云南白药

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 品牌 | 商品评论特征 | | | |
| 云南白药 | 品牌 | 价格 | 牙龈 | 清香 |
| 中华 | 实惠 | 美白 | 服务态度 | 清新 |
| 佳洁士 | 美白 | 便宜 | 牙釉质 | 清新 |
| 黑人 | 味道 | 清新 | 实惠 |  |
| 欧乐B | 牙龈 | 味道 | 清新 | 性价比 |
| 皓乐齿 | 小朋友喜欢 | 薄荷味道 | 泡沫细腻 | 清新 |
| 舒适达 | 牙龈 | 清新 | 牙齿敏感 | 品牌 |
| 纳美 | 泡沫细腻 | 小苏打美白 | 清新 |  |
| 高露洁 | 实惠 | 泡沫细腻 | 治口气 | 清新 |
| 冷酸灵 | 牙龈敏感 | 品牌 | 清新 | 美白 |
| 舒客 | 清新 | 樱花口味 | 美白 |  |

1. SNOWNLP进行文本情感指数的分析，
   1. 用pandas库将评论存为dataframe类型的数据
   2. 用dropna去除评论中可能存在的空值
   3. 用unique方法去除评论中可能存在的重复值
   4. 用snownlp库将得出文本的情感值，并以此分为好评（情感指数>0.9），差评（情感指数<0.3）
   5. 将评论的情感曲线用matplotlib库建立核子图与直方图，得到各品牌的评价情感曲线进行横向分析
   6. 用JIEBA库进行分词,为LDA模型做准备
   7. 去除评论中的停顿词，如“啊”，“的”，“。”，‘京东’，“物流”等对评论分析无意义的词。将去除后的文本类型改为dataframe（分词会将其存为单列的series）
   8. 用d步骤中得到的去除停顿词后的文本字典（已被分词）建立语料库
   9. 用e步骤得到的语料库建立LDA模型并存为html文件
   10. 分别对好评、差评评论重复步骤a-g