

数据科学实验报告

汽车行业用户观点主题及情感识别

- 1. 小组成员
- 2. 分工情况
- 3. 比赛情况
- 4. 问题定义
 - 4.1. 背景
 - 4.2. 任务
- 5. 解决思路
- 6. 解题过程
 - 6.1. 数据处理
 - 6.1.1. 流程概述
 - 6.1.2. 中文分词
 - 6.1.3. 词性标注
 - 6.2. 模型训练
 - 6.2.1. 多种方法效果
 - 6.2.2. 额外数据爬取
 - 6.2.3. 后续尝试改进
 - 6.3. 模型预测
- 7. 收获总结
- 8. 附代码
- 9. 参考资料

1. 小组成员

- 朱河勤 PB16030899
- 王博 PB16020870
- 周宇淮 PB16021099
- 邵毅诚 PB16021405


2. 分工情况


- 朱河勤: 负责整个数据的处理, 模型的训练与优化, 报告的撰写
- 王 博: 负责额外数据的爬取, 以及尝试基于 `tf-idf` 的分类方法, 报告的撰写
- 周宇淮: 负责尝试其他模型, 尝试过 `gauss-regression`
- 邵毅诚: 负责联络


3. 比赛情况

我们小组选取的是 DataFountain 比赛


- 比赛名称: 汽车行业用户观点主题及情感识别
- 队伍名: 对不对



**对不对**

)

队伍 ID: 65261




mbinary

20

单身 (•̀▽•́)

PythonC/C++


发送私信



浪浪浪浪儿

--

发送私信




UNION

22

喜欢各种东西

MMA

发送私信



DF1538985590202

--

发送私信

- 比赛排名(截止时间 2018-10-25)
 - A榜: 397/1701
 - B榜: 413/1701
 - 比赛成绩: 0.609

397	↓ 128	对不对	0.60918770	22	2018-10-12 09:57:17
-----	-------	-----	------------	----	---------------------

4. 问题定义

4.1. 背景

随着政府对新能源汽车的大力扶植以及智能联网汽车兴起都预示着未来几年汽车行业的多元化发展及转变。汽车厂商需要了解自身产品是否能够满足消费者的需求，但传统的调研手段因为样本量小、效率低等缺陷已经无法满足当前快速发展的市场环境。因此，汽车厂商需要一种快速、准确的方式来了解消费者需求。

4.2. 任务

本赛题提供一部分网络中公开的用户对汽车的相关内容文本数据作为训练集，训练集数据已由人工进行分类并进行标记，参赛队伍需要对文本内容中的讨论主题和情感信息来分析评论用户对所讨论主题的偏好。讨论主题可以从文本中匹配，也可能需要根据上下文提炼。

5. 解决思路

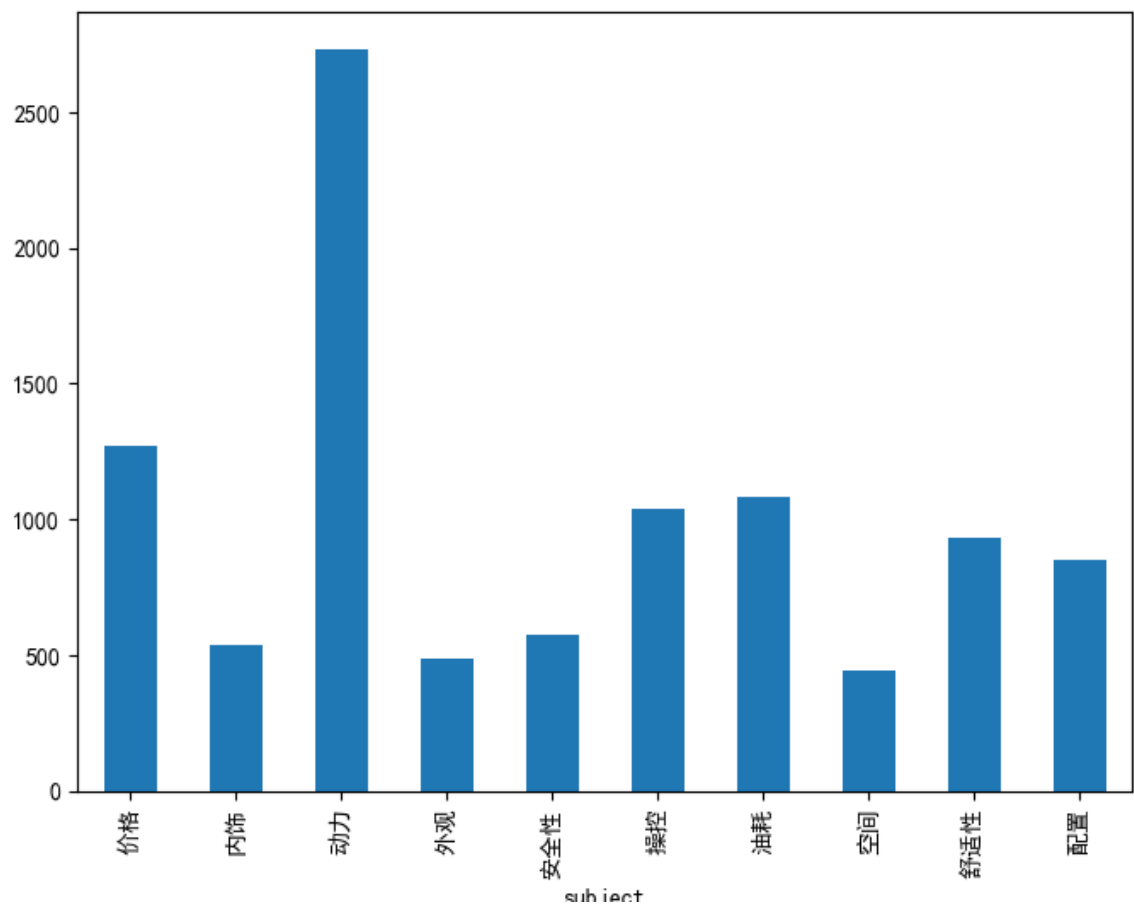
对于是情感计算, 与主题多标签分类, 我们首先使用了 `tfidf` 进行向量化, 然后用 `Logistic Regression` 模型进行拟合, 更进一步的, 我们尝试了词性标注, 停用词处理等等步骤 在效果不是很好的情况下, 我们考虑了是否数据太少, 所以爬取了一些数据, 进行模型的训练. 之后我们又进行了深度学习模型的训练, 使用了多层感知机.

6. 解题过程

6.1. 数据处理

6.1.1. 流程概述

首先观察数据, 如对于主题的分类, 有如下的数据



然后我们考虑数据的清洗

- 去除停用词^[1]
- 中文分词
- 词性标注
- `tfidf` 向量化

6.1.2. 中文分词

我们使用python 中的 jieba 程序包进行中文分词, 并且输出词性信息. 经过停用词去除和分词, 原文和分词结果的对比如下:

96A5VQrB	什么都没有, 就是现金优惠和给了点小东西, 你最好先问问看有没有车	价格	0	优惠	什么(r) 现金(n) 优惠(vn) 小东西(nr) 最好(a) 先(d) 问问看(l) 有没有(v)
JXIEzBcKG	原来最低也就优惠15000至16000	价格	0	优惠	原来(d) 最低(a) 优惠(vn)
P5A3evWt	我们这里, 优惠2000, 我呵呵了	价格	-1	呵呵	优惠(vn)
RJe23zrhZ	遵义优惠6000, 送十万公里保养	价格	0	优惠	遵义(v) 优惠(vn) 送(v) 公里(q) 保养(v)
zYfd7eDQ	深圳这破地方居然只优惠5K, 看大家说的竟然有1万5的	价格	-1	只优惠	深圳(ns) 破(v) 地方(n) 居然(d) 优惠(vn) K(eng) 竟然(d)
LOCJwcQR	我只看到优惠20K, 不错。呵呵	价格	0	优惠	看到(v) 优惠(vn) K(eng) 不错(a)
SqulQghlw	本来谈到2.5豪华优惠1万, 因为想要10月份以后的车, 销售推荐了特装。	价格	1	值	本来(t) 谈到(v) 豪华(a) 优惠(vn) 想要(v) 月份(n) 以后(f) 销售(vn) 推荐(v) 特装(n) 尊贵(a) 差(a) 轮毂(n) 值(v)
ACrPWTU	优惠1万, 上5年保险延保两年, 不交钱	价格	0	优惠	优惠(vn) 保险(n) 延保(nr) 不(d) 交钱(n)
YVfZamz3	哦, 套路。。。送的礼包也只有保养实惠点, 现金优惠搞到最大, 再	价格	-1	套路	套路(n) 送(v) 礼包(n) 保养(v) 实惠(vn) 现金(n) 优惠(vn) 搞(v) 最大(a) 弄(v) 保养(v)
MsCmRYy	特装确实不错, 配置很实用	配置	1	实用	特装(n) 确实(ad) 不错(a) 配置(v) 实用(v)
bhpH5MC	没cd, 没后视镜记忆。之家配置单里写有的	配置	-1	没	cd(eng) 后视镜(n) 记忆(n) 家(q) 配置(v) 单里(n) 写有(v)
KOD82cjZ	长城配置好! 性价比高!	配置	1	好	长城(ns) 配置(v) 性价比(n) 高(a)
KOD82cjZ	长城配置好! 性价比高!	价格	-1	高	长城(ns) 配置(v) 性价比(n) 高(a)
IkRcwBU9z	特装我觉得就是多了几个不实用的配置, 除了电动尾门, 其他都没意义	配置	-1	不实用	特装(n) 觉得(v) 不(d) 实用(v) 配置(v) 除了(p) 电动(n) 尾门(n) 意义(n)
1WpNIYyx	摩雷的听爱卓和优特声人声都非常OK, 而且价格不算高。DLS的低频	配置	1	OK	摩雷(nrt) 听(v) 爱卓(nr) 优特(a) 声(q) 听(v) 人声(n) 非常(d) OK(eng) 价格(n) 不算(v) 高(a) DLS(eng) 低频(b) 应该
tij956MLse	这个山寨导航屏幕很背光 换任何导航软件 我都提不起兴趣	配置	-1	背光	山寨(ns) 导航(v) 屏幕(n) 背光(n) 换(nz) 导航(v) 软件(n) 提不起(v) 兴趣(n)
q9wYl6kV	原车的卡打开 显示所有文件 拷贝到16G卡(先	配置	0	nan	原车(n) 卡(n) 打开(v) 显示(v) 文件(n) 拷贝(n) 文件(n) 除了(p) ini(eng) 文件(n) 拷贝到(nz) G(eng) 卡(nr) 先(d) 格式
GI4JT7Rvn	个人感觉 改导航 没什么用 还不如手机, 2.0改那个中控小彩屏真心不错,	配置	1	不错	个人感觉(n) 改(v) 导航(v) 没什么(l) 手机(n) 改(v) 中控(l) 彩屏(n) 真心(d) 不错(a)
A6NxsYld5	还是没有手机导航更新快 而且还有实时路况	配置	-1	没有	手机(n) 导航(v) 更新快(l) 实时(d) 路况(n)
sz5SD8Qw	不仅16款, 斯巴鲁的导航全是到港装国产山寨导航, 分辨率正常啊	配置	-1	山寨	导航(v) 全(a) 港装(n) 国产(n) 山寨(ns) 导航(v) 分辨率(n) 正常(d)
9nVacW5l	别卖了兄弟, 森林人好开不保值。买的时候尊贵值钱, 卖的时候尊贵不	配置	-1	不值钱	卖(v) 兄弟(n) 开(v) 不(d) 保值(n) 买(v) 尊贵(a) 值钱(v) 卖(v) 尊贵(a) 不值钱(n) 导航(v) 不值钱(n) 少(a) 公里(q) 良
gyRioxAM	XT性价比不高, 太贵啦! 还是买2.5豪华导航版吧 (2.5低配)。	配置	-1	贵	XT(eng) 性价比(n) 不(d) 高(a) 太贵(nr) 买(v) 豪华(a) 导航(v) 版(n) 低(a) 配(v)
rf7K8CHQ	个人意见 改导航 还不如手机好用, 音响改装就是花无钱。	配置	-1	比不上	个人(n) 意见(n) 改(v) 导航(v) 还不如(l) 手机(n) 音响(n) 改装(v) 花(v) 无(v) 钱(n)
8qlz60Ch5	我16款开导航一会儿屏幕就黑了	配置	-1	黑	开(v) 导航(v) 屏幕(n) 黑(a)
aA6qlSkot	2015款特装的导航准升级了?	配置	0	nan	特装(n) 导航(v) 升级(vn)
n2pkFOY6	原车CD音质好的很, 我把4S送的导航拆了又换回CD机头了。纯粹, 简单	配置	1	好	原车(n) CD(eng) 音质(n) S(eng) 送(v) 导航(v) 拆(v) 换回(v) CD机(n) 头(n) 纯粹(a) 简单(a) 实用(v) 精神(n) 开(v) 玩
hL8JZfmm	垃圾导航 都不晓得怎么退出 比如听收音, 退到主界面还是收音在响,	配置	-1	垃圾	垃圾(n) 导航(v) 不(d) 晓得(v) 退出(v) 听(v) 收音(n) 退到(v) 主(b) 界面(n) 收音(n) 响(v) 换成(v) USB(eng) 切换(v) 主
QPWzrFte	我的导航从没有死机过。	配置	0	nan	导航(v) 死机(n)
hUAmRW	显示屏幕应该不支持蓝光的, 原装的导航太差	配置	-1	差	显示(v) 屏幕(n) 应该(v) 不(d) 支持(v) 蓝光(nr) 原装(n) 导航(v) 太(d) 差(a)
zsueijMfuK	说实话, 基本上用不上车上导航, 用手机更方便! 音响效果不用纠结, 主	配置	0	nan	说实话(l) 基本上(n) 用不上(l) 车上(s) 导航(v) 手机(n) 更(d) 方便(a) 音响效果(n) 不用(v) 纠结(v) 毕竟(d) 想(v) 成

6.1.3. 词性标注

创建许多额外基于文本的特征有时可以提升模型效果。比如下面的例子:

- 文档的词语计数—文档中词语的总数量
- 文档的词性计数—文档中词性的总数量
- 文档的平均字密度 -- 文件中使用的单词的平均长度
- 完整文章中的标点符号出现次数 -- 文档中标点符号的总数量
- 整篇文章中的大写次数—文档中大写单词的数量
- 完整文章中标题出现的次数—文档中适当的主题（标题）的总数量
- 词性标注的频率分布
- 名词数量
- 动词数量
- 形容词数量
- 副词数量
- 代词数量

6.2. 模型训练

6.2.1. 多种方法效果

考虑到是多标签^[2]进行分类, 我们将结果标签化, 使用 0~9 分别代替各个类. 然后在选择模型时, 我们考虑了 Logistic Regression 和 support vector machine, 两种效果相差不大, LR 稍微好一点.

如下是训练 LR 模型的评估结果

training: subject					
	precision	recall	f1-score	support	
0	0.83	0.82	0.82	400	
1	0.62	0.47	0.54	177	

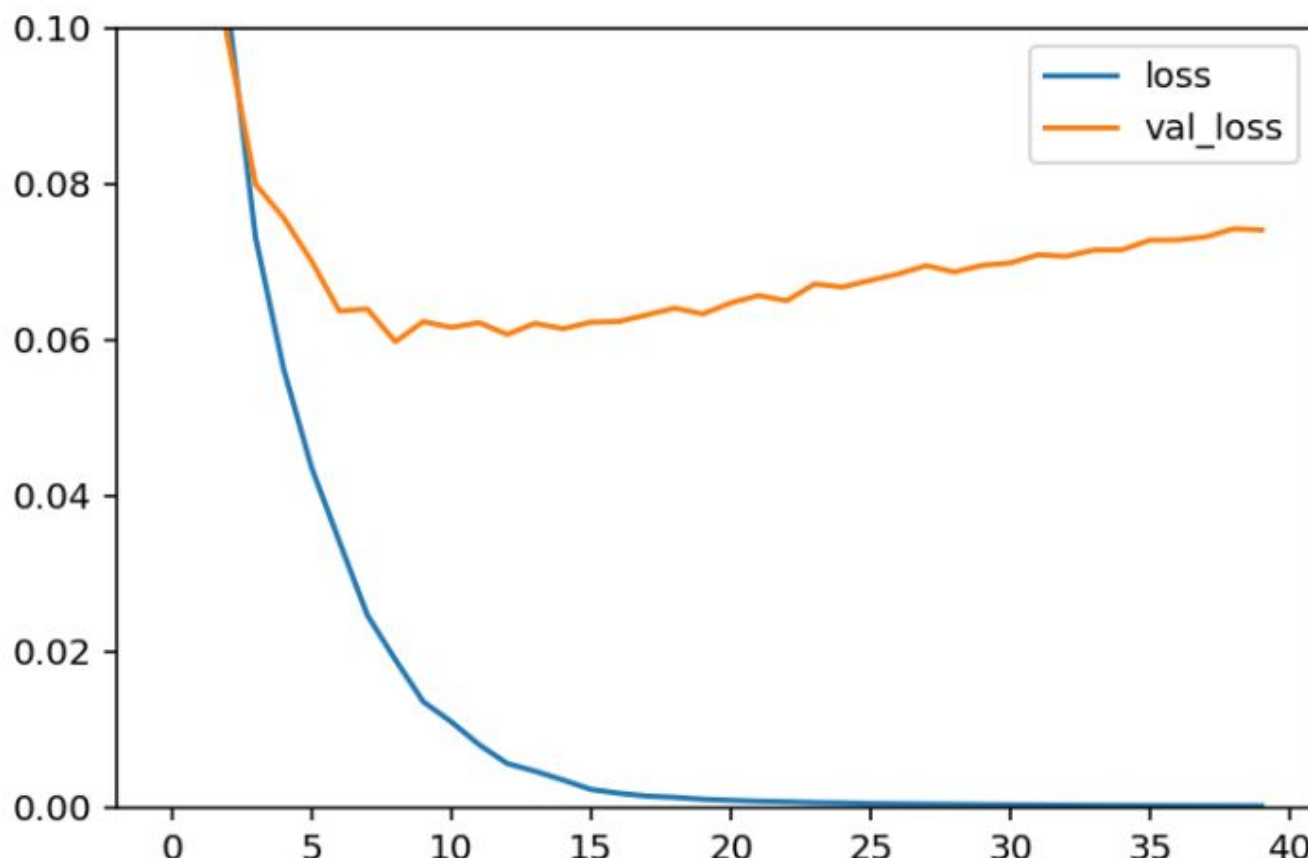
2	0.64	0.84	0.72	797
3	0.58	0.46	0.51	147
4	0.71	0.59	0.64	166
5	0.55	0.57	0.56	283
6	0.85	0.77	0.81	328
7	0.57	0.53	0.55	135
8	0.73	0.53	0.62	281
9	0.79	0.68	0.73	271
avg / total	0.70	0.69	0.69	2985
training: sentiment_value				
	precision	recall	f1-score	support
0	0.68	0.08	0.15	480
1	0.72	0.97	0.82	2021
2	0.54	0.23	0.32	484
avg / total	0.68	0.71	0.63	2985

后来, 我们又尝试了基于 tf-idf 的分类方法, 类似检索的方式, 找到最有可能的分类. 得到索引文件如下:

word	价格_归一	配置_归一	操控_归一	舒适性_归一	油耗_归一	动力_归一	内饰_归一	安全性_归一	空间_归一	外观_归一
发动机	0.1556539	0.1626063	0.5227827	0.6243098	0.4578267	0.7004717	0.3573554	0.2190111	0.2390919	0.3917024
油耗	0.1319012	0.1889728	0.3047019	0.4184228	0.9547039	0.3269701	0.2631428	0.1875437	0.3950828	0.1236149
机油	0.2885627	0.1688805	0.180764	0.7813941	0.7189438	0.6654399	0.2553214	0.3016859	0.2281415	0.1031069
动力	0.2469523	0.4819457	0.5679892	0.4415679	0.7747883	0.3013934	0.5714731	0.2250828	0.9300898	0.5110105
价格	0.9963592	0.506178	0.259496	0.1225053	0.1957594	0.2534631	0.2887788	0.0710873	0.4607806	0.3998332
换	0.636305	0.8609786	0.8965431	1.1832676	0.7985373	1.2294537	0.4432925	0.7518878	0.3665232	0.3533813
买	1.6713612	0.9875931	0.8965431	0.6728384	0.598903	0.9329617	1.0276326	0.5466078	1.6615719	1.2810074
开	0.2520589	0.4410241	1.217524	0.0679117	1.5134458	1.0894417	0.6812109	0.6372234	1.101445	0.4751633
高	1.1905942	0.9476367	0.8561024	0.7235357	1.857309	0.867082	0.7959343	0.7053526	1.1430063	0.7117234
内饰	0.293255	0.5082361	0.8136738	0.6208742	0.1892055	0.3526317	0.9194441	0.3152454	1.1170527	0.2410142
感觉	0.3993253	0.5798385	1.6046457	1.3724193	0.9777257	1.0761842	0.8971329	0.7193168	1.0879259	1.2924166
空间	0.3327473	0.4794618	1.0151208	0.6903161	0.3644866	0.3475183	1.0082818	0.2804033	1.11366	0.925972
配置	0.9922318	0.4048717	0.5334084	0.2035088	0.1605155	0.4565601	1.3255597	0.495986	0.9644796	1.0009316
烧	0.2163437	0.1899218	0.1563738	0.5742328	0.494095	0.8819027	0.3324694	0.0848185	0.2199139	0.1656475
操控	0.2429169	0.2819633	0.0079587	0.7934733	0.4286966	0.4464689	0.9294944	0.4197464	1.7490563	0.5621144
太	0.7898452	0.8419636	0.9185408	1.465704	0.8960843	0.9003725	1.7753886	0.7520366	1.056169	1.2851137
刹车	0.1297499	0.2366663	1.1868829	0.5125267	0.2035386	0.3963366	0.2054373	1.1049873	0.0830425	0.1876523
跑	0.3762212	0.3455169	0.6934309	0.9101365	0.30814113	1.2203832	0.4123964	0.5786504	0.5001006	0.3390252
高速	0.1905811	0.3281762	0.5944551	1.2227708	0.30701363	1.2022558	0.5222655	0.5862499	0.675558	0.3053135
导航	0.3155583	0.5083596	0.2400338	0.2876518	0.1767916	0.1470373	0.1784407	0.1669184	0.1298338	0.1564732
低	1.1277686	0.9377044	1.0294232	0.660879	0.1229567	0.9158609	0.6122173	0.3937209	1.0672249	0.4194126
外观	0.3235539	0.6035816	0.5566001	0.1990847	0.0761338	0.3844452	0.6126974	0.2156461	0.9318644	1.0612963
底盘	0.2637578	0.2706185	0.4438682	0.8565401	0.1745537	0.4224699	0.8221826	0.7690923	0.9020801	0.6437188
觉得	0.6816316	0.7939538	1.2257044	1.1820829	0.5867975	0.9838237	1.0660884	1.0341837	1.2928131	1.9908715
公里	0.4200982	0.4012454	0.4749056	0.9880023	0.3016298	1.2606204	0.3192746	0.4853197	0.1935873	0.2187259
空调	0.0675061	0.3274202	0.1244236	0.2454793	0.6750917	0.4875533	0.4008172	0.412429	0.3888471	0.0439342
不错	0.8456521	0.7824612	1.4963145	1.0175439	0.8755391	0.7959592	1.606739	0.7139193	1.7535993	1.1007312
变速箱	0.2932341	0.2316794	0.5298759	0.4952945	0.4667609	0.5799914	0.3686987	0.7664243	0.1490366	0.22452
驱	0.3898054	0.555295	0.3093093	0.5572264	0.6462313	0.8090939	1.0099502	0.7479155	1.7350427	0.78414

通过该文件, 可以比较清楚的知道单个单词在不同类别中的比例. 在自己的训练/测试集划分上, 达到F1=0.74, 但是在线测试 F1 是 0.55.

后来我们尝试了 多层感知机 (MLP) 模型, 下面是训练 40 个 episodes 的 loss, val_loss 数据



出现了过拟合现象, 训练集的 `loss` 已经降到 0 了, 但是验证集的 `val_loss` 一直在上升, 因此这不是一个很好的模型, 因为它太过拟合了.

我们猜测可能是数据太少的原因, 后来我们尝试了在相似的汽车评论网站自己爬取一些数据.

6.2.2. 额外数据爬取

在 <https://auto.news18a.com/> 这个网站上, 我们找到了有类似格式的数据, 我们打算将网站上的评论数据爬去下来.

我们需要的是对汽车各个指标的评论数据. 这个网站对每种型号的车有一个页面, 每个页面下方, 通过点击可以拉取评论.

爬取的流程如下:

1. 在这个网站上找到足够数量的, 各种品牌的汽车页面
2. HTTP请求下载汽车页面中, 获取拉取评论所需的参数
3. 根据参数, 发送HTTP请求, 调用网站API拉取评论
4. 对拉取的原始数据进行处理, 将有用的按照格式写入csv文件

通过观察网页URL参数, 该网站的每种型号的车对应一个4位数ID号. 但是有些ID号码不对应汽车.

于是, 我首先写了一个脚本, 检查了ID号码为2000-4000的页面, 忽略404, 得到了一个可用ID列表.

接下来, 通过另外一个脚本, 从这些ID拉取全部评论. 由于评论时间跨度长, 较老老的评论和现在的格式不同, 因此适当舍弃一些或者转换一部分.

爬取的部分数据如图所示:

3407-3497.csv - Excel

文件 开始 插入 页面布局 公式 数据 审阅 视图 帮助 团队 告诉我想要做什么

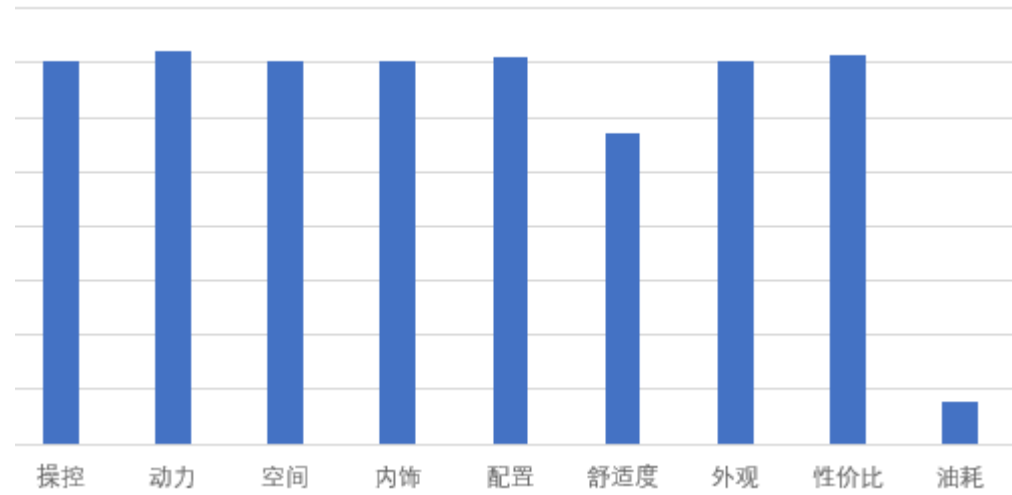
A1

✕ ✓ f_x

content

	A	B	C	D	E
1	content	subject	sentiment	score	
2	这款车的设计还是非常时尚的,侧身上挑的腰线看起来非常有型,尾部富有动感。	外观	0	4	
3	此款的内饰,做工确实不错,但是样子总觉得一般般,用料也一般般,整体比较简洁。	内饰	0	4	
4	虽然轴距只有两米五,但坐在里边感觉这车还是属于紧凑实用型的,值得提的是他的后	空间	0	4	
5	北汽的操控一直就不差,它的操控整体给我的感觉是底盘很扎实,通过性也非常不错,操控		1	5	
6	1.5L发动机一直是北汽的杀手锏,动力非常不错,而且北汽的这个1.5L发动机稍微踩点油	动力	0	4	
7	配置应该一直是北汽的短板,跟韩国车和咱们的自主品牌来比,那配置简直不是一个级	配置	1	5	
8	如果配置再增加一些的话,这车的性价比就高了。另外车的油耗也是一大亮点,确实不	性价比	0	4	
9	虽然是织物的座椅,但包裹性和支撑性还是不错的,后排的3个头枕枕上去挺舒服的。	舒适度	0	4	
10	北汽E系列这款自动版新设计的车身轮廓让它看起来更耐看,车身更饱满;车身侧面,外观		0	4	
11	从方向盘到中控台,典型的国产系严谨细腻无时不在显露。内饰最漂亮的莫过于中控台	内饰	1	5	
12	前排空间还好,由于轴距的原因,后排空间不是那么的理想,不过作为一辆小型代步车	空间	0	4	
13	操控性能是北汽不得不提的一大优点。硬朗的悬挂支撑,搭配精准的转向以及扎实的底	操控	0	4	
14	1.5L的动力总成只能说是中规中矩,没有太大亮点,好在是自动挡的,起步一脚油门就	动力	0	4	
15	配置一般般吧,没有什么亮点,不过好像太花哨了咱也用不到,如果出问题了还得花钱	配置	1	5	
16	性价比还不错,主要是外观非常讨人喜欢,作为家庭用车非常合适,外出旅行也挺好的	性价比	0	4	
17	舒适上称不上豪华,不过也能过得去,虽然悬挂有些硬,不过还是控制在了可接受范围	舒适度	0	4	
18	外观没说的,车头再长一点就好了	外观	0	4	
19	做工粗糙,中控台和车门缝隙,右边半公分,左边一公分半,左后车门B柱上的密封条	内饰	0	3	
20	空间不错,后排比老捷达好多了,亲戚的速腾后排也不过如此。就是后排稍矮一点,1.8	空间	1	5	
21	方向盘稍重,不过跑高速是好事	操控	0	3	
22	日常代步还可以,1.3的发动机带1.1吨的车,不知道时间久了会不会费油	动力	0	3	
23	低配的,配置谈不上好,但是这个价格,安全配置是满意的	配置	0	3	
24	如果性价比不好就不买它了	性价比	1	5	

爬取完成后,对爬取条目进行抽样统计,大致得到爬取数据的各个标签比例:



出现这个问题,是由于该网站老版本评论是大部分,但是老版本的评论不包括油耗类别.

因此,在筛选训练数据时,除油耗外,选择较少比例的数据,尽量使得最终用于训练的数据平衡.

6.2.3. 后续尝试改进

在这个过程中我们尝试了调整一些超参数,比如神经网络的层数, dropout 层数, 输入结点数量等等. 然而效果仍然不佳, 于是我们提交结果的时候仍然用的 LR 模型

训练好的模型要保存^[3]下来, 以供后面环节使用.

6.3. 模型预测

上面训练好的模型, 导入测试数据, 进行预测, 得到提交的结果 并提交了很多次, 如下图

历史记录

我的提交

队伍提交

1. submit_merge_with_sentiment_value.csv

所在赛程: 初赛 - B 榜

状态 / 得分: 0.58689120000

查看日志

提交时间:2018/10/20 23:45

备注: 无备注信息

2. subject-0.7184-value-0.7373-no-tag.csv

所在赛程: 初赛 - B 榜

状态 / 得分: 0.60292643000

查看日志

提交时间:2018/10/19 23:55

备注: 无备注信息

3. 0.73-0.73.csv

所在赛程: 初赛 - A 榜

状态 / 得分: 0.60119840000

查看日志

提交时间:2018/10/13 08:44

备注: 无备注信息

4. subject-0.7-value-0.72.csv

所在赛程: 初赛 - A 榜

状态 / 得分: 0.59986687000

查看日志

提交时间:2018/10/13 08:23

备注: 无备注信息

5. subject-0.7-value-0.72.csv

所在赛程: 初赛 - A 榜

状态 / 得分: 文件异常

查看日志

提交时间:2018/10/13 08:21

备注: 无备注信息

6. subject-0.7021-value-all-zeros-no-tag.csv

所在赛程: 初赛 - A 榜

状态 / 得分: 0.59986687000

查看日志

提交时间:2018/10/13 08:18

备注: 无备注信息

7. submit-other.csv

所在赛程: 初赛 - A 榜

状态 / 得分: 0.60918770000

查看日志

提交时间:2018/10/12 09:57

备注: 无备注信息

同样的办法, 我们先进行了主题的预测, 后来又进行了评论情感的预测^[4]. 考虑到数据的特别, 0(即中性) 非常多, 我们最开始提交的是全是0 的数据, 后来进行了预测, 结果相差不大. 这里体现可以根据数据的特点进行启发式的预测

7. 收获总结

在参与这个比赛的过程中, 我们对数据科学的认识更进一步了. 我们知道数据科学

- 最耗时的步骤在数据处理部分, 是否有好的数据, 是否结合了特定场景的信息, 是否进行了数据集的划分
- 关键点在模型的训练, 模型是否合适, 该怎样有方向地调参(而不出盲目地), 是否尝试多种模型等等

- 模型的改进与评测是不可或缺的环节. 如何评测这个模型的好坏, 如何对比不同模型的优劣, 如何根据训练结果改进模型等等

8. 附代码

- lr.py: 使用了 tfidf, LR 模型
- kfold.py: 使用了 k 折交叉验证, 分别预测了多个结果, 取最可能的结果
- mlp.py: 多层感知机模型
- tagging.py: 词性标注模块
- multi_label_evaluation.py: 多标签的评测方法, 包括 `f1`, `hamming distance`, `jaccard`

9. 参考资料

1. [中文常用停用词](#)
2. Shivam Bansal [手把手教你在 Python 中实现文本分类](#) (附代码、数据集)
3. 拾毅者 [机器学习 - 训练模型的保存与恢复](#)
4. [wiki 文本情感分析](#)