NLPartners-在线舆情自动检测系统

2020.05.14

目录

项目介绍

本项目的主要目的是检测美团杭州区域的各个商家情况，通过各个商店的用户评论，来对商店进行评价，评分为1～5(1表示最低评价，5表示最高评价)。本项目统计整个平台中对所有商家的评价以及各类评价的占比。同时选出了对评价较好的前十位商家，列举了这些商家的用户评论，提取了评论的关键词。用户可通过该平台观测整个舆情的走向，也可作为商家后续改进的参考。

本平台同时提供用户管理的操作，不同的用户可在平台注册查看”美团”平台最近的评论的情况。

数据处理

数据处理主要分为两个部分：

1. 平台数据处理

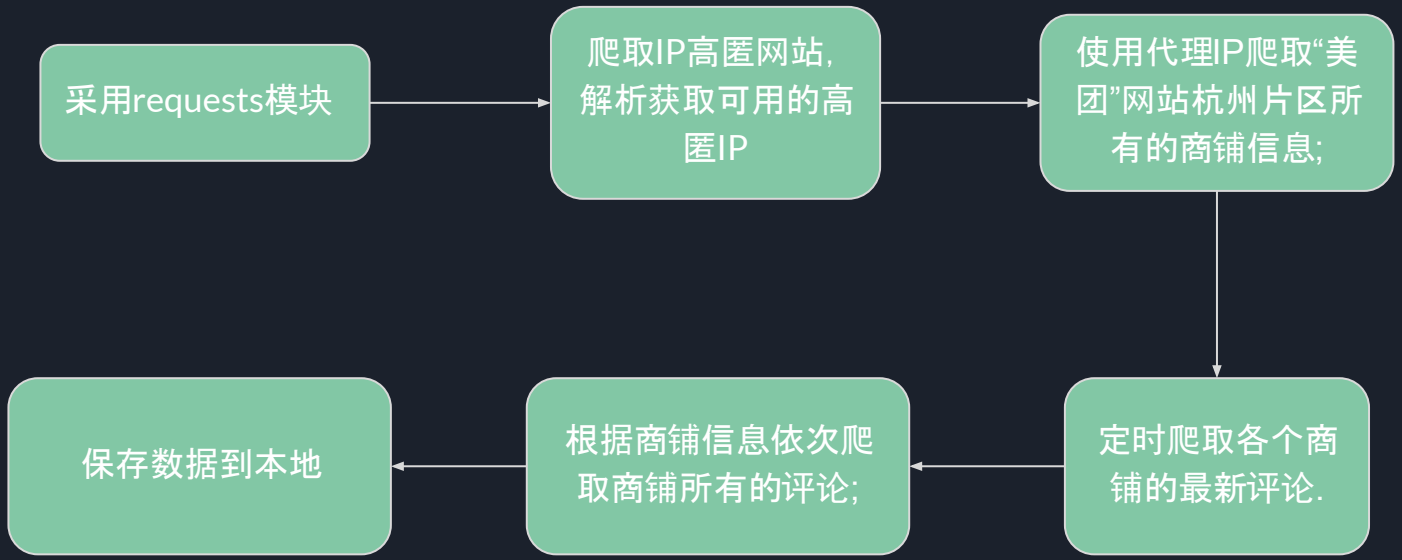
   实时从美团的商家页面爬取评论等数据，然后进行处理

2. 分类模型数据处理

   对每条评论进行评分分类的模型训练用的数据的处理，主要采用已有的数据集

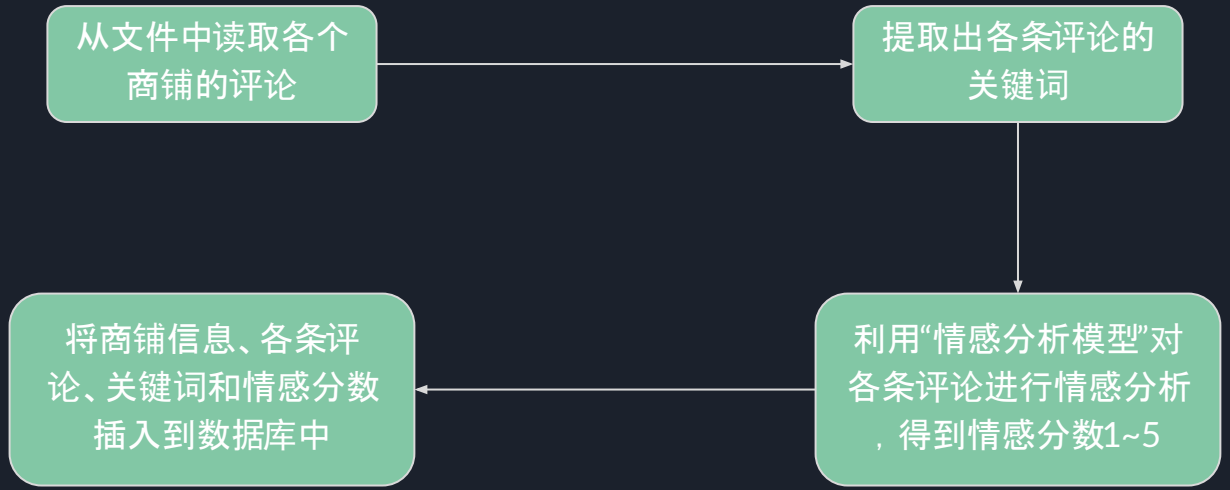
平台数据处理

1. 数据爬取



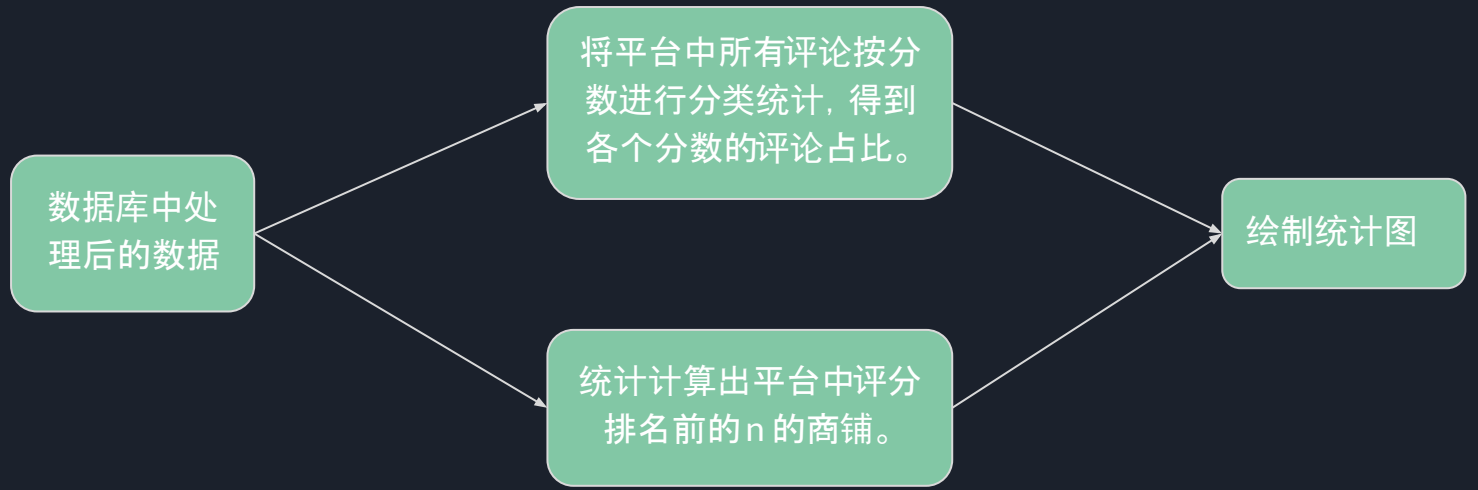
平台数据处理

2. 数据处理

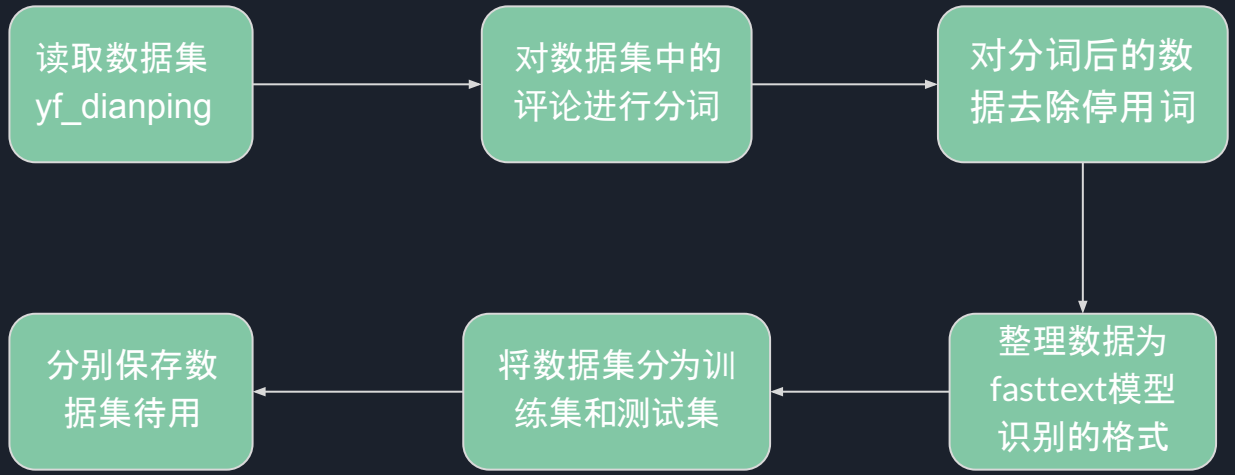


平台数据处理

3. 数据统计



分类模型数据处理


读取数据集
yf_dianping
对数据集中的
评论进行分词
对分词后的数
据去除停用词
整理数据为
fasttext模型
识别的格式
将数据集分为训
练集和测试集
分别保存数
据集待用

情感分类模型

项目中需要对各条评论的情感进行分析，需要对采集的商铺评论进行打分分类。在分类模型的选择中，直接采用了fasttext分类模型进行处理。

FastText是Facebook于2016年开源的一个词向量计算和文本分类工具，FastText模型架构和word2vec的CBOW模型架构非常相似。

FastText模型架构如有图所示。其相比CNN等模型，其训练的速度相对来说快的多。

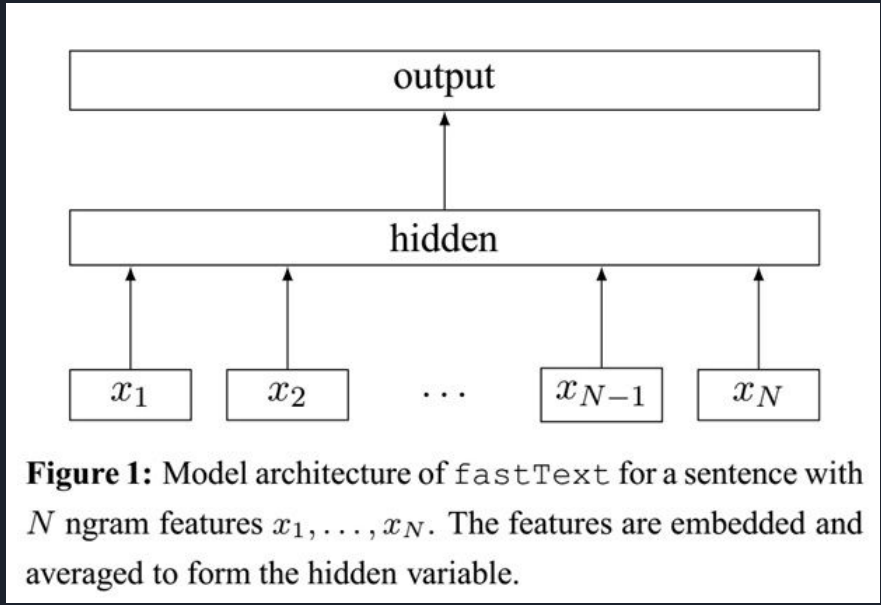


Figure 1: Model architecture of `fastText` for a sentence with N ngram features $x_1, \ldots, x_N$. The features are embedded and averaged to form the hidden variable.

FastText原理介绍

和CBOW一样，fastText模型也只有三层：输入层、隐含层、输出层（Hierarchical Softmax），输入都是多个经向量表示的单词，输出都是一个特定的target，隐含层都是对多个词向量的叠加平均。不同的是，CBOW的输入是目标单词的上下文，fastText的输入是多个单词及其n-gram特征，这些特征用来表示单个文档；CBOW的输入单词被onehot编码过，fastText的输入特征是被embedding过；CBOW的输出是目标词汇，fastText的输出是文档对应的类标。

从隐含层输出到输出层输出，会发现它就是一个softmax线性多类别分类器，分类器的输入是一个用来表征当前文档的向量；模型的前半部分，即从输入层输入到隐含层输出部分，主要在做一件事情：生成用来表征文档的向量。那么它是如何做的呢？叠加构成这篇文档的所有词及n-gram的词向量，然后取平均。叠加词向量背后的思想就是传统的词袋法，即将文档看成一个由词构成的集合。

fastText的核心思想就是：将整篇文档的词及n-gram向量叠加平均得到文档向量，然后使用文档向量做softmax多分类。这中间涉及到两个技巧：字符级n-gram特征的引入以及分层Softmax分类。

FastText 主要参数说明

序号	参数	说明
1	lr	学习率，默认值0.1
2	dim	词向量维度，默认为100
3	ws	窗口大小，默认为5
4	epoch	训练轮询的次数，默认为5
5	minCount	word出现的最小次数，默认为1
6	minCountLabel	label出现的最小次数，默认为1
7	minn	char gram的最小长度，默认为0
8	maxn	char ngram 的最大长度，默认为0

FastText 主要参数说明

序号	参数	说明
9	neg	负采样的数量，默认为5
10	wordNgrams	word ngram的最大长度，默认为5
11	loss	损失函数，有ns,hs,softmax,ova可选，默认为softmax
12	thread	线程数，默认为cpu核数
13	UpdateRate	学习率的更新频率，默认为100
14	t	数据采样的阈值，默认为0.0001
15	label	数据的标签前缀，默认为'__label__'
16	verbose	日志显示，默认为2(为每个epoch输出一行记录)

模型调参记录(节选)

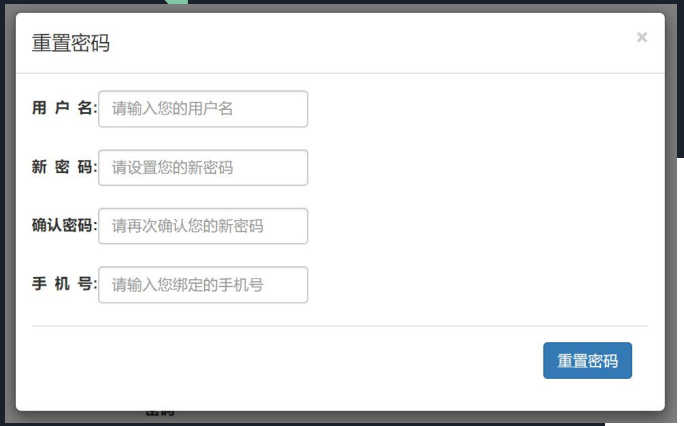
lr	dim	loss	P	R	epoch	ws
0.001	150	1.087576	0.558	0.558	5	5
0.005	150	0.983858	0.582	0.582	5	5
0.005	150	0.963282	0.585	0.585	10	5
0.005	150	0.962978	0.585	0.585	10	10
0.05	150	0.941118	0.584	0.584	10	10
0.5	150	0.959422	0.576	0.576	10	10
0.5	150	0.973034	0.579	0.579	5	5
0.5	150	0.970885	0.581	0.581	5	3
0.1	100	0.954375	0.585	0.585	5	5

模型说明

根据调参的记录来看，Fasttext模型在数据集一定的情况下，调节模型的参数，如学习率、epoch，窗口尺寸等，对精度和召回率的提升并不是特别明显，还需从其他方面进行提升，可通过扩充数据集的方式来改进。

Fastext在模型的训练速度上比其他模型有个比较大的优势就是训练的速度很快，可以在几分钟内完成模型的快速迭代，这是本项目中采用该模型的一个主要原因。另外其训练简单，易于实现，能够满足基本的功能需求。

项目效果展示



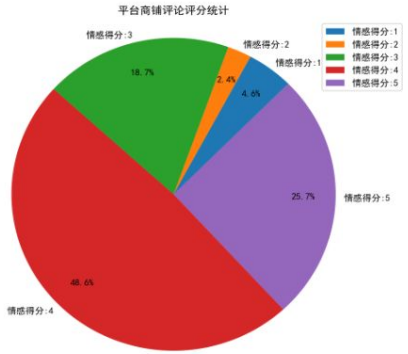
密码重置页面

登录页面

账户注册页面

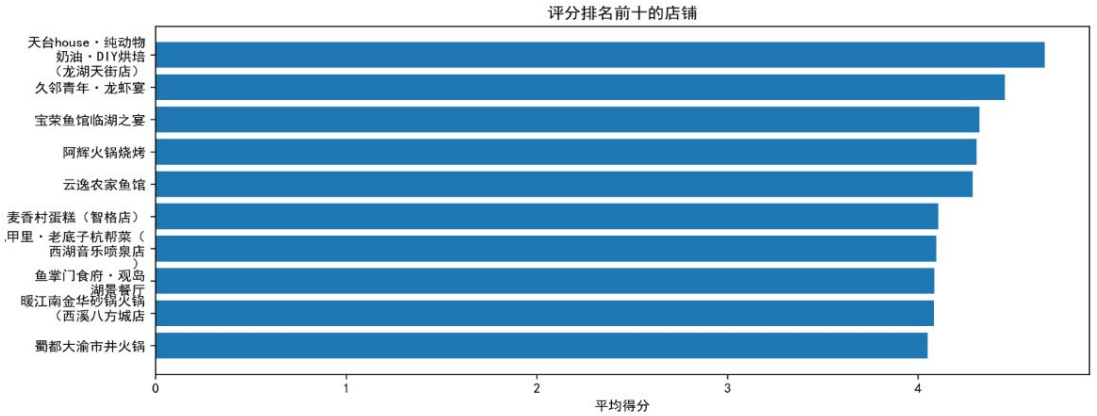
项目效果展示

平台商铺评分总体统计



了解详情

好评排名前十商铺



了解详情

项目效果展示

情感评分	评论总数	关键词
1	2180	味道、进去、难吃、菜品、非常、真的、服务态度、服务、龙虾
2	1132	环境、差不多、宝宝、味道、披萨、好吃、口感、临时、奶油
3	8832	味道、口味、鱼头、菜品、上菜、环境、价格、还行、牛排
4	22903	味道、不错、好吃、环境、口味、菜品、蛋糕、服务、千岛湖
5	12105	好吃、非常、味道、老板、菜品、超级、服务、千岛湖、环境

各类评分关键词

项目效果展示

序号	店铺名称	平均得分	好评率	关键词
1	天台house·纯动物奶油·DIY烘培（龙湖天街店）	4.6655	100.0%	非常、老板、店家、超级、老板娘、不错、好吃、朋友、体验
2	久邻青年·龙虾宴	4.4571	97.14%	味道、龙虾、分量、老板、麻辣、老板娘、一家、菜品、去年
3	宝荣鱼馆临湖之宴	4.3227	98.0%	味道、老板、非常、服务、鱼头、鱼汤、新鲜、不错、总体
4	阿辉火锅烧烤	4.3077	100.0%	味道、一家、晚上、老板、一个、烤鱼、烧烤
5	云逸农家鱼馆	4.2879	100.0%	老板、味道、千岛湖、鱼头、菜品、食材、一个、上菜、处于
6	麦香村蛋糕（智格店）	4.1073	95.64%	蛋糕、不错、非常、好吃、味道、老板娘、口味、性价比、已经

排名靠前的各个店铺的平均得分以及评论的关键词显示

项目效果展示

序号	评论	情感得分	关键词	时间
1	虽然没吃过油焖的，但今天这个味道确实不错。很入味。	4		2020-05-10 18:54:12
2	麻辣小龙虾，真的超级好吃，老板娘和老板也超级好，还送了一瓶啤酒。第一次给别人这评论，不是有过人之处我是不愿意浪费几分钟编辑文字的。强烈推荐。	5	麻辣	2020-05-10 18:54:12
3	老板娘太热情了，两瓶百威还给我免单啦，爱你＾3＾ 套超级划算的，分量超级足，两份套餐足够吃饱了，土豆片超级好吃，强烈推荐! 下次一定还会再来的!	5	老板娘	2020-05-10 18:54:12

序号	评论			
1	非常满意的体验 小姐姐很耐心 从开始的千层版糕成品 体验很棒 很开心👌👌👌			
2	网红蛋糕，老板娘会教，上面的插件随意搭配，			18:47:06
3	店家姐姐真的很温柔 自由发挥的蛋糕 夹心水果很新鲜	5	店家	2020-05-10 18:47:06
4	很好吃，小姐姐教的很用心	5	好吃	2020-05-18:47:06
5	很好吃!	4	好吃	2020-05-18:47:06
6	店长人很好，还帮忙拍照哈哈，千层水果蛋糕挺好吃的，超喜欢吃千层！第一次diy蛋糕，都舍不得吃了哈哈哈	5	店长	2020-05-18:47:06

序号	评论	情感得分	关键词	时间
1	你值得拥有，闻起来是非常香浓的巧克力味，不飞粉显色度很好，就是有点贵你值得拥有，闻起来是非常香浓的巧克力味，不飞粉显色度很好，就是有点贵	4	值得	2020-05-10 18:51:11
2	很好的蛋糕，物美价廉，强烈推荐哦！老板炒鸡好	5	蛋糕	2020-05-10 18:51:11
3	小猪佩琪，小朋友很喜欢	4	小猪	2020-05-10 18:51:11

店铺具体评论的显示

项目评价

优点：

1.实现了对平台评论的实时爬取和评价等的处理和存储；

2.完成了用户的注册登录等功能；

3.根据评论列出了排名靠前的商家的特点；

4.界面操作简单明了，情感分类模型实现简单。

项目不足：

1. 实现了基本功能后，未对其他分类模型作过多探索；

2. 数据的爬取限于网站反爬虫的原因，每个商家只爬取了 500条的数据。

谢谢！