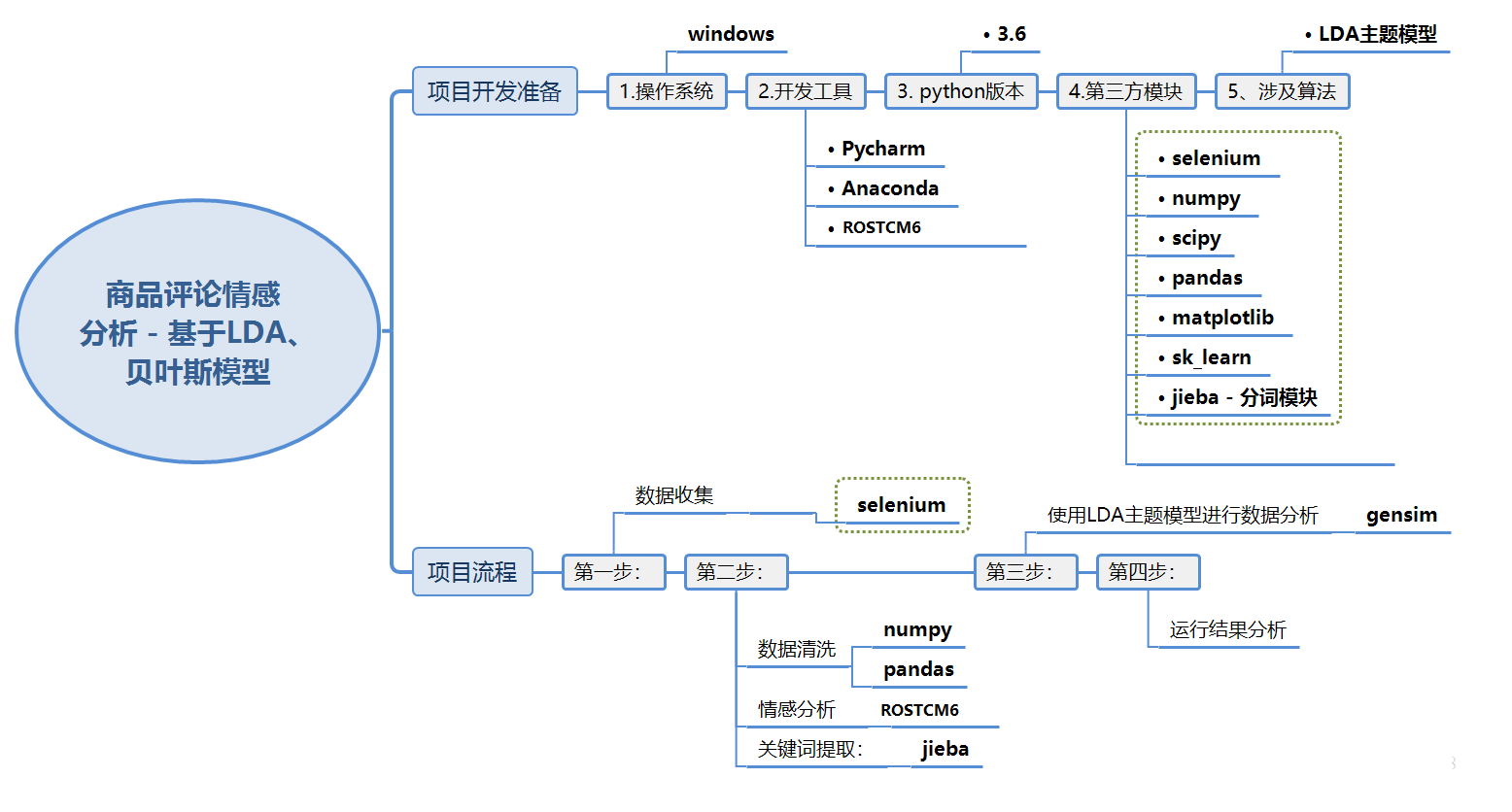
Readme

针对京东商城20W条美的热水器评论数据进行统计清洗分析。

分析出所有正面评论和负面评论。

分词统计热词出现频率。

分析出销售问题所在。



开发环境

本项目的文本情感分析使用的是基于情感字典的文本情感分析。

为了能够正确标注一段中文文本的情感。需要如下几个情感字典：

①停用词字典：用于过滤掉一段文本中的噪声词组。

②情感词字典：用于得到一段文本中带有情感色彩的词组及其评分。

③程度副词字典：代表情感词的强烈程度，相当于情感词的权重。

④否定词字典：用于判断其后情感词的意思究竟是好（正极性）还是坏（负极性），若情感词前有否定词，则情感得分-1。

情感字典以及评分通常由手工标注完成，而标注是一项费时又费力的活，因此这四个字典都是由网络搜集而来。

项目文本名称对应内容：{Huizong.csv :“所有评论信息”,  meidj\_jd.txt:“提取后的评论”,

 meidi\_jd\_process\_1.txt:“去重后的评论信息”, meidi\_jd\_neg.txt：“负面评论信息”， meidi\_jd\_pos.txt :“”正面评论信息”,stop\_list:“停用词表”}

数据采集

首先要进行数据采集

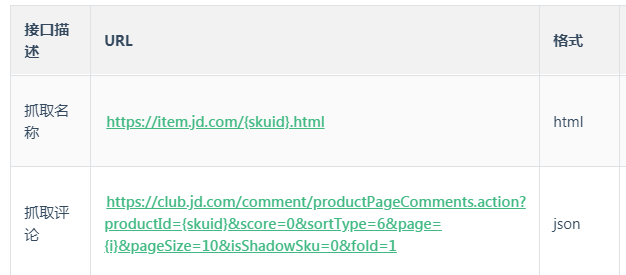
由于京东商品评论信息是由JS动态加载的。

直接抓取商品详情页的URL并不能获得商品评论的信息。

因此我需要先找到存放商品评论信息的文件。

这里我使用Chrome浏览器里的开发者工具进行查找。

捕获到的接口为：



发现根据更改page的值可以选择对应评论页。

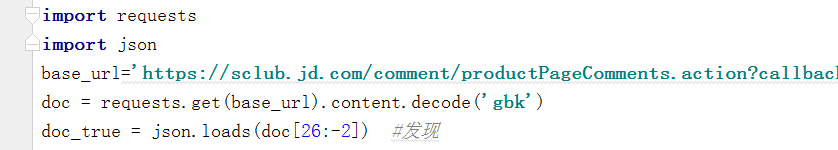
先测试下：

请求一页并提取对应评论信息：

通过requests来测试，发现获取到的content内容：

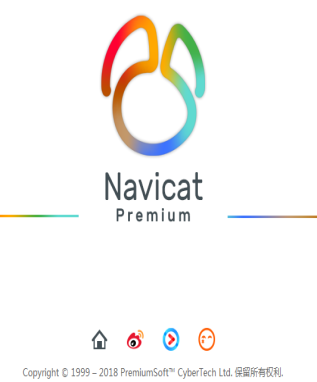
开头为:'fetchJSON\_comment98vv1204'，结尾为: ' ); '

发现content内容不是完整的json格式，所以要把doc的前面25位和后两位去掉；

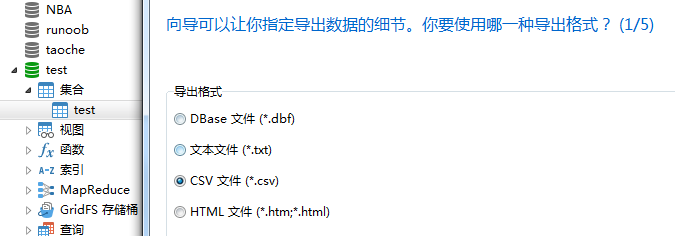
即：doc\_true = json.loads(doc[26:-2])

然后循环提取comment 即可得到评论内容；把提取到的数据保存到mongodb中。

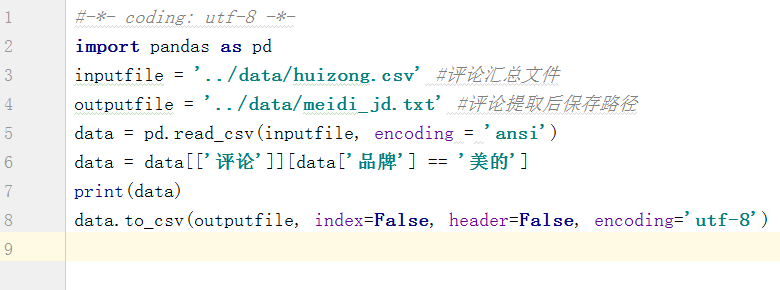
数据导出

首先在mongodb中使用去重语句aggregate 来进行集合内去重。

再通过图形界面Navicat来操作Mongodb数据库，进行数据导出。



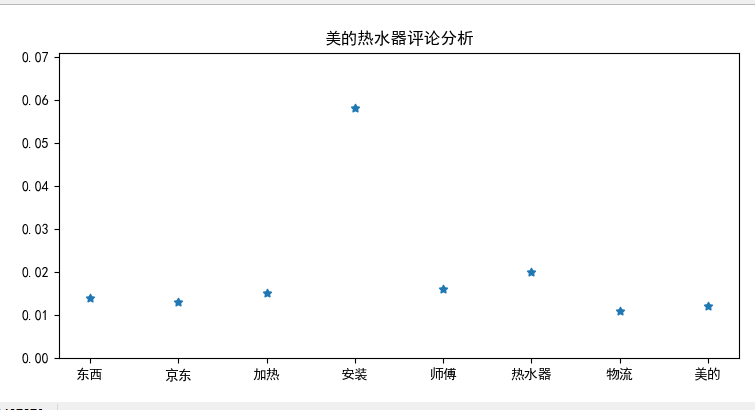
保存为csv后，使用pandas提取评论，并保存为meidi\_jd.txt；



|  |  |  |
| --- | --- | --- |
|  | 数据去重  此时已经提取出了名为meidi\_jd.txt的评论文件。  接下来要开始数据清洗工作，先进行数据去重。  我使用了pandas中的unique方法来进行去重，并保存为meidi\_jd\_process\_1.txt；        可发现，经过unique去重后数据减少了6371条。  数据分词  ROSTCM6是武汉大学沈阳教授研发编码的国内目前唯一的以辅助人文社会科学研究的大型免费社会计算平台。该软件可以实现微博分析、聊天分析、全网分析、网站分析、浏览分析、分词、词频统计、英文词频统计、流量分析、聚类分析等一系列文本分析。    通过ROSTCM6工具来提取出正面与负面评价(待分析的文件编码必须是ANSI)。  例：data=pd.read\_csv(…) data.to\_csv(r'xixi.txt', encoding='ANSI')  正面情感词：    负面情感词：    去除标签：  然后使用结巴分词进行简单分词。  先进行参数初始化，然后通过jieba.cut进行自定义分词函数；  通过“广播”形式分词，加快速度然后保存结果；  通过wordcloud库，生成简单词云图来查看正面情感。    查看负面情感词：    主题模型  根据上面的简易词云图，并不能有效的分析出销售问题所在。  所以我开始对分词之后的文档建立主题模型。    LDA（Latent Dirichlet Allocation）是一种文档主题生成模型，也称为一个三层[贝叶斯](https://baike.baidu.com/item/%E8%B4%9D%E5%8F%B6%E6%96%AF)概率模型，包含词、主题和文档三层结构。  所谓生成模型，就是说，我们认为一篇文章的每个词都是通过“以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语”这样一个过程得到。  文档到主题服从多项式分布，主题到词服从多项式分布。 |  |
|  | 结果分析  **输出负面主题后：**    通过上述分析可以看出在被判定为负面评论的关键词中，‘安装’的系数是最大的，其次主要就是‘商品本身’，可见在热水器销售的过程中，对用户评论有关键影响的主要是商品的安装服务和商品的质量，另外可以看出，‘客服’、‘物流’也是比较重要的；  **输出正面主题后：**    ‘好’、‘不错’等类似的关键词反应了顾客对应商品还是比较满意的，另外师傅’、‘安装’  的系数还是比较大的，另外‘及时’、‘很快’也一定程度上反应了物流的快慢对于用户好的购  物体验的影响。 |  |

我将输出出来的两种主题信息保存在到result\_id.txt中。

提取相同的数据然后把对应的数进行相加，再使用matplotlib绘制出散点图



根据结果总体分析来看商品的售后安装和商品的质量是顾客最为关键的考察点。

虽然本例中情感分析、分词的准确率并不能百分百的正确预测，但也在一定程度上反应了具体

由哪些因素主导了用户的购物体验及评论的好坏。