文章编号: 2095-2163(2021)11-0128-03 中图分类号: TP393.092 文献标志码: A

医药商品评论的情感分析

郭小磊
(山西中医药大学 健康服务与管理学院, 太原 030619)

摘 要: 本文利用 Python 语言获取某平台医药商品评论文本,采用正负预料库进行 LDA 模型训练和情感分值的计算,并使用 Matplotlib 方法和 Wordcloud 对处理之后的数据进行可视化。通过这种方法可以有效、精准获取京东商品评论,并对其进行情感分析,对提高工作效率和数据分析成效均具有积极的作用。

关键词: Python; LDA 模型; 可视化; 情感分析

Sentiment analysis of pharmaceutical commodity reviews based on Python

GUO Xiaolei
(School of Health Service and Management, Shanxi University of Traditional Chinese Medicine, Taiyuan 030619, China)

【Abstract】 In this paper, Python is used to obtain the review text of a medical product on JD.com, LDA model training and emotion score calculation are carried out by using positive and negative prediction library, and the processed data are visualize by using Matplotlib method and Wordcloud. This method can effectively and accurately obtain jd product reviews and analyze the emotional tendency of users who review them, which is of positive help to improve work efficiency and data analysis effectiveness.

【Key words】 Python; LDA model; visualization; sentiment analysis

0 引 言

网上购物已经成为中国消费的主要方式,其模式越来越成熟,产品评论数量也非常巨大。本文利用 Python 语言获取某平台医药商品评论文本,通过有效的方法精准获取商品评论数据,采用 LDA 模型分析评论用户的情感倾向。通过数据分析结果帮助平台和厂商及时调整营销策略和产品改进。

1 京东商品评论数据获取

1.1 网络爬虫技术原理

网络爬虫是一种按照一定规则自动提取 Web 网页中应用程序或脚本的技术[1]。首先,从 URL 中对页面源文件进行分析; 然后准确抓取新 Web 链接,并以此为基础寻找新 Web 链接,直至完成全部页面的准确抓取和分析[2]。Python 是一款开源的编程语言,自带众多第三方库,其中的 requests 库可以帮助用户抓取 URL 内容。

1.2 文本信息获取

本文以某购物网站平台下补肾类药品为例,产品的评论解析过程及文本爬取思路如下:

(1) 准备 Requests 库、random 库和 time 库。其中,Requests 库用于实现爬虫功能; random 库、time 库用于伪装用户抓取的随机时间,防止网站封锁。

(2) 登录并获取评论。

(3) 检查,重新加载,搜索评论,在 Headers 中找到 Request URL。对商品信息进行爬取,并设置循环抓取数据。利用 Python 爬取网页信息需要伪装浏览器 user-agent 和用户,cookie 发送请求。

(4) 循环请求每页评论网址,采用正则表达式爬取结果。由于该网站对爬虫的限制,只能爬取 100 页评论,共 1 076 条评论数据。

2 基于 LDA 模型的情感分析

情感分析是一个带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程[3]。利用机器提取分析文本的情感色彩,发现文本中人们对某人或事物的态度,从而发现潜在的问题用于推演改进或舆情预测。一篇文档通过文本的数据特征提取后,形成若干个主题,主题包含了单个或数个词所不能体现出有价值的信息,由共同出现频率最高的词组成。主题提取的相关算法,包括 TF-IDF 算法、LSI 算法、PLSI 算法和 LDA 算法[4]。LDA 模型是一种典型的词袋模型,即一篇文档是由一组词构成,词与

基金项目: 山西省高等学校教学改革创新项目(J2021453); 山西省高等学校科技创新计划(2020L0430); 山西中医药大学科技创新能力培育计划(2019PY-032)。

作者简介: 郭小磊(1981-),女,硕士,讲师,主要研究方向: 计算机应用、医学信息数据挖掘。

收稿日期: 2021-08-15

词之间没有先后顺序的关系，是一种简单有效的情感分析方法。

2.1　数据清洗

获取的评论不能直接利用分析，要通过数据的清洗。具体包括清除缺失数据、数据去重、定义机械压缩去词函数和短句过滤等方法。

(1) 清除缺失数据;

(2) 去掉第一列的重复数据;

(3) 对重复的评论进行压缩，通过定义函数，调用 apply 方法进行机械压缩去词;

(4) 过滤评论信息只有 4 个字符的无参考意义的评论。经过一系列操作后有 1 060 条数据符合要求。

2.2　由评论内容生成的词云

词云是对文本数据中出现频率较高的词在视觉上的突出呈现，从而直观领略文本数据的主要表达意思[5]。利用 python 中 jieba 库对评论内容进行分词，排名前 20 的词分别为 “京东”，“效果”，“包装”，“感觉”，“肾宝”，“购买”，“服用”，“正品”，“物流”，“没有”，“值得”，“希望”，“价格”，“活动”，“收到”，“快递”，“信赖”，“东西”，“品牌”，“身体”；再利用 wordcloud 模块绘制词云图，如图 1 所示，其中使用停用词表筛选掉一些词频较高的无效词，以求词云图达到更好的呈现效果。

图 1　评论词云图

Fig. 1　Comment wordcloud

2.3　文本情感分析

将获取评论文本转换为 csv 格式，这部分评论不仅包含了评论者的主观情绪，还可能因为被其他用户点赞而暗含点赞用户的情感倾向，但是这些评论的情感分值也可能因为其他用户的点赞发生偏差。基于评论文本文档进行 LDA 模型的分析，Python 的 snowNLP 情感分析是基于情感词典实现的，其简单的把文本分为两类: 积极和消极，返回值为情绪的概率，也就是情感评分在 [0,1] 之间，越接近 1，情感表现越积极，越接近 0，情感表现越消极[6]。

2.3.1　模型训练

首先创建正负语料库，本文参考网上提供的汉语商品评论语料库，同时还根据该商品评论的特点进行增加和删减等调整，以增加评论的准确率。通过训练正向和负向情感数据集训练模型，找到外部库 snownlp 中 sentiment 模块，并将训练完的模型进行替换。

2.3.2　情感分析

本文在此基础上先词典化，接下来将文档表示成词袋向量，最后进行 LDA 模型训练。通过进行情感分析，对主题词进行聚类和输出如图 2 所示。

```
[['京东'], ['速度'], ['用后'], ['效果'], ['追评']]
(0, '0.315*"用后" + 0.307*"速度" + 0.158*"追评"')
(array([[0.33955488, 1.3187417 , 0.34170252],
        [1.312477  , 0.34041408, 0.347108  ],
        [1.3161466 , 0.3405596 , 0.34329292],
        [0.33831003, 1.3199264 , 0.3417627 ],
        [0.34722543, 0.33862263, 1.3141512 ]], dtype=float32), None)
(1, '0.330*"效果" + 0.324*"京东" + 0.118*"用后"')
(array([[0.33955294, 1.3187531 , 0.34169313],
        [1.3124701 , 0.34041402, 0.34711504],
        [1.316141  , 0.34056067, 0.34329745],
        [0.33830974, 1.3199288 , 0.34176072],
        [0.34716132, 0.33861846, 1.3142195 ]], dtype=float32), None)
(2, '0.375*"追评" + 0.170*"速度" + 0.155*"用后"')
(array([[0.3395538 , 1.3187485 , 0.34169692],
        [1.3124604 , 0.3404169 , 0.34712178],
        [1.3161464 , 0.34055948, 0.34329322],
        [0.33831012, 1.3199261 , 0.3417629 ],
        [0.34712344, 0.33861622, 1.3142595 ]], dtype=float32), None)
```

图 2　主题词聚类

Fig. 2　Keyword clustering

通过调用 sentimentslist 方法逐行读取评论文本信息，进行情感倾向值计算，统计各情感分数段出现的评论并输出结果。图 3 为情感分析概率图，横坐标表示情感概率，区间为 [0,1]；纵坐标表示频率，每一个区间有 10 条评论量，柱状图表示各情感分数出现的频率。由图 4 中柱状分布可以看出 [0.9,1.0] 之间的分值出现频率更高，可以认为该评论的情感倾向是偏向正向的积极的。而低于 0.1 的分值也占有不小的频率，表明评论当中也有一小部分差评。

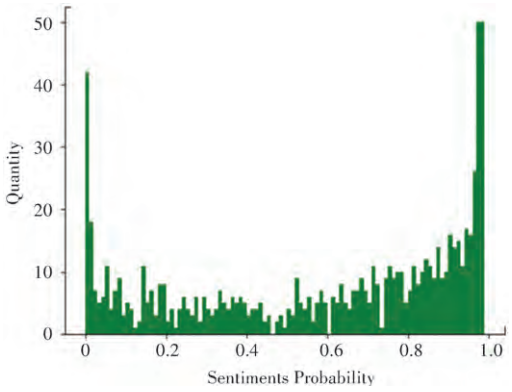

图 3　情感分析概率图

Fig. 3　Sentiment analysis probability graph

通过对情感评分进行逐条分析，部分结果见表 1。将值大于 0.6 设置为喜欢、值小于 0.1 设置为不喜欢，其余值设置为一般。情感分析可视化结果如

图 3 所示，该产品好评数量约为 750，占比 75%；差评数量约 100，占比 10%，一般占比约 15%，可视化结果如图 4 所示，用户总体对于京东平台的此商品的满意率较高，对于该商品评论关注点集中在 3 个方面，首先是产品本身的效果，其次是物流，最后是产品的包装、品牌和价格等因素，生产商可以在这些方面进行提高和改进。

表 1 情感分析结果(部分)

Tab. 1 Results of sentiment Analysis (part)

情感分析结果	倾向	商品评论
0.9220590183915549	likes	非常好用，功效也很好，一直在用。
0.998599087751205	likes	多次购买了！赞一个！希望越做越好！
0.0010048637034076	unlikes	垃圾，买了两天价格相差 47 元。
0.0176742825949318	unlikes	吃了几天，好像肠胃有点难受！
0.2103224101633364	common	最近有点累，买来吃吃看看有没有效果！
0.3347688969555753	common	不知道效果怎么样，用完再来评价。

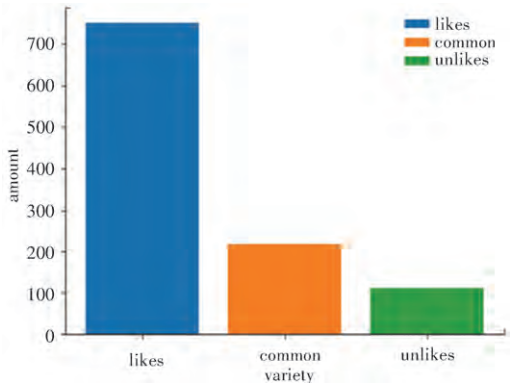

图 4 情感分析可视化结果

Fig. 4 Sentiment analysis visualization results

3 结束语

本文进行情感分析研究的语料库是中文语料库，而在互联网快速发展迭代的今天，不能排除英语、表情符号以及网络新造词出现在评论文本当中，本文没有考虑其它语言、医药术语和表情符号等对情感值的影响。另外，本文只是进行了情感分值的分析，在后续的实验当中，可以构建中医药术语语料库从而进行更为复杂和具体的文本情感强度分析。希望通过此方法拓展的其它领域如对中医药数据进行爬取和情感分析，提升中医药事业发展。

参考文献

[1] 杜晓旭，贾小云. 基于 Python 的新浪微博爬虫分析 [J]. 软件，2019，40(4)：182-185.
[2] 王金峰，彭禹，王明，等. 基于网络爬虫的新浪微博数据抓取技术 [J]. 中小企业管理与科技(上旬刊)，2019(1)：162-163.
[3] 王英杰. 基于 Python 的微博数据爬虫程序设计研究 [J]. 信息与电脑(理论版)，2018(23)：93-94.
[4] 叶春蕾，冷伏海. 基于概率模型的主题识别方法实证研究 [J]. 情报科学，2013，31(2)：135-139.
[5] 郭贵洲，余磊，张寒梅. Python 语言在晕渲制作中的应用 [J]. 地理空间信息，2012，10(4)：159-161，184.
[6] 卢伟胜，郭躬德，陈黎飞. 基于词性标注序列特征提取的微博情感分类 [J]. 计算机应用，2014，34(10)：2869-2873.

(上接第 127 页)

结构分别对原始算法的 CNN 和 RNN 单元进行改进；结合加入正则项后的 CTC 损失函数，实现更快、更准确地识别特殊环境下的车牌；算法设计在降低网络模型复杂度的同时确保了识别的准确性，缩小计算量的同时又引入正则项以尽量避免发生过拟合现象，使用批量归一化算法加速训练过程。大量的实验数据集上的测试结果表明，改进后的 CRNN+CTC 算法在车牌识别率和识别速度上均优于改进前的方法和其它主流方法。然而，本文在训练时仍出现了少量的过拟合现象，算法在单张车牌平均识别时间方面与传统方法相比仍有一定的提升空间，这是后续工作主要方向。

参考文献

[1] 林哲聪. 基于卷积神经网络的车牌识别系统设计和算法实现 [D]. 杭州：浙江工业大学，2018.
[2] 陈丹. 低分辨率车牌识别算法研究 [D]. 西安：西安理工大学，2019.
[3] GRAVES A. Connectionist temporal classification [M] //Supervised Sequence Labelling with Recurrent Neural Networks. Springer，Berlin，Heidelberg，2006，385：61-93.
[4] WANG D，TIAN Y M，GENG W H，et al. LPR-Net：recognizing Chinese license plate in complex environments [J]. Pattern Recognition Letters，2020，130：148-156.