关于文本分类(情感分析)的中文数据集汇

总 原创 樱与刀 ① 于 2019-07-14 16:08:14 发布 ② 41213 🏠 收藏 414 版权 分类专栏: NLP 文章标签: NLP dataset

NLP 专栏收录该内容 4 订阅 6 篇文章 订阅专栏

文本分类 (情感分析)中文数据集汇总

这段时间在公司NLP组里实习,相应的开始学习了一些NLP的知识,并搜索了 一些关于NLP中文本分类领域的相关数据集,本文主要列举一些中文数据集。 关于英语数据集,且听下回分解。

1.THUCNews 数据集 :

THUCNews是根据新浪新闻RSS订阅频道2005~2011年间的历史数据筛选过滤 生成,包含74万篇新闻文档(2.19 GB),均为UTF-8纯文本格式。我们在原始 新浪新闻分类体系的基础上,重新整合划分出14个候选分类类别:财经、彩 票、房产、股票、家居、教育、科技、社会、时尚、时政、体育、星座、游 戏、娱乐。使用THUCTC工具包在此数据集上进行评测,准确率可以达到 88.6%。

数据集地址: http://thuctc.thunlp.org/

数据规模:共382688条,分布于15个分类中。

2.今日头条新闻文本分类数据集:

数据来源: 今日头条客户端

数据格式: 6552431613437805063_!_102_!_news_entertainment_!_谢娜为李 浩菲澄清网络谣言,之后她的两个行为给自己加分_!_佟丽娅,网络谣言,快乐大本营,李浩 菲,谢娜,观众们

每行为一条数据,以_!_分割的个字段,从前往后分别是 新闻ID,分类code(见下文),分类名称(见下文),新闻字符串(仅含 标题),新闻关键词

数据集地址: https://github.com/fate233/toutiao-text-classfication-dataset

3.全网新闻数据(SogouCA):

来自若干新闻站点2012年6月—7月期间国内,国际,体育,社会,娱乐等18个 频道的新闻数据,提供URL和正文信息

数据格式为

1 <doc> 2 <url>页面URL</url> 3 <docno>页面ID</docno> 4 <contenttitle>页面标题</contenttitle> 5 <content>页面内容</content> 6 </doc>

注意:content字段去除了HTML标签,保存的是新闻正文文本

数据集地址: https://www.sogou.com/labs/resource/ca.php

4.搜狐新闻数据(SogouCS):

来自搜狐新闻2012年6月—7月期间国内,国际,体育,社会,娱乐等18个频道 的新闻数据,提供URL和正文信息

1 数据格式为 2 <doc> 3 ∣ <url>页面URL</url> 4 <docno>页面ID</docno> 5 <contenttitle>页面标题</contenttitle> 6 <content>页面内容</content> 7 </doc>

注意:content字段去除了HTML标签,保存的是新闻正文文本

数据集地址: https://www.sogou.com/labs/resource/cs.php

5.ChnSentiCorp_htl_all数据集:

7000 多条酒店评论数据,5000 多条正向评论,2000 多条负向评论

1 数据字段:

2 Label: 1表示正向评论, 0表示负向评论 3 Review: 评论内容

房间小得无法想象,建议个子大的不要选择,一般的睡觉脚也伸不直.房间不超过10平方,彩电是14. 7321 0 我们一家人带孩子去过"五.一",在协程网上挑了半天才选中的酒店,但看来还是错了。1.酒店除了... 周六到西山去采橘子,路过这家酒店的时候就觉得应该不错的,采好橘子回来天也晚了,就临时决定住在... 交通很便利,到渔人码头和港澳码头都在步行的范围之内。CHECKIN和CHECKOUT的速度都... 很不错的一个酒店,床很大,很舒服.酒店员工的服务态度很亲切.

数据集地址:

https://raw.githubusercontent.com/SophonPlus/ChineseNlpCorpus/master/dat asets/ChnSentiCorp_htl_all/ChnSentiCorp_htl_all.csv

6.waimai_10k数据集:

某外卖平台收集的用户评价,正向4000条,负向约8000条

1 数据字段: 3 Label: 1表示正向评论, 0表示负向评论 4 Review: 评论内容

label review 25 1 送餐特别快,态度也好,辛苦啦 6632 0 点了热带雨林披萨+饮料,和BBQ鸡肉披萨+饮料,送来的是两个奥尔奥披萨+两个银耳冰粥,冰凉冰... 8849 0 难吃川油死了,味道烂 11114 0 今天菜太咸,连着定了3天吃,一天比一天难吃。 11661 0 送的太慢了,菜都凉了。 9571 0 没有满减! 10614 0 差评! 定的时间是12点一刻,结果刚11点就送来了! 果断退单。送餐前不看时间吗? 7585 0 羊肉串太咸,还有些不新鲜。鸡心和鸡胗烤的太老 6919 0 仲議吊紐好 速度狂体

数据集地址:

https://raw.githubusercontent.com/SophonPlus/ChineseNlpCorpus/master/dat asets/waimai_10k/waimai_10k.csv

7.online_shopping_10_cats数据集:

10 个类别(书籍、平板、手机、水果、洗发水、热水器、蒙牛、衣服、计算 机、酒店),共6万多条评论数据,正、负向评论各约3万条

字段说明 字段 说明 cat 类别:包括书籍、平板、手机、水果、洗发水、热水器、蒙牛、衣服、计算机、酒店 label 1 表示正向评论 , 0 表示负向评论 review 评论内容

cat label review 11194 平板 0 什么玩意。刚用一天,就充不上电,开不开机,返厂老麻烦, 17794 水果 1 买了几次了,价格实惠,口感不错,保鲜好! 29529 洗发水 1 挺值得购买的,有包装买回去送家人,毛巾质量不错。小块的可以拿来当擦手帕。 24976 水果 0 真的就算后悔了。两天才拿到货。还不如水果店买!还都发霉不新鲜了!以后不买了 28447 洗发水 1 一般般,薄荷洗发水没想象中的凉快 264 书籍 1 这本书有别于以往看过的早款书籍,结合了说明文的写实,散文的情致和图册的一目了然。特别是读过几...

数据集下载地址:

https://github.com/SophonPlus/ChineseNlpCorpus/raw/master/datasets/online _shopping_10_cats/online_shopping_10_cats.zip

8.weibo_senti_100k数据集:

10 万多条,带情感标注 新浪微博,正负向评论约各 5 万条。 字段说明

字段 说明 1表示正向评论,0表示负向评论 review 微博内容

62050 0 太过分了@Rexzhenghao //@Janie_Zhang:招行最近负面新闻越来越多呀... 68263 0 希望你?得好?我本 " ?肥血?史 " [晕][哈哈]@Pete三姑父 81472 0 有点想参加????[偷?]想安排下时间再决定[抓狂]//@黑晶晶crystal: @细腿大羽... **42021** 1 [给力]感谢所有支持雯婕的芝麻![爱你] 7777 1 2013最后一天,在新加坡开心度过,向所有的朋友们问声:新年快乐!2014年,我们会更好[调...

数据集下载地址:

https://github.com/SophonPlus/ChineseNlpCorpus/blob/master/datasets/weibo _senti_100k/intro.ipynb

9.simplifyweibo_4_moods数据集:

36 万多条,带情感标注 新浪微博,包含 4 种情感,其中喜悦约 20 万条,愤 怒、厌恶、低落各约5万条

字段说明

label 0 喜悦 , 1 愤怒 , 2 厌恶 , 3 低落 review 微博内容

回复美国看起来很美,对别人比较狠!对付哪国人,就用哪国人做他的腿,简称狗腿落后的祖宗挨过打! 258815 2 我表示压力狠大。1哇。犀利妹!偶尔街拍,其实姐只是一个你永远无法超越的传说。 可怜,帮这孩子转下,希望不会因为涉嫌联系,不过不会更好的,就是想别丝想疯了什么情况啊?想… 哦也~~~!得瑟哈哈哈耶~~~!新logo。 山 ,们的logo 会不会抢了的... 351395 3 我发现真的是最齐全的一张。这是去看北方儿子的时候啊。怀念。对了,我怎么穿那件破衬衫。。好难看...

举报

数据集下载地址:

10.dmsc_v2数据集:

https://pan.baidu.com/s/16c93E5x373 o zyWevITg

★ 72 📭 🏡 414 🖹 📮 8 【 专栏目录

https://blog.csdn.net/alip39/article/details/95891321

2022/11/27 16:12 (65条消息) 关于文本分类(情感分析)的中文数据集汇总_樱与刀的博客-CSDN博客_中文情感分析数据集

28部电影,超70万用户,超200万条评分/评论数据

字段说明



| | movield | title | title_cn |
|---|---------|---------------------------------|----------|
|) | 0 | Avengers Age of Ultron | 复仇者联盟2 |
| 1 | 1 | Big Fish and Begonia | 大鱼海棠 |
| 2 | 2 | Captain America Civil War | 美国队长3 |
| 3 | 3 | Chinese Zodiac | 十二生肖 |
| 4 | 4 | Chronicles of the Ghostly Tribe | 九层妖塔 |
| | | | |

数据集地址: https://pan.baidu.com/s/1c0yn3TlkzHYTdEBz3T5arA

原始数据集地址:

https://www.kaggle.com/utmhikari/doubanmovieshortcomments

11.yf_dianping数据集:

24 万家餐馆,54 万用户,440 万条评论/评分数据 字段说明

| 字段 | 说明 |
|--------|------------------|
| restId | 餐馆 id (从 0 开始,连续 |
| name | 餐馆名称 |
| | |

| | restld | name |
|--------|--------|-------------|
| 210902 | 210902 | NaN |
| 124832 | 124832 | NaN |
| 26766 | 26766 | 香锅制造(新苏天地店) |
| 91754 | 91754 | NaN |
| | | |

-----数据集地址: https://pan.baidu.com/s/1yMNvHLl6QYsGbjT7u51Nfg 原始数据集地址: http://yongfeng.me/dataset/

12.yf_amazon数据集:

52 万件商品,1100 多个类目,142 万用户,720 万条评论/评分数据 字段说明

| 1 | | |
|---|-----------|---------------------|
| | 宇段 | 说明 |
| | productid | 产品 id (从 0 开始,连续编号) |
| | nama | 辛日 夕和 |

catids 类别 id (从 0 开始,连续编号,从左到右依次表示一级类目、二级类目、三级类目)

| | productid | name | catids |
|--------|-----------|-------------------------------------|--------------|
| 331420 | 331420 | 欧意金狐狸 女式 皮手套 QT602 | 802,143,996 |
| 130945 | 130945 | YESO TOT 中性 单肩包/斜挎包 均码 9411 | 1111,864,781 |
| 179886 | 179886 | 李斯特论柏辽兹与舒曼 | 832,552,337 |
| 504123 | 504123 | Tuscarora 途斯卡洛拉 中性 烈焰驰骋无缝头巾 PSU3083 | 1111,522,720 |
| 387785 | 387785 | 我们的故事:一百个北大荒老知青的人生形态 | 832,519,599 |

原始数据集地址: http://yongfeng.me/dataset/

数据集地址: https://pan.baidu.com/s/1SbfpZb5cm-g2LmnYV_af8Q

13.Datahub数据中心:

包含文本分类、情感分析以及知识图谱的数据集

相关地址: http://www.datahub.ileadall42.com/data/list? category=2&parent_category=1

14.知乎看山杯数据集:

数据集下载地址: https://pan.baidu.com/s/1qUr6IQQn6DzrMlbaAUZsIQ 提取码: qbiw

15.Al_challenger情感分析数据集:

态度"、"排队等候时间"等细粒度要素。

数据集分为训练、验证、测试A与测试B四部分。数据集中的评价对象按照粒度 不同划分为两个层次,层次一为粗粒度的评价对象,例如评论文本中涉及的服 务、位置等要素;层次二为细粒度的情感对象,例如"服务"属性中的"服务人员

数据集下载地址:

https://github.com/nju161250102/AI_challenger/tree/master/data

16.复旦中文文本分类语料库

数据链接:https://pan.baidu.com/s/1833mT2rhL6gBMIM0KnmyKg 密码: zyxa

情感分析数据集(正面10000条,负面5000条) 吸收了 谭松波 非平衡酒店评论语料库(7000条正面,3000条负面,有部分重复数据)...

ChnSentiCorp中文情感分析数据集 04-09

ChnSentiCorp 是一个中文情感分析数据集,包含酒店、笔记本电脑和书籍的网购评论。

评论 8条> 写评论 张某文的博客_Lambda 热评 谢谢你 🥳

中文情感分析数据_BrainEditor的博客_情感分析数据

中文情感词汇本体可以用于解决多类别情感分类的问题,同时也可以用于解决一般的倾向... 基于Bert的中文情感分析代码及分析_SurverDevin的博客_bert...

基于Bert的中文情感分析代码及分析 众所周知,Google 前段时间开源了他们的 BERT – B... Sentiment Analysis 中文情感分析-数据集

该<mark>数据集</mark>包括了四个领域的<mark>中文</mark>评论:笔记本电脑、汽车、相机和手机,评论属于 二分... 情感分析及数据集 最新发布 yingzi的技术博客 💿 282

由于情感可以被分类为离散的极性或尺度(例如,积极的和消极的),我们可以将<mark>情感</mark>...

...SIMS:中文多模态情感分析数据集与细粒度的模态注释_Yang SiChe... 11-21 CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotat...

中文文本情感分类及<mark>情感分析</mark>资源大全_Paulzhao6518的博客 摘要:20世纪初以来,文本的<mark>情感分析</mark>在自然语言处理领域成为了研究的热点,吸引了众多...

中文文本分类项目数据集.rar 08-06

Bert中文文本分类全网最详细教程+源码+数据集 热门推荐 clnjust的博客 © 1万+

"复旦大学计算机信息与技术国际数据库中心自然语言处理小组"提供的小样本中文文本分...

概述 Bert是当前最先进的自然语言处理工具之一。最起码要会使用这个强大的工具。 为... 中文情感分析语料库_noter16的博客_nlp情感分析语料库

中文情感分析的语料库非常少,这五个中文语料库是我在网上的搜集的。 数据集1:情感挖... 中文情感分析任务数据集

1 NLPCC 1.1 NLPCC-2012 下载网址: http://tcci.ccf.org.cn/conference/2012/pages/pag... 文本分类数据集汇总 Dataset Classes Type Samples Best Method Performance AG News 4 Topic Train:1200...

可用于文本分类的中文数据集 m0_50553633的博客 ◎ 812

可用于文本分类的中文数据集(已经标注好类) 博客的连接: https://blog.csdn.net/alip39/... 机器学习之文本分类(附带训练集+数据集+所有代码) 天才幻想家 💿 9183

我本次对4类文本进行分类((所有截图代码和<mark>数据集</mark>最后附带免费下载地址)) 主要... 情感分析常见数据集以及任务进展 xxr233的博客 🧿 3168

情感分析 原文: https://github.com/sebastianruder/NLP-progress/blob/master/english/se... 机器学习-教育客服相关场景下的50组中文普通话对话文本语料-机器... 08-12 机器学习-教育客服相关场景下的50组中文普通话对话文本语料-机器学习数据处理必看-...

中文文本情感分析案例和数据集 07-08 中文文本情感分析案例 课程配套程序 该文件夹是本次课程配套的程序,打来即可使用 ...

ocr中文数据集_文本分类(情感分析)中文数据...weixin_39874366的博客 ◎ 1565

来自我的博客: https://blog.csdn.net/alip39blog.csdn.net1.THUCNews数据集: THUCN...

最新最全自然语言处理之情感分析常用数据集 Tiger数学之家的博客 © 1万+ 本文介绍情感分析领域最常见的一些数据集。 目录 TOC {:toc} SemEval SemEval-2014 ...

NLP实战 -中文文本分类问题: THUCNews数据集 qq_36047533的博客 ◎ 6088 中文文本分类问题: THUCNews数据集1 THUCNews数据集与预处理1.1 数据集下载1.2 ...

中文文本情感分类实战(weibo_senti_100k为... 千万利器,莫过于你的信念 💿 2156

中文文本情感分类数据准备 数据准备 使用jieba分词 import jieba data_path = "sources/... 文本分类数据集_PU-Learning/文本分类/文本聚... weixin_39797780的博客 ◎ 969

本文为PU-Learning/文本分类/文本聚类/情感分析相关研究提供部分常用数据集下载地址.. Sentiment Analysis 中文情感分析-数据集

该<mark>数据集</mark>包括了四个领域的<mark>中文</mark>评论:笔记本电脑、汽车、相机和手机,评论属于 二分...

"相关推荐"对你有帮助么?

🔀 非常没帮助 😧 没帮助 🙂 一般 😮 有帮助 😝 非常有帮助 ©2022 CSDN 皮肤主题:精致技术 设计师: CSDN官方博客 返回首页

公安备案号11010502030143 京ICP备19004658号 京网文〔2020〕1039-165号 经营性网站备案信息 北京互联网违法和不良信息举报中心 家长监护 网络110报警服务 中国互联网举报中心 Chrome商店下载 账号管理规范 版权与免责声明 版权申诉 出版物许可证 营业执照 ©1999-2022北京创新乐知网络技术有限公司

> **樱与刀** 码龄5年 **♥** 暂无认证 31 6万+ 67万+ 12万+ 🥰 原创 周排名 总排名 访问 等级 1142 53 111 17 619 积分 粉丝 获赞 评论 收藏 关注 私信 り 搜博主文章 Q 关于文本分类(情感分析)的中文数据集汇 总 🧿 41201