



文献 CSTR:

32001.14.11-6035.csd.2022.0026.zh



文献 DOI:

10.11922/11-6035.csd.2022.0026.zh

数据 DOI:

10.11922/sciencedb.01623

文献分类: 信息科学

收稿日期: 2022-03-25

开放同评: 2022-07-13

录用日期: 2024-12-26

发表日期: 2025-02-24

针对北京旅游景点的评论文本情感标注数据集

朱登赞¹, 万福成^{1*}, 国旗², 张东娇¹, 于洪志¹

1. 西北民族大学语言与文化计算教育部重点实验室, 兰州 730030

2. 西北民族大学甘肃省民族语言文化智能信息处理重点实验室, 兰州 730030

摘要: 在旅游业高速发展背景下, 潜在出行用户想在大量信息中寻找合适的旅游景点不仅耗费时间且缺乏可信度。各类旅游平台的用户评论具有很高的参考价值, 本研究基于携程网提取北京部分旅游景点评论文本, 通过文本爬取、文本预处理、文本情感倾向标注过程, 制作了北京部分旅游景点评论文本情感数据集。实验所爬取旅游景点评论文本数据量为 5 万条。通过与用户在线查看旅游评论对比验证表明, 本数据集具有较高可靠性。本数据集可为做旅游数据文本的情感分析提供数据支撑。

关键词: 旅游景点; 情感分析; 评论文本; Python 爬虫技术

数据库(集)基本信息简介

数据库(集)名称	针对北京旅游景点的评论文本情感标注数据集
数据作者	朱登赞, 万福成, 国旗, 张东娇, 于洪志
数据通信作者	万福成 (wanfucheng@126.com)
数据时间范围	2019–2021
地理区域	北京
数据量	5 万条
数据格式	*.xlsx
数据服务系统网址	http://doi.org/10.11922/sciencedb.01623
基金项目	国家自然科学基金(62366046)、甘肃省基础研究创新群体项目(24JRRA154)
数据库(集)组成	数据集从携程网中提取北京部分旅游景点评论文本, 实验所爬取旅游景点评论文本数据量 5 万条, 以 xlsx 格式保存到数据库中。将积极的评论文本标注为 1, 消极的评论文本标注为 0。

引言

近些年随着互联网的发展与普及, 旅游业发展迅速。相对于其他产业而言, 它具有抽象、高信息量、高宣传的特点, 因此出现了许多旅游出行门户网站^[1-3], 如携程、飞猪等, 促进了旅游业的高速发展。而近年来人们的生活水平不断提高, 旅游度假在人们的生活休闲中必不可少。这些平台的出现, 让用户的出行更加方便快捷^[4], 提高了出去旅行的欲望。通过网络可以实现商家与消费者以及消费者之间的双向交流, 产生了很多关于旅游方面的信息和评论^[5], 这些信息具有透明

* 论文通信作者

万福成: wanfucheng@126.com

化、公开化、真实化的特点，会给有旅游意向的人提供多种选择和多种建议。随着使用人数以及出行次数的增多，相关网站下的用户评论，出行线路，门票信息，季节特色等信息呈现指数级增长^[6]，各类信息的层层堆叠覆盖造成了信息过载，用户在海量未经处理的信息中找到自己感兴趣的目标非常耗费精力，而现在的各种旅游网站并没有提供信息分类和处理的功能，这些信息冗余造成的问题使用户无法难以做出理想的决策^[7]。面对网络上日益丰富的旅游评论信息，如何能快速有效获取并使用其中的有效信息成为人们关注的问题。

网络评论作为很多用户了解各个景点的第一手信息，由大众消费者贡献并且具有高度可信性以及认同感，对用户的旅游决策有着重要的影响。旅游者在出行前通常会通过各种网站来获取目的地信息以及其他旅游者对于该目的地的评价反馈。消费者发布的评论不仅对众多旅游者选择目的地提供参考依据并且有着重要的影响^[8-10]。事实上，人们不仅在外出旅行之前收集各种旅游信息作出决策，在到达目的地之后常常也有对这些信息的需求。旅游文本评论在本质上属于用户生成内容之一，这些评论文本真实反映用户的旅游体验、感受，体现了用户对项目的褒贬态度和情感倾向^[11]。旅游行业的日益发展壮大使人们对出行目的地有了更多选择，因此，充分理解旅游者的所知所想所感对行业发展就更加重要。

网络文本评论已成为旅游景区的重要评价标准，不仅是用户真情实感的真实反馈，同时旅游评价文本分析也能给景区提供有效意见，提高游客的满意度。旅游评价相比于普通的商品评价，通常字数较多，并且包含多个评价方面以及评价层次。针对这些网络评论文本进行深入分析有十分重要的意义^[12]。通过网络爬虫工具获取平台的所有评论文本信息^[13]，对这些文本评论进行深入挖掘，从而获取用户对于某个地方的产品及服务的感受，智能分析用户对这个地点的情感倾向，可对旅游者聚焦点、情感倾向等方面有一个更加深刻的认识。让旅游者可以从处理过的评论文本信息中快速准确地找到自己需要的有效信息^[14]，确定出行的目的地，制定出行计划，避免浪费时间。旅游景点评论文本情感数据集的存在必不可少。

1 数据采集和处理方法

1.1 数据来源

实验数据集是从携程网中提取北京部分旅游景点评论文本。爬虫获取旅游景点评论文本数据。携程作为旅游服务平台，用户评价是其重要业务部分。为维护自身信誉和竞争力，平台鼓励用户真实评价，因此本研究数据源具备一定客观性基础。而且携程用户众多，不同用户基于真实体验评价旅游景点，负面评论能真实呈现，最大程度保证了数据的原始性与真实性。

1.2 数据爬取

第一阶段：获取种子 URL 地址，提取网站源码中的种子 URL 加入待爬取的任务列表中，例如本研究爬取的携程网页地址。之后自动获取符合条件的新地址加入 URL 队列中，重复爬取的过程。

第二阶段是获取目标 URL 后，模拟浏览器对服务器发送请求，如果得到服务器正常响应，可以下载相应页面获得响应内容。

第三阶段是获取解析内容，网页内容包含大量无效冗余的数据，采用正则表达式解析网页旅游景点评论文本数据实现对网站有效数据的提取，获得目标数据。本实验所提取的文本信息属性包括景点名、评论文本、评论条数、评论时间等字段。

第四阶段是数据存储。将采集到的信息存储到数据库中，保证数据的持久化。本实验爬取旅游景点评论文本数据量为 5 万条，因此保存成 xlsx 格式。网络爬虫流程如图 1 所示。

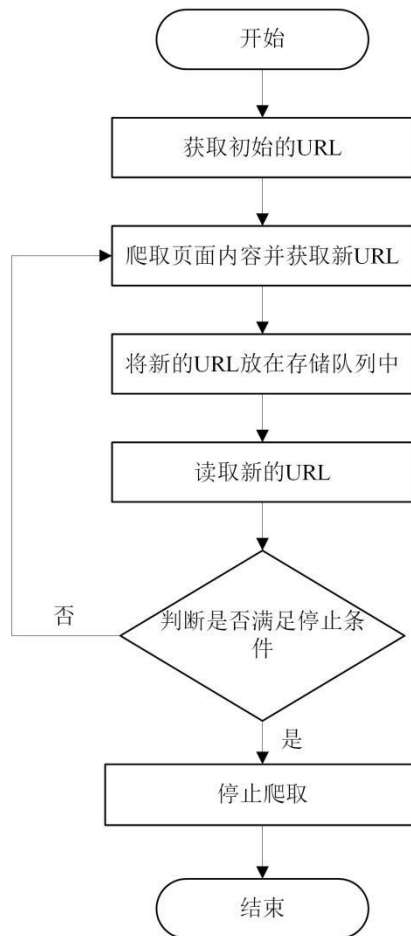


图 1 网络爬虫过程图

Figure 1 The web crawler process

1.3 文本预处理

本数据集是在携程网上爬取的数据，具有较多的噪声，需要进行结构化处理才能使用。文本预处理主要分为文本清洗、文本分词和停用词过滤三部分，通过文本预处理将数据集中无效内容进行过滤，从而保证最终结果准确率。文本预处理过程如图 2。

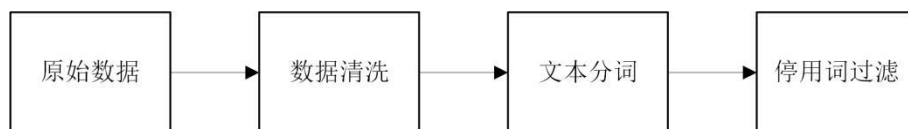


图 2 文本预处理过程

Figure 2 Text pre-processing process

1.3.1 文本清洗

旅游景点评论文本是游客根据自身经历真实表达的，没有统一的标准。从携程网爬取的旅游景点评论文本，包含各种冗余信息，如广告词句、特殊字符、多余的空格等，会占用一定的存储空间。经过文本清洗后，从携程网爬取的数据文件大小有所降低，为 4.36 MB。在预处理的文本清洗阶段，

主要是通过正则匹配过滤掉网址超链接、俚语和特殊符号等无意义的文本数据。俚语通常是特定地区、群体或行业中使用的非正式语言表达，通过分析其上下文来进行判断。特殊符号通常包括标点符号以外的各种符号，如表情符号、图形符号、乱码等，可以使用正则表达式来匹配它们的特定形式进行识别。这种嘈杂的数据会使从文本中提取主题变得更加困难，也会影响最终结果。

例如旅游景点评论文本：“北京海洋馆 <https://www.bj-sea.com> 绝绝子，内容丰富，活动精彩，孩子喜欢(*^▽^*)”。在文本清洗过程中，可以进行以下处理：

对于“绝绝子”这个俚语，可以根据具体情况决定是否保留或进行适当的转换。如果认为它对情感倾向的判断有较大影响，可以保留；如果觉得它比较模糊或难以理解，可以考虑删除或替换为更正式的表达。

对于表情符号“(*^▽^*)”和超链接，可以将其视为无意义文本进行过滤。

经过清洗后变为“北京海洋馆内容丰富，活动精彩，孩子喜欢”。去除冗余数据可以提高文本数据质量，对评论文本进行噪声处理是非常重要的。

1.3.2 文本分词

文本分词是将评论文本中的每条评论分割成词语。对于英文而言分词比较容易，因为英文单词间由空格分隔，而中文文本分词相对较难，中文句子没有明显的分割符号，不能直观进行分词，需要借助结巴分词、哈工大分词等分词工具进行分词操作。本研究采用的是结巴分词 <https://github.com/fxsjy/jieba>。为了避免分词后实验数据集对实验的影响，针对旅游景点评论文本的具体内容增加了个性化的用户词典，个性化词典包含不能被分词算法分开的重要词。

1.3.3 停用词过滤

停用词过滤就是去掉句子中不必要的单词，去掉以后对理解整个句子的语义没有影响。例如“你”、“我”、“他”、“的”等单词，出现频率较高、没有具体的含义，还会消耗大量不必要的训练时间，因此常采用停用词表进行去停用词处理。为使旅游景点评论文本主题提取效果更好，本研究根据所处理数据集的实际内容来对停用词表进行个性化定制。结合哈工大、百度通用停用词表作为初步的停用词候选列表，在此基础上进行个性化的调整。对数据集中的文本进行初步分析，统计每个词汇的出现频率，将出现频率高且对文本分析意义不大的词汇加入停用词候选列表中。不同领域的停用词表，会对文本预处理后的结果产生一定影响，它决定了在文本挖掘的过程中需要使用的数据，以及最终呈现的效果。例如：“哇塞！北京后花园的（白虎涧）风景区还不错嘞，人挺少的，景色优美，好玩呢”处理后变为“北京后花园（白虎涧）风景区还不错，人挺少，景色优美，好玩”。

2 数据样本描述

2.1 数据组成

本数据针对北京部分旅游景点的游客评论进行收集与提取。经过文本爬取、文本预处理、文本情感倾向标注过程，数据集可以清晰明了地展现用户对该景点的喜好与评价，方便其他潜在游客的对于信息的接收。其中包括游客对于票价、人数、规格场地、当地特色等全方位的评价，将这些评论按照游客的情感倾向进行数据标注，给游客提供一些有用的建议和选择，能够制定更好的计划。

2.2 样本数据描述

经过文本预处理之后的旅游景点评论数据共有 5 万条。景点风景历史文化的评论文本比例是 56.9%，针对旅游体验感、性价比的评论文本比例是 25.2%，针对交通、天气、时间预约等其他评论的比例是 17.9%。

对预处理后的数据集进行人工标注，具体过程：挑选语言学专业学生进行人工标注，将预处理后的评论文本随机分配，确保每个文本至少由两人独立标注。根据文本的内容、语气、用词等判断其情感倾向，将文本标注为积极或消极。对于存在争议的文本，组织人员进行讨论，共同确定其情感倾向。数据集中将评论文本情感倾向为积极的文本标注为 1，评论文本情感倾向为消极的文本标注为 0。如表 1 和表 2 所示。

表 1 旅游景点积极评论文本数据

Table 1 Positive review texts about tourist attractions

评论	情感
天坛很壮观，惊叹我们大中国的建筑古代人的智慧	1
北海公园秋季的北海，太美了！刚好赶上菊花展	1
天坛性价比高，总体超赞，	1
故宫雨中的故宫也别有一番风味，在李秀才的带领下，江山美景，历史古韵，尽收眼底。	1
北京海洋馆总体超赞，有趣好玩，景色不错，性价比高，	1

表 2 旅游景点消极评论文本数据

Table 2 Negative reviews of text data about tourist attractions

评论	情感
簋街小吃辣的太多，不太适合我的口味，一般般。	0
北京动物园网上购票都没有告诉海洋馆几点闭馆 性价比低，有待改进，	0
北京海洋馆感觉很不值，海洋馆照相还很贵，表演就二十分钟，很差感觉不值。	0
北京动物园没啥动物，不过门票也便宜。海洋馆去多了也就那样。没啥独特的地方。总之去过一次就行了。	0
八达岭野生动物世界无聊没有动物在车外边	0

3 数据质量控制和评估

本数据集的质量取决于网站信息和游客评论的真实性、可靠性，数据采集和数据处理的科学性。所有数据的采集是通过在携程网上爬取北京部分旅游景点的用户评论获取的，并通过数据爬取、文本预处理等操作来进行有效数据的提取。

游客的不同情感倾向的评论给用户提供了真实可靠的信息，以评论文本作为参考，可以为游客提供较好的旅游路线，更好的旅游体验让用户尽可能选出适合自己旅游的最优解；同时也为旅游景区提供了游客的情感倾向与主要感受，能够更精准地改进，提供更好的服务。

4 数据使用方法及建议

网络上的旅游评论海量且碎片化，不利于用户快速直接找到自己需要的评论文本和情感倾向。然而，本数据集是根据携程网中的点评内容也就是订单完成之后用户的真实反映，从网上爬取这些真实的信息进行汇总分析，直观地反映了游客对于各个旅游景点的情感倾向，有助于加深潜在游客对于该景点的了解，制定更好的旅游规划。

本数据集主要是针对北京部分旅游景点的游客评论，主要适用于近期有在北京地区进行旅游计划的用户，可以快速地从海量旅游评论中获得旅游者对于旅游过程中某一方面的正负面评价，能更有效地帮助用户进行决策，同时可以了解什么样的评论对用户帮助最大，尽可能给予用户更多的有用信息，让用户有一个完满的旅行体验。本数据集为 xlsx 格式，可直接用 excel 查看分析数据，用户可以根据实际情况下载研究数据。

致 谢

感谢携程网对本研究获取北京部分旅游景点评论文本数据支持，感谢西北民族大学语言与文化计算教育部重点实验室对本研究的支持。

数据作者分工职责

朱登赟（1983—），女，河南省开封市人，博士研究生，研究方向为自然语言处理。主要承担工作：数据抓取，数据清洗，LDA 模型构建。

万福成（1985—），男，辽宁省台安市人，教授，博士生导师，研究方向为自然语言处理、信息抽取、自动问答技术。主要承担工作：总体质量管控，论文指导与修改。

国旗（1998—），女，黑龙江省哈尔滨市人，硕士研究生，研究方向为信息抽取。主要承担工作：数据质量控制，论文撰写。

张东娇（1996—），女，黑龙江省哈尔滨市人，硕士研究生，研究方向为自然语言处理。主要承担工作：数据集的预处理和整合。

于洪志（1947—），女，山东省龙口市，教授人，博士生导师，研究方向为自然语言处理、人工智能、互联网+多语言信息技术。主要承担工作：数据审核。

参考文献

- [1] 郭庆, 贾郭军. 基于 tree LSTM 的旅游评论情感分析方法研究[J]. 计算机应用研究, 2020, 37(S2): 63 - 65. [GUO Q, JIA G J. Research on sentiment analysis method of travel reviews based on tree LSTM[J]. Application Research of Computers, 2020, 37(S2): 63-65.]
- [2] 何雪琴, 杨文忠, 吾守尔·斯拉木, 等. 融合句法规则和 CNN 的旅游评论情感分析[J]. 计算机工程与设计, 2019, 40(11): 3306 - 3312. DOI:10.16208/j.issn1000-7024.2019.11.041. [HE X Q, YANG W Z, Wushouer ·Silamu, et al. Sentiment analysis of tourist reviews combined with syntactic rules and CNN[J]. Computer Engineering and Design, 2019, 40(11): 3306 - 3312. DOI:10.16208/j.issn1000-7024.2019.11.041.]

- [3] 夏震. 基于 Bi-LSTM 的旅游评价情感分析模型[J]. 自动化应用, 2020(5): 129 - 131. DOI:10.19769/j.zdhy.2020.05.053. [XIA Z. Bi-LSTM-based sentiment analysis model for tourism evaluation[J]. Automation Application, 2020(5): 129 - 131. DOI:10.19769/j.zdhy.2020.05.053.]
- [4] 周道华, 古鹏飞, 曾俊. 一种旅游网络评论情感分析及服务质量评价方法: CN109034893A[P]. 2018-12-18. [ZHOU D H, GU P F, ZENG J. A sentiment analysis and service quality evaluation method for tourism network reviews: CN109034893A[P]. 2018-12-18.]
- [5] 覃国蓉, 叶志成, 庄槟豪, 等. 旅游网络评论情感分析方法研究及系统实现[J]. 深圳信息职业技术学院学报, 2015, 13(3): 57-62. DOI:10.3969/j.issn.1672-6332.2015.03.012. [QIN G R, YE Z C, ZHUANG B H, et al. Research of tourism network review sentiment analysis method and system implementation[J]. Journal of Shenzhen Institute of Information Technology, 2015, 13(3): 57 - 62. DOI:10.3969/j.issn.1672-6332.2015.03.012.]
- [6] 曹斌. 互联网上旅游评论的情感分析及其有用性研究[D]. 哈尔滨工业大学, 2012.[CAO B. Sentiment analysis and its usefulness of travel reviews on the Internet[D]. Harbin Institute of Technology, 2012.]
- [7] 拓凯渊. 基于文本挖掘的线上旅游评论分析[J]. 西部皮革, 2020, 42(2): 121 - 123. [TUO K Y. Analysis of online travel reviews based on text mining[J]. West Leather, 2020, 42(2): 121-123.]
- [8] 李炜. 基于情感词典和神经网络的中文旅游评论分析[D]. 中国科学院大学, 2018. [LI W. Analysis of Chinese tourism reviews based on emotion dictionary and neural network[D]. University of Chinese Academy of Sciences, 2018.]
- [9] 郑文英. 旅行目的地中文评论的情感分析研究[D]. 哈尔滨工业大学, 2010. [ZHENG W Y. Sentiment analysis of Chinese reviews of travel destinations[D]. Harbin Institute of Technology, 2010.]
- [10] 杨艳霞. 基于本体的旅游网络评论情感分析与预警系统[J]. 计算机与数字工程, 2016, 44(4): 649-652. DOI:10.3969/j.issn.1672-9722.2016.04.020. [YANG Y X. Tourism network comments sentiment analysis and pre-warning system based on ontology[J]. Computer & Digital Engineering, 2016, 44(4): 649 - 652. DOI:10.3969/j.issn.1672-9722.2016.04.020.]
- [11] FORIS D, CRIHALMEAN N, FORIS T. Exploring the environmental practices in hospitality through booking websites and online tourist reviews[J]. Sustainability, 2020, 12(24): 10282. DOI:10.3390/su122410282.
- [12] CHEN L J, KENG G. Feature expansion using lexical ontology for opinion type detection in tourism reviews domain[J]. International Journal of Advanced Computer Science and Applications, 2020, 11(8): 257 - 269. DOI:10.14569/ijacsa.2020.0110877.
- [13] AFZAAL M, USMAN M, FONG A. Predictive aspect-based sentiment classification of online tourist reviews[J]. Journal of Information Science, 2019.
- [14] FANG B, YE Q, KUCUKUSTA D, et al. Analysis of the perceived value of online tourism reviews: influence of readability and reviewer characteristics[J]. Tourism Management, 2016, 52: 498-506. DOI:10.1016/j.tourman.2015.07.018.

论文引用格式

朱登赞, 万福成, 国旗, 等. 针对北京旅游景点的评论文本情感标注数据集[J/OL]. 中国科学数据, 2025, 10(1). (2025-02-24). DOI: 10.11922/11-6035.csd.2022.0026.zh.

数据引用格式

语言与文化计算教育部重点实验室, 西北民族大学, 针对北京旅游景点的评论文本情感标注数据集 [DS/OL]. V2. Science Data Bank, 2022. (2025-02-24). DOI: 10.11922/sciencedb.01623.

附 录

数据来源 URL:

<https://you.ctrip.com/sight/beijing1/2045497.html>
<https://you.ctrip.com/sight/beijing1/229.html>
<https://you.ctrip.com/sight/beijing1/230.html>
<https://you.ctrip.com/sight/beijing1/62722.html>
<https://you.ctrip.com/sight/beijing1/231.html>
<https://you.ctrip.com/sight/beijing1/5174.html>
<https://you.ctrip.com/sight/beijing1/233.html>
<https://you.ctrip.com/sight/beijing1/5208.html>
<https://you.ctrip.com/sight/beijing1/243.html>
<https://you.ctrip.com/sight/beijing1/5173.html>
<https://you.ctrip.com/sight/beijing1/107469.html>
<https://you.ctrip.com/sight/beijing1/228.html>
<https://you.ctrip.com/sight/beijing1/5153.html>
<https://you.ctrip.com/sight/beijing1/52626.html>
<https://you.ctrip.com/sight/beijing1/234.html>
<https://you.ctrip.com/sight/beijing1/5181.html>
<https://you.ctrip.com/sight/beijing1/5170.html>
<https://you.ctrip.com/sight/beijing1/235.html>
<https://you.ctrip.com/sight/beijing1/232.html>
<https://you.ctrip.com/sight/beijing1/64955.html>
<https://you.ctrip.com/sight/beijing1/107621.html>
<https://you.ctrip.com/sight/beijing1/52627.html>
<https://you.ctrip.com/sight/beijing1/5265.html>
<https://you.ctrip.com/sight/beijing1/107634.html>
<https://you.ctrip.com/sight/beijing1/54198.html>
<https://you.ctrip.com/sight/beijing1/107837.html>

<https://you.ctrip.com/sight/beijing1/107846.html>
<https://you.ctrip.com/sight/beijing1/5315.html>
<https://you.ctrip.com/sight/beijing1/65081.html>
<https://you.ctrip.com/sight/beijing1/54291.html>
<https://you.ctrip.com/sight/beijing1/2476712.html>
<https://you.ctrip.com/sight/beijing1/142572.html>
<https://you.ctrip.com/sight/beijing1/62704.html>
<https://you.ctrip.com/sight/beijing1/54196.html>
<https://you.ctrip.com/sight/beijing1/2476712.html>
<https://you.ctrip.com/sight/beijing1/140217.html>
<https://you.ctrip.com/sight/beijing1/67588.html>
<https://you.ctrip.com/sight/beijing1/242.html>

A dataset of sentiment annotation the review texts for Beijing tourist attractions

ZHU Dengyun¹, WAN Fucheng^{1*}, GUO Qi², ZHANG Dongjiao¹, YU Hongzhi¹

1. Northwest Minzu University Key Laboratory of Linguistic and Cultural Computing Ministry of Education, Lanzhou 730030, P.R.China

2. Northwest Minzu University Key Laboratory of China's Ethnic Languages and Intelligent Processing of Gansu Province, Lanzhou 730030, P.R. China

*Email: wanfucheng@126.com

Abstract: With the rapid development of tourism industry, it is time-consuming and unreliable for potential travel users to sift through vast amounts of information to find suitable tourist attractions, often with little trust in the sources. User reviews on various tourism platforms are of great reference value. Based on Ctrip, this study extracted review texts for some tourist attractions in Beijing, and produced a sentiment dataset through the process of text crawling, text preprocessing and text sentiment annotation. The dataset contains 50k samples. A comparison with online tourism reviews shows that this dataset is of high reliability and feasibility. The dataset can provide data support for sentiment analysis of tourism data texts.

Keywords: tourist attraction; emotion analysis; review text; Python crawler

Dataset Profile

Title	A dataset of sentiment annotation the review texts for Beijing tourist attractions
Data authors	ZHU Dengyun, WAN Fucheng, GUO Qi, ZHANG Dongjiao, YU Hongzhi
Data corresponding author	WAN Fucheng (wanfucheng@126.com)
Time range	2019–2021
Geographical scope	Beijing
Data volume	50,000 items

Data format	*.xlsx
Data service system	< http://doi.org/10.11922/sciencedb.01623 >
Source(s) of funding	The National Natural Science Foundation of China (62366046); the Basic Research Innovation Group Project of Gansu Province(24JRRA154)
Dataset composition	The dataset extracted review texts from some tourist attractions in Beijing on Ctrip, with a total of 50,000 crawled and saved in the database in xlsx format. Positive review texts are labeled as 1 and negative comment texts are labeled as 0.