

Homework 3：挖掘用户关注 实验报告

Stu: 18级自动化 181870078 黄廷基

Date: 2021.8.9

1. 任务描述

1.1 要求

针对淘宝平台中部分商品的用户评论数据，分别使用文本关键词提取方法以及主题模型，提取其中的用户关注点。

(1) 关键词提取：对于所有数据，进行文本关键词提取，挖掘该品类整体上用户所关注的焦点。

(2) 主题模型：面对浩如烟海的文档，把相似的文章聚合起来，并且提取描述聚合后主题的重要关键词，通过主题模型，挖掘在该品类中用户关注的几个主要话题及对应的话题内容。

1.2 需求分析

- 选取数据：服装品类 + 自爬数据
 - 关键词提取模型：Word2Vec + Kmeans
 - 主题模型选择：LDA模型
-

2. 数据爬取

2.1 运行环境

软件：Pycharm2020 (Python 3.9) 、MySQL Workbench

文件：data_get.py、taobaocookie.py、create_database.sql

扩展包：selenium、pymysql

模拟浏览器版本：Chrome/92.0.4515.131

2.2 爬取数据展示

| Result Grid | | | | | | | |
|---|--------|---------------|---------------|------------|------------------------|-----------------------------------|--------|
| Filter Rows: | | | | | | | |
| Edit: Export/Import: Wrap Cell Content: Fetch rows: | | | | | | | |
| | tb_num | url | cus_id | cus_name | cus_rank | re_content | re_pic |
| ▶ | 1 | https://it... | 1142442625103 | m***t (匿名) | //img.alicdn.com/tp... | 买了很多次了也买过初旭的还是最爱这个牌... | |
| | 1 | https://it... | 1123886443778 | z***0 | //www.alicdn.com/a... | 很好满意持续购买 | |
| | 1 | https://it... | 1141441705413 | k***n (匿名) | //img.alicdn.com/tp... | 15天内买家未作出评价,' | |
| | 1 | https://it... | 1127773690108 | t***1 | //www.alicdn.com/a... | 买了几多了 | |
| | 1 | https://it... | 1130048699256 | t***6 | //www.alicdn.com/a... | 包装品质: null商品分量: null保质期: null新... | |
| | 1 | https://it... | 1123569526701 | 沈***卿 (匿名) | //img.alicdn.com/tp... | 评价方未及时做出评价,系统默认好评! | |
| | 1 | https://it... | 1141665416679 | 我***! | //www.alicdn.com/a... | 很好吃哦可以选择在有活动的时候买,会比... | ','//m |
| | 1 | https://it... | 1141249896132 | 我***7 | //www.alicdn.com/a... | 东西好大一包比超市便宜多了赞赞赞好评以... | |
| | 1 | https://it... | 1133822011582 | 消***0 | //www.alicdn.com/a... | 给弟弟买的他很喜欢 | |
| | 1 | https://it... | 1142571419492 | 黑***1 | //www.alicdn.com/a... | 夹在小番茄里面味道还不错 | ','//m |
| | 1 | https://it... | 1142282025464 | t***8 | //www.alicdn.com/a... | 速度真的非常快了,不管是发货还是物流,... | |
| | 1 | https://it... | 1133481517417 | t***4 | //www.alicdn.com/a... | 产品好价格便宜 | |

2.3 一些修订

- data_get.py里面的pymysql.escape_string()方法, 最新的版本需要 from pymysql.converters import escape_string 来直接调用。
- taobaocookie.py里, getTaobaoCookie()里循环的browser.quit()需要去掉, 否则只爬取一条webdriver就关闭了。

3. 关键词提取——Word2Vec + Kmeans

3.1 运行环境

软件: Pycharm2020 (Python 3.9)

文件: word_break.py、get_vector.py、train_word2vec.py、get_keyword.py

扩展包: selenium、pymysql

模拟浏览器版本: Chrome/92.0.4515.131

3.2 模块功能

```
## word_break.py: 去除停用词
#
## get_vector.py: jieba分词、去重等数据处理
#
## train_word2vec.py: 训练word2vec模型, 得到词向量
#
## get_keyword.py: 进行Kmeans聚类, 按照组内距离之和, 得到Top3的关键词
#
```

3.3 结果展示

- 去除停用词结果:

data_seg.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

一个月 退款 服务态度 挺
衣服 挺 好看 做工 稍微 粗糙
衣服 质量 春秋 皆宜 加 一件 毛衣 初冬 OK 尺码 标准 173 60 公斤 选 M 合身 穿 得体 做工 精细 面料 舒适 喜欢 那种
布料 亲切感 无色差
卖家 衣服 合适 很快 换 一件
质量 不错
挺 好看 滴
堂弟 学期 区 参加 模特 赛 获奖 特意 送 一套 大获全胜 西装 合适 质量 不错 值
适合 春天
天气 穿 刚刚
穿 好看 噫 搭配 款式 码数 买小一码
厚实 保暖 转冷 穿 送 男朋友 喜欢
修身 很棒
看着 不错 发货 速度 很快
衣服 质量 不错 上身 效 好看 做工 完美 面料 柔软 亲肤 码数 合适
ok
好看 大小 合适 服务态度 好评
好评 好看
真的 好看 版型 超好
整体 评价 完美 男朋友 穿 这家 店 衣服 买 两件 尺码 推荐 身高 155 体重 103 码数 S 合身 面料 品质 舒服 厚实
太 好看
姐妹 买

- jieba分词、去重等数据处理结果：

data_result.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

一个月 退款 服务态度 挺
衣服 挺 好看 做工 稍微 粗糙
衣服 质量 春秋 皆宜 加 一件 毛衣 初冬 尺码 标准 公斤 选 合身 穿 得体 做工 精细
面料 舒适 喜欢 那种 很快 换 一件
卖家 衣服 合适
质量 不错
挺 好看 滴
堂弟 学期 区 参加 模特 赛 获奖 特意 送 一套 大获全胜 西装 合适 质量 不错 值
适合 春天
天气 穿 刚刚
穿 好看 噫 搭配 款式 码数 买小一码
厚实 保暖 转冷 穿 送 男朋友 喜欢
修身 很棒
看着 不错 发货 速度 很快
衣服 质量 不错 上身 效 好看 做工 完美 面料 柔软 亲肤 码数 合适
好看 大小 合适 服务态度 好评
好评 好看
真的 好看 版型 超好
整体 评价 完美 男朋友 穿 这家 店 衣服 买 两件 尺码 推荐 身高 体重 码数 合身 面
料 品质 舒服 厚实
好看 买
姐妹 真不错 面料 舒适 质量上乘 穿外 身上 舒服 品牌 好多 满意 体恤 这家 买
衣服 材质 穿 宽松 型 春天 大码 好看
姐妹 买

- Word2Vec模型参数设置：

```
model = word2vec(Linesentence(inp), window=5, vector_size=100, min_count=5,
sg=1, hs=1, workers=25)
```

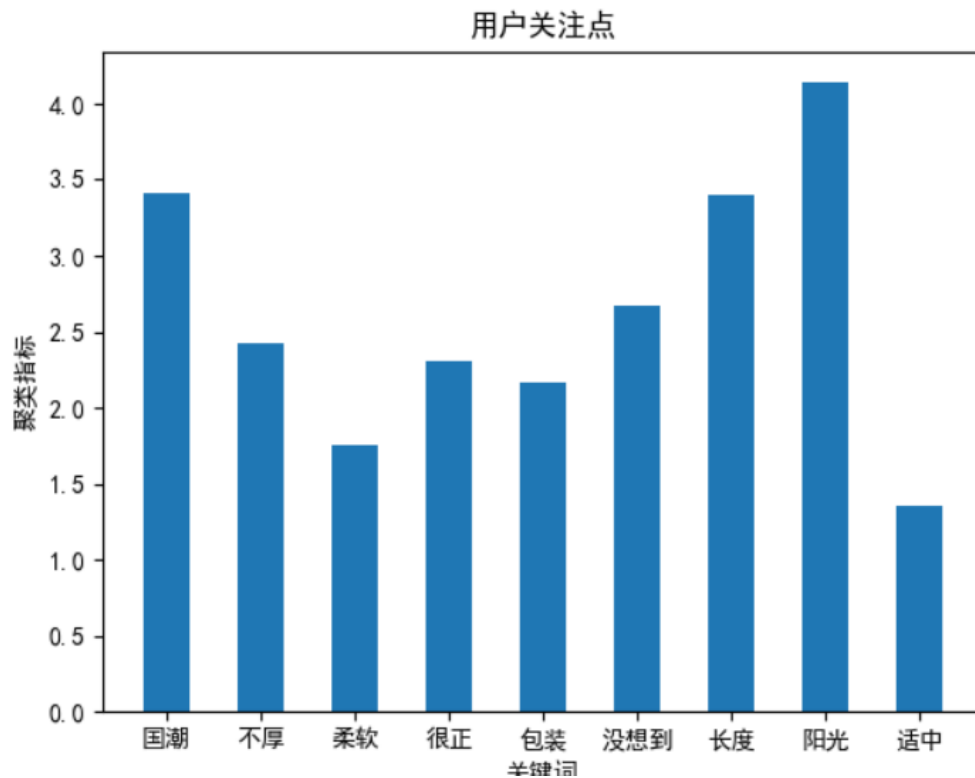
- Word2Vec词向量部分展示：

```
wordvector = (list: 578) [array([-0.15091664, 0.22360222, 0.06165149, -0.01365666, -0.03156255, \n
> 000 = (ndarray: (100,)) [-0.15091664 0.22360222 0.06165149 -0.01365666 -0.03156255 -0.26031724, 0.13314614, 0.32982013, -0.06481525, -0.1408048 -0.05...View
> 001 = (ndarray: (100,)) [-2.07006633e-01 3.01348329e-01 9.15216729e-02 1.42821699e-01, 2.60444300e-04 -3.13130111e-01 1.52687922e-01 4.13871408...View
> 002 = (ndarray: (100,)) [-1.6664900e-01 1.5320301e-01 -3.4515336e-02 -9.0154544e-02, 1.6314458e-02 -2.7336058e-01 8.1435353e-02 3.0579746e-01, -6.6...View
> 003 = (ndarray: (100,)) [-0.11353467 0.16551855 -0.0428595 -0.01063439 0.0484028 -0.31240475, 0.05528349 0.31187958 -0.00755838 -0.34608656 0.098...View
> 004 = (ndarray: (100,)) [-0.17536221 0.2066586 -0.01322915 -0.03186357 -0.02899127 -0.22125015, 0.13101956 0.36704352 -0.08495189 -0.17794192 -0.06...View
> 005 = (ndarray: (100,)) [-0.2184721 0.09380061 -0.06439135 -0.1079153 -0.04184093 -0.24799488, 0.11324737 0.3133007 -0.05324604 -0.3424686 -0.0075...View
```

- Kmeans聚类中心：

```
['国潮', '不厚', '柔软', '很正', '包装', '没想到', '长度', '阳光', '适中']
```

- 统计结果：



- 排除掉英文数字及语气词汇，可以看出关键词主要排序为：
阳光 > 国潮 > 长度 > 不厚 > 很正 > 包装 > 柔软 > 适中
- 从评价关键词可以看出，消费者对服装商品主要的评价集中在衣服的款式、质量上，这与常理相符合。

4. 主题模型——LDA

4.1 运行环境

软件：Pycharm2020 (Python 3.9)

文件：lda.py

扩展包：gensim.models.ldamodel

4.2 模型结果

- 数据使用第三部分中处理后的数据
- 模型参数设置：

```
dictionary = corpora.Dictionary(texts)
corpus = [dictionary.doc2bow(text) for text in texts]

lda = gensim.models.ldamodel.LdaModel(corpus=corpus, id2word=dictionary,
num_topics=20)
# 将单个主题作为格式化字符串
# 返回: 主题的字符串表示, 如'-0.340 *"类别"+ 0.298 *"$ M $" + 0.183 *"代数"+
...'.
# topicno: 主题ID, 这里是10
# topn: 将使用的主题中的单词数
print(lda.print_topic(10, topn=5))
```

- 模型输出结果:

```
0.128*"好看" + 0.064*"衣服" + 0.045*"喜欢" + 0.032*"超级" + 0.028*"质量"
```

- 与上面第三节的关键词结果不同, 主题模型呈现的结果显得更加“浓缩”, 能直接指出了评价的主题为“衣服”、“质量”。

5. 总结与思考

5.1 问题思考——应用场景

- 对于“消费者评价关键词”的任务需求, 可应用到商家服务方面, 比如: 给商品智能取名, 在商品后缀上添加评价关键词, 更容易吸引消费者;
- 对于“评价主题”的任务, 则可以用于平台智能筛选不相关评价, 比如在数据中常常见到“这是一条凑字数的评价”类似的评论, 可以通过每条评价与主题相关程度进行屏蔽等操作;

5.2 不足与改进

- 数据清洗效果不够理想: 在清洗时只保留了中文, 但后来发现符号、英文字母并不是都没有意义, 比如‘nice’、‘!’等, 都表达了消费者的评价; (这可能需要增加模型难度来改进)
- 分词结果不好: 本次作业是直接使用了中科院的中文分词, 但在商品评价上直接使用有些欠妥, 可能要考虑在此基础上进行一些更贴合场景的增删;