第二问的e小问:

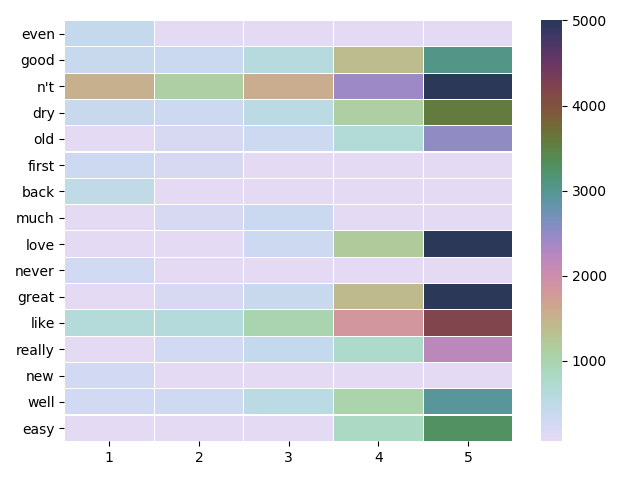
题目要求根据文本的评论的特定质量描述符, 和评级相关联, 挖掘其中是否有深层次信息. 为此本文通过对review\_body的具体内容进行挖掘, 提取其相关描述信息, 并通过统计和绘图等方式进行了挖掘.

具体步骤如下:

1. 建立特定的质量描述符词典. 通过查阅资料和相关数据, 本文建立了用户情感-评论-程度词典, 词典中包含英文的各种描述情感、评论词汇、程度词汇等等, 具体词汇表见附录.
2. 将吹风机, 微波炉, 尿不湿数据合并, 同时提取review\_body内容. 按用户评级分类, 分别从review\_body中统计出现在词典中的词语, 通过词频统计方法得到各个词语出现的频次, 将结果按频次降序排列后取出前20个词语. 得到如下表格xxx:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 |
| old | 272 | 225 | 353 | 677 | 2512 |
| never | 298 | 0 | 0 | 0 | 0 |
| long | 241 | 186 | 327 | 563 | 1392 |
| new | 286 | 0 | 0 | 0 | 0 |
| good | 392 | 368 | 591 | 1402 | 3058 |
| much | 228 | 234 | 370 | 606 | 2079 |
| back | 460 | 212 | 0 | 0 | 0 |
| bad | 222 | 0 | 0 | 0 | 0 |
| high | 0 | 0 | 232 | 0 | 0 |
| however | 0 | 0 | 265 | 0 | 0 |
| need | 230 | 157 | 252 | 566 | 1645 |
| receive | 222 | 0 | 0 | 0 | 0 |
| keep | 0 | 196 | 261 | 479 | 1695 |
| cute | 0 | 0 | 0 | 0 | 1479 |
| perfect | 0 | 0 | 0 | 0 | 1600 |
| star | 0 | 0 | 0 | 440 | 0 |
| hot | 0 | 187 | 0 | 0 | 0 |
| easy | 0 | 0 | 0 | 832 | 3261 |
| dry | 386 | 326 | 535 | 1108 | 3578 |
| n't | 1509 | 1102 | 1594 | 2445 | 5932 |
| really | 237 | 268 | 417 | 784 | 2209 |
| want | 255 | 157 | 274 | 437 | 0 |
| love | 0 | 159 | 359 | 1184 | 8867 |
| great | 221 | 237 | 388 | 1405 | 5797 |
| nice | 0 | 0 | 0 | 558 | 1405 |
| small | 0 | 165 | 279 | 549 | 0 |
| like | 642 | 602 | 984 | 1851 | 4186 |
| best | 0 | 0 | 0 | 0 | 1448 |
| even | 401 | 208 | 233 | 0 | 0 |
| first | 355 | 223 | 0 | 0 | 1363 |
| still | 228 | 0 | 255 | 506 | 0 |
| set | 0 | 163 | 272 | 607 | 1378 |
| well | 285 | 302 | 531 | 1019 | 2932 |

利用表格数据绘制热力图(图xxx)进行直观分析.



图中横坐标为星级, 纵坐标为单词, 从图和表中可以看出, 评分为一星的评价出现的单词最多的是” n’t ”是个否定词, 出现了1509次, 接下来的是never, like等; 而评分为5星的出现最多的单词为love(8867次), 其次为great, like等等, 该单词在一星或二星中几乎未出现. 这表明特定的描述词汇与评级是密切相关联的. 同时, 我们也注意到了, “ n’t ”在各个星级中都出现了多次. 通过具体的review\_body数据分析, 在一星级中其多与其他积极词汇同时出现用以表达否定情感, 例如” n’t like ” 等等. 而在五星数据中其常形成许多积极情感的短语, 例如” like n’t allow finger print”等等.

通过上述分析, 我们总结出如下的星级常用词汇表xxx:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 |
| not like | not good | not like | like | love |
| never | not like | not good | well | great |
| back | old | much | great | easy |
| even | not well | old | not | like |
| bad | dry | however | need | well |