大数据传播与新媒体分析

DATA130036.01

大数据视角下的网络新闻跟帖中地域刻板印象分析

——以网易新闻跟帖数据为例

15级大数据 何占魁

15级大数据 冉诗菡

14级大数据 周之烁

2018.5.16

PART 1

引言

问题描述

选题意义

研究现状

问题描述

项目以**网易新闻跟帖**中的**地域刻板印象**作为研究对象，希望呈现出各省份之间的**印象标签**和**关系**，以及针对这些刻板印象背后的规律给出我们的分析与解释。

选题意义

新闻跟帖能够真实反映**网民的社会心理**

人气最高、最受关注的跟帖中往往夹杂着地域歧视、**相互谩骂**的跟帖

地域歧视一直是媒体与社会公众关注的**热点问题**

对我国**网络空间治理**、**舆情检测与管理**提供有效的分析方法和建议

研究现状

一、网络、传媒地域歧视现象研究

《天涯论坛中的苏北地域歧视现象研究》《网络新闻跟帖中地域歧视现象的现实解读与理性反思》《浅析媒介中的地域歧视现象》《事件新闻报道命名与地域歧视》《新闻报道中的地域歧视性语言现象探析》

二、新闻跟帖研究

《对门户网站新闻跟帖特点的分析———以网易新闻为例》《网民新闻跟帖中的语言暴力研究》

PART 2

项目框架与主要技术


新闻
标题：筛选地区
正文：提取每段关键句用主题模型确定文章主题
探究新闻是否会对回帖产生倾向性引导
回帖
利用回帖用户ip判断回帖者地区
通过地域词典提回帖提到的地域
回帖者对目标地域的情感倾向
项目框架

PART 3

数据

数据来源

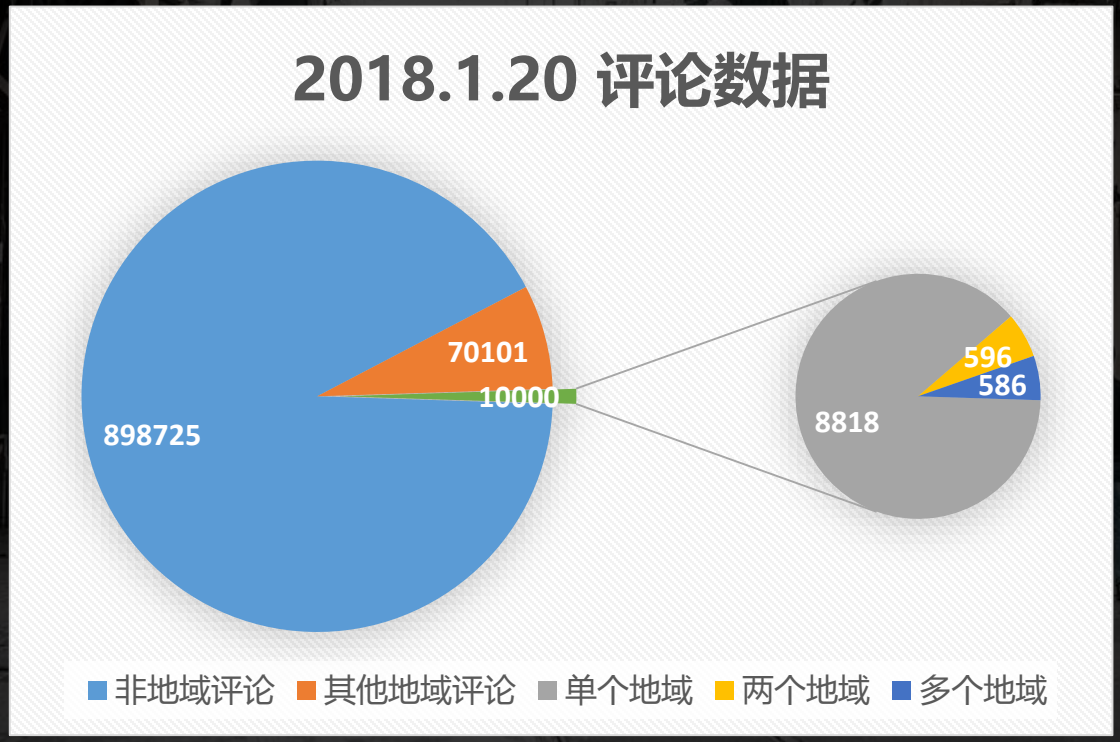
数据预处理

IP定位
地域词典
结果

数据来源

课程提供的网易新闻跟帖文本信息(未完成)

- 为了绕过数据瓶颈，完成模型流程，我们暂时使用了小规模数据集和人工生成的文本数据。



数据格式

文章ID docid	新闻的唯一键值
频道名 channel	新闻所属版块
文章日期 realtime	新闻的发出日期
文章正文 post	新闻的正文内容
文章标题 title	新闻的标题

帖ID tie_id	评论帖的唯一键值
评论内容 content	评论帖的正文

文章ID url docid	所评论新闻的唯一键值
帖ID tie_id	评论帖的唯一键值
用户ID passport	发帖用户的唯一键值
楼层数 floor	发帖的楼层，默认为1
发帖人IP ip	发帖人的IP明码
发帖人省份	发帖人的省份
发帖人市区	发帖人市区，未知为0
第一楼IDf1_tie_id	本帖的第一楼ID
上一楼ID ptie_id	本帖的上一楼ID
发帖时间 realtime	发帖的时间
用户状态 userprofile	用户状态的五位数值
发帖来源 source	发贴来源，wb即pc发贴，ph-ios即iphone端发贴，ph-android即安卓端发贴，其他无法确认的默认均为ph。

IP定位

- 根据网易回帖用户的IP地址，利用Python调用淘宝IP库获取IP归属地返回省份和城市

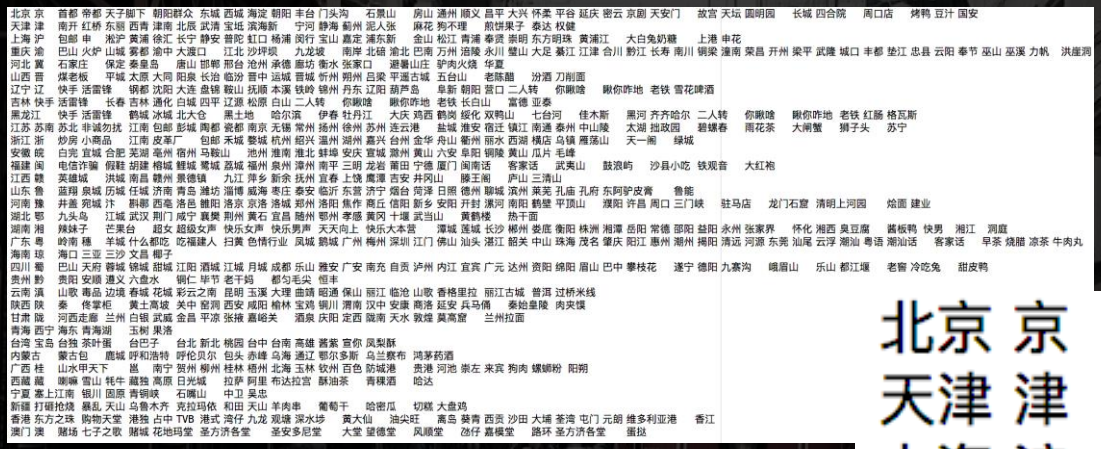
  淘宝IP地址库 http://ip.taobao.com/

  省准确度超过99.8%，市准确度超过96.8%。

地域词典

通过收集各省份具有唯一性的中性关键词，构造各省的地域字典。

一条评论中可能提到多个地域，因此将地域数量上限设置为3。



定位结果

content	province	city	region_1	region_2	region_3
当时外资逃跑的时候你们说中国要崩，如今来看外资就是个屁啊，跑了一点也不影响增长，苏州哈哈哈	吉林	通化	江苏	0	0
一说到苏州肯定有撕逼大戏，赶紧搬凳子过来 [微笑]	广东	广州	江苏	0	0
大郑州才不穷，他们有富士康啊	江苏	苏州	河南	0	0
浙江人务实 不像有些地方人那口气都能上天	江苏	苏州	浙江	0	0
新疆温饱解决了吗	江苏	南京	新疆	0	0
比如苏州世界知名品牌..... 啥?	XX	XX	江苏	0	0

content	province	city	region_1	region_2	region_3
赣州拿财政总收入和惠州一般预算收入比，笑死人了，惠州 16 年财政总收入 800 多亿，还不包括非税...	广东	广州	广东	江西	0
我去过湖州，感觉比上海差一个地球周期	上海	上海	上海	浙江	0
广东外嫁江苏，说实话，苏北真不穷，比广东非珠三角地方富有不知道多少，物价低，生活质量好很多	广东	广州	江苏	广东	0
我是宿迁的，我们村十年前就说要拆迁，到现在也没拆，真服也没个准话，自己家的房子自己家的地就是...	江苏	南京	江苏	浙江	0
搞错了吧，16 年惠州跟赣州差不多 369 亿财政收入。	江西	赣州	广东	江西	0
进来看到浙江被苏北吊打啊。好虚啊 浙江人远离此帖。苏北雄起了。	浙江	嘉兴	江苏	浙江	0

PART 4

主要模型

情感分析模型

主题分析模型

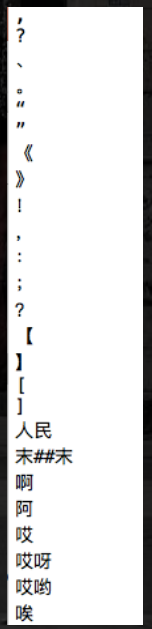
情感分析模型

1) 简单统计正面词汇与负面词汇的个数并比较大小

2) 通过正向词、负向词、程度词和否定词计算情感分数

3) 神经网络训练分类器

移除停用词



情感词典1

	词语	词性种类	词义数	词义序号	情感分类	强度	极性	辅助情感分类
2	脏乱	adj	1	1	NN	7	2	
3	糟报	adj	1	1	NN	5	2	
4	早衰	adj	1	1	NE	5	2	
5	责备	verb	1	1	NN	5	2	
6	贼眼	noun	1	1	NN	5	2	
7	战祸	noun	1	1	ND	5	2	NC
8	招灾	adj	1	1	NN	5	2	
9	折辱	noun	1	1	NE	5	2	NN
10	中山狼	noun	1	1	NN	5	2	
11	清峻	adj	1	1	PH	5	0	
12	清莹	adj	1	1	PH	5	1	
13	轻倩	adj	1	1	PH	5	1	
14	晴丽	adj	1	1	PH	5	1	
15	求索	adj	1	1	PH	3	1	
16	热潮	noun	1	1	PH	5	1	
17	仁政	noun	1	1	PH	5	1	
18	荣名	noun	1	1	PH	5	1	
19	柔腻	adj	1	1	PH	5	1	
20	瑞雪	noun	1	1	PA	5	1	
21	擅名	noun	1	1	PD	7	1	
22	神采	adj	1	1	PA	5	1	PH

情感词典2

作者简介
知网简介
理论与实践
知网论坛
下载中心
相关文章
相关网站
清华论坛

《知网》更新公告

- 2007年10月22日 知网发布“情感分析用词语集（beta版）”

知网从即日起发布“情感分析用词语集（beta版）”，共有12个文件。

1. “中文情感分析用词语集”，现包含如下6个子文件：

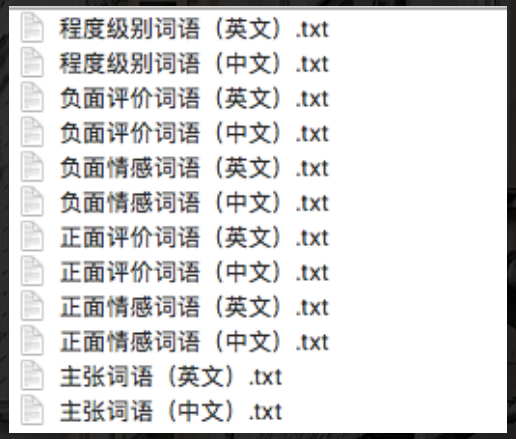
“正面情感”词语，如：爱，赞赏，快乐，感同身受，好奇，喝彩，魂牵梦萦，嘉许 ...
“负面情感”词语，如：哀伤，半信半疑，鄙视，不满意，不是滋味儿，后悔，大失所望 ...
“正面评价”词语，如：不可或缺，部优，才高八斗，沉鱼落雁，催人奋进，动听，对劲儿 ...
“负面评价”词语，如：丑，苦，超标，华而不实，荒凉，混浊，畸轻畸重，价高，空洞无物 ...
“程度级别”词语，
“主张”词语

2. “英文情感分析用词语集”；现包含词语8945，其中包括如下6个子文件：

“正面情感”词语，如：happy，be jealous，admiration，consent，welcome，look forward to ...
“负面情感”词语，如：defy，disappointed，fear，criticize，regret，pull a long face ...
“正面评价”词语，如：good-looking，high-quality，effective，tranquility，safe and sound ...
“负面评价”词语，如：grotesqueness，inferior，expensive，expensively，brutal，false，gawky，low ...
“程度级别”词语，
“主张”词语

3. “中英文情感分析用词语集”现包含词语约17887。

“情感分析用词语集（beta版）”即日起向公众免费提供下载。
下载地址：《知网》情感分析用词语集（beta版）。



中文程度级别词语 219

1. “极其|extreme / 最|most” 69
百分之百
倍加
备至
不得了
不堪
不可开交
不亦乐乎
不折不扣
彻头彻尾
充分
到头
地地道道
非常
极
极度
极端
极其

情感词典3

```
########################################
        BosonNLP情感词典
########################################

BosonNLP情感词典是从微博、新闻、论坛等数据来源的上百万篇情感标注数据当中自动构建的情感极性词典。因为标注包括微博数据，该词典囊括了很多网络用语及非正式简称，对非规范文本也有较高的覆盖率。该情感词典可以用于构建社交媒体情感分析引擎，负面内容发现等应用。

在BosonNLP情感词典中，文本采用UTF-8进行编码，每行为一个情感词及其对应的情感分值，以空格分隔，共包括114767个词语。其中负数代表偏负面的词语，非负数代表偏正面的词语，正负的程度可以由数值的大小反应出。

格式：
[[词语]] [[情感值]]

例：
最尼玛 -6.70400012637
扰民 -6.49756445867
fuck... -6.32963390433
RNM -6.21861284426
wcnmlgb -5.96710044003
2.5: -5.90459648251
Fxxk -5.87247473641
MLP -5.87247473641
吃哑巴亏 -5.77120419579

来源：
http://bosonnlp.com
```

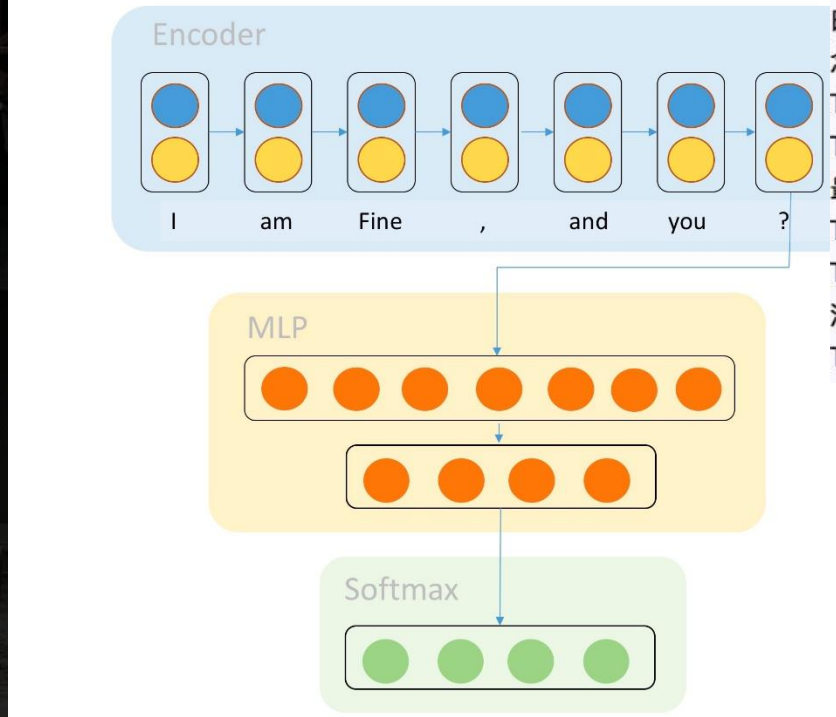
情感分数

1）每一个情感词对应不同的情感分数（存于SentiDict中）。

2）将情感词的位置记录下来，两个情感词之间存在的程度词和否定词决定权重W的大小和正负。

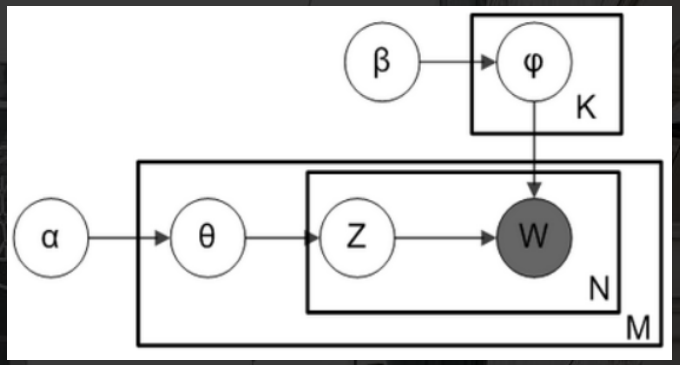
3）计算公式：Score = Score + W * (SentiDict[word])

神经网络模型



Type in a source sequence:
可惜了，红颜薄命。这些动不动就杀人的禽兽，下地狱吧
The emotion is: 生气😠
Type in a source sequence:
就这一个孩子，可悲的政策
The emotion is: 生气😠
Type in a source sequence:
人渣！！！
The emotion is: 开心😄
Type in a source sequence:
生在最安全国家却出现这样的事，丢人啊。
The emotion is: 其他
Type in a source sequence:
曰本AV片毒害天朝淫民！！太深了！！基本上这些淫民都是受日本
念置之脑后！！！不惜手段达到目的！！此就是成语（色迷心窍）
The emotion is: 生气😠
Type in a source sequence:
最近国内实时新闻都直接报？不用走官方？估计都去搞美国事了
The emotion is: 厌恶！
Type in a source sequence:
沙井那边搞化工的很多，这时候环保安检都要出来走几步了
The emotion is: 开心😄

主题模型 LDA



在LDA模型中，一篇文档生成的方式如下：

1° 从狄利克雷分布α中取样生成文档i的主题分布θ_i

2° 从主题的多项式分布θ_i中取样生成文档i第j个词的主题z_(i,j)

3° 从狄利克雷分布β中取样生成主题z_(i,j)的词语分布ϕ_(z_(i,j))

4° 从词语的多项式分布ϕ_(z_(i,j))中采样最终生成词语w_(i,j)

其中主题和词采用Gibbs抽样。

PART 5

阶段性成果

已完成工作

情感分析结果

已完成工作

相关文献阅读

多次小组讨论、与老师邮件沟通

搭建团队Github Repo、服务器

整理、人工生成数据

完成IP定位与文本地域探测程序

搭建LSTM情感分类器

多类情感词典尝试

样本数据计算与统计

初步数据可视化

情感分析结果

情感词统计

content	polar	pos-words	neg-words
结婚，找成都妹子结婚，要有本地户口的，未生育，和前任没有联系	1	7	3
最后一幅 我看到总府皇冠假日了 那个时候是成都最新潮的外来酒店品牌	1	8	2
现在峨眉雪，在乐山还有卖的	1	4	1
看来还是成都好	1	2	0

content	polar	pos-words	neg-words
福建人的骗，连有些地方银行都不批贷款。	-1	0	4
福建人就这点抵抗力也敢在网易混？[微笑]	-1	2	2
吧，忠告: 看病千万远离莆田甚至福建，那群牲畜什么钱都敢赚，毫无下限	-1	5	6
路者跟快餐店老板说过我还会差你这点钱啊这句话，我赌五毛是东北老铁	-1	4	9

情感分析结果

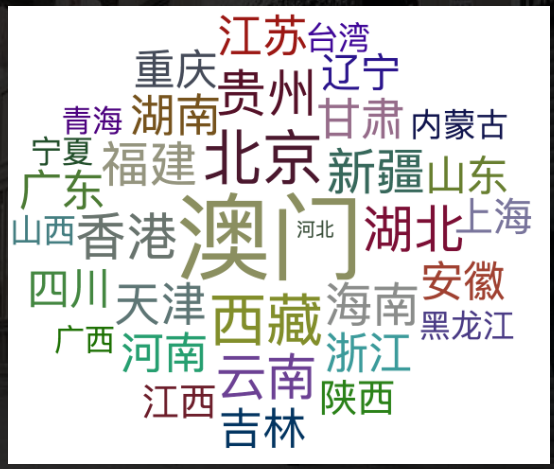
各省地域印象

	src	dist	word	polar	freq
955	陕西	陕西	贪官	-1	70
957	陕西	陕西	贪污	-1	28

	src	dist	word	polar	freq
76831	江苏	河南	勾引	-1	8
89109	湖北	河南	拐	-1	48
89119	湖北	河南	偷	-1	9

地域关注度词云

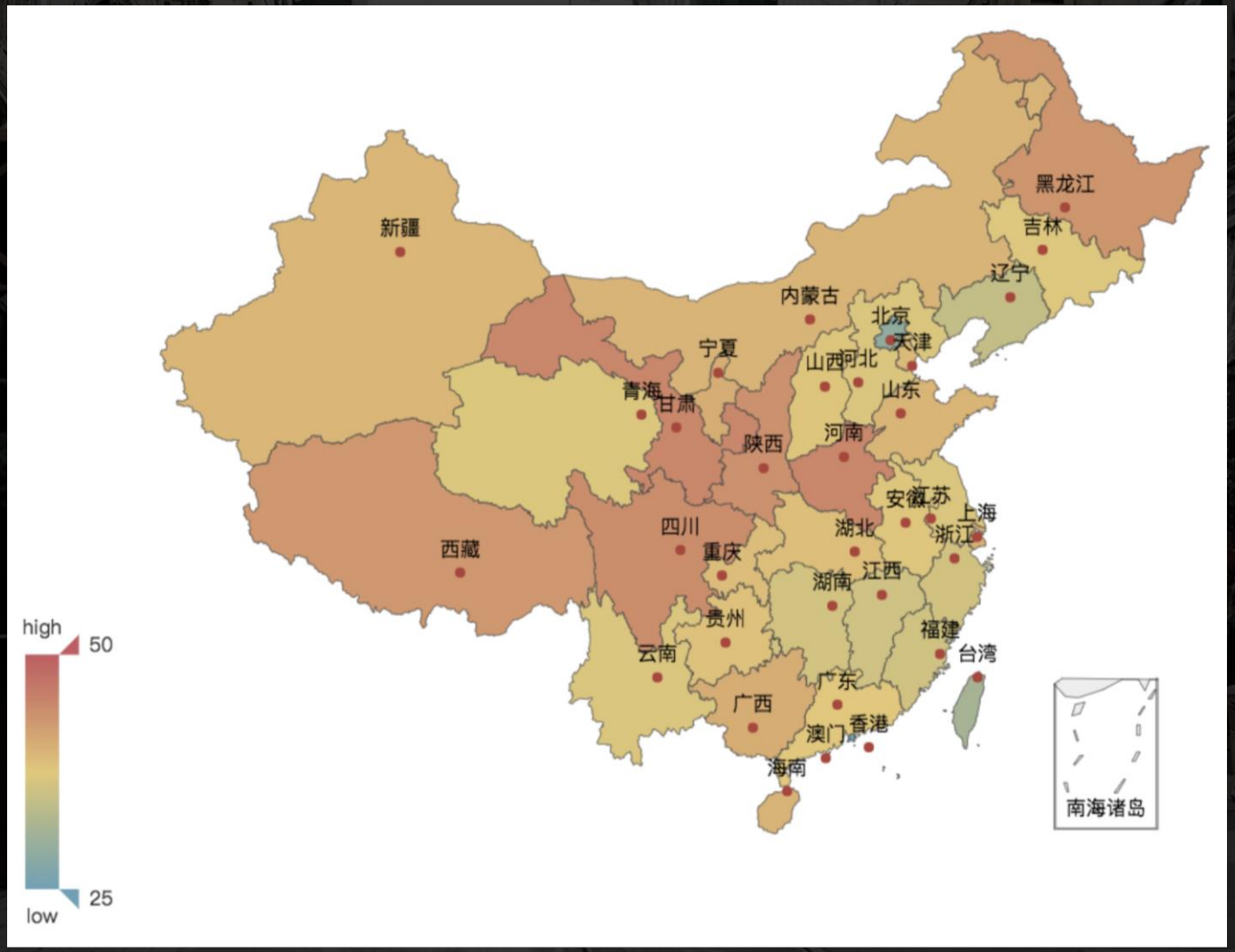
地域好评度（他评）词云



地域好评度（自评）词云

情感分析结果

全国地域负面评价图



PART 6

下阶段展望

遇到的问题

下阶段工作

预期结果

遇到的问题

1 数据

不充足，还没拿到具体的数据，
只有一个*sample*

遇到的问题

2 模型

地域探测

自己构造的地域辞典可以筛选出绝大部分，但仍然不够准确。

ip探测

调用的接口对于访问的频率有限制，尽管程序已经是多线程的，ip获取的时间仍然占据了90%及以上的时间。

情感分析

对于多地域的时候不太能区分哪个情感词针对哪个地域。筛选出来的情感词不够具有代表性。

遇到的问题

3 主观影响

来自我们

在筛选结果时可能会受到主观因素影响

来自网易

删帖和禁言

来自回帖者

人为刷楼

下阶段工作

数据来源

与老师沟通，获取到网易方数据

地域探测

完善地域字典，提高地域筛选精度
提高IP查询精度，或者连夜跑完所有IP地址

情感挖掘

提高神经网络模型精度
代表性情感词提取

数据可视化

预期结果

1. 各省份的典型刻板地域印象以及相互之间的印象评价
   a) 地域歧视语境下的各省(或代表性省份)的刻板印象标签
   b)各省(或代表性省份)之间的地域印象评分矩阵
   c)可视化(甚至可交互式)的省份“互黑”地图
   d)详述其他与预期现象或直观感受不符的发现
2. 地域形象时间–评分关系图，以及对于演变模式的分析
3. 对异常波动进行典例内容分析


Thanks!
微信二维码