

数字化记忆灾害 Report

人员分工

- 赵伟栋：数据收集整理，剪辑视频
- 陶文晖：后端代码编写，情绪分析，制作后端组视频
- 刘天恩：数据收集整理
- 邱天润：前端数据展示部分框架搭建，部分图表制作
- 郑青禾：前端地图部分展示制作
- 李玉金：前端部分首页制作

项目流程

- (1)数据爬取
- (2)清洗数据并做成能导入数据库的表格
- (3)统计分析数据、数据可视化

代码组成

后端部分

- 收集数据：**使用 **selenium** 模拟用户使用浏览器的过程，使用 **etree** 解析 html 结构，获取用户 id，性别，帖子内容，点赞评论转发，粉丝数，地址等。
- 情绪分析：** 使用 **hanlp** 语言模型，人工标签的方式进行训练，进行基本的正面，负面，中性的情绪识别。
- 整理数据：**将数据整理成为 CSV 数据集，便于前端分析调用。

前端部分

总述

我们的前端部分使用 **Vue** 和 **T-Design** 搭建，分为首页概览和各个具体分析功能两个部分，主要实现了以下界面和功能，来更好地交互式展示和分析我们所获取到的数据。

Home（主页部分）	App.vue	主页和总框架
Map（地图部分）	MapView.vue	地图面板

	mainMap.vue	地图组件
Analytics（分析部分）	Analytics.vue	分析部分界面
	data.js	处理数据
	histogram.js	柱状图组件
	piechart.js	饼图组件
	word-cloud.js	词云组件
AI Chat（大模型部分）	AI.vue	大模型交互界面
	claude.js	与大模型（Claude）交互

主页部分详述

在主页中，根据现有数据文件分析数据并可视化了热度趋势，集聚效应，讨论度，访问渠道和热点排行等具体数据，借助数据大屏的形式实现直观可视化。

分析部分详述

在这一部分中，我们使用 **d3** 的类库，通过代码实现了柱状图、饼图、词云的效果。同时，我们给出了一个控制套件，可以选择关键词、分析的内容，并支持选择对指定时间范围进行分析。数据来源为请求获取的 CSV 格式数据文件。

对于具体的内容，我们以基于人工智能模型的分词工具 **jieba** 为基础进行中文分词，增加了可以分析的内容丰富程度。

地图部分详述

引入 **echart** 中国地图包，使用热力图模式，便于展示数据在地区之间的差异。解析后端整理的 csv 文件，提取出时间、地点、类型等信息，并在前端可视化。

关键点和功能包括：

日期范围选择器： 使用了第三方的日期范围选择器，允许用户选择特定的时间范围进行数据可视化。

关键词选择下拉菜单： 提供了一个下拉菜单，用户可以从预定义的关键词列表（tagsList）中选择关键词。

地图容器： 使用 ECharts 库初始化地图容器，展示中国地图，并提供交互功能。地图上的各个省份具有数据点，数据点的值决定了该省份的颜色深浅。

异步数据获取： 通过异步请求获取 CSV 格式的数据文件，解析数据后更新地图图表，反映出选定关键词和日期范围内的信息量分布。

AI 大模型部分详述

在这一项目中，我们还使用了 **Claude** 大模型来对我们所获取的信息进行分析。

我们接入了 Claude API，并构建了一个问答界面，通过流式输入输出确保良好的用户体验。由于 token window 所限，我们暂时支持的是对某一个特定主题进行分析（这也是选择接入 Claude 模型而非 GPT 的原因，因为 GPT 所能提供的 token window 更小）；不过从后续效果展示截图所见，我们事实上也得到了令人满意的结果，并支持多轮问询。

数据部分详述

收集要求

1.关键词：

三组关键词之间的关系是 A 并 B 并 C。

主题：上海疫情	
时间段：2022.3.1-2022.6.30	
A 组	新冠 新型冠状病毒 疫情 上海市疫情防控 上海疫情 上海市新冠疫情防
B 组	防控政策 管理措施 联防联控机制综合组 治疗指南 实时动态 地图 感染情况 故事 讲述 口述 日记 知识 建议 解答 辟谣 专家 医生 中医 求助 互助 热线 救援 拼团 外卖 快递 事迹 纪实 宣布
C 组	团长 封校 可乐 咖啡 硬通货 刘耕宏 健身 合围区 移民 外国人 外籍人士 辟谣 压茬 历史无阳 研判 无差别 保供 哄抬物价 必需品 小区 楼组长 全力以赴 点式复工 有序解封 有序放开 有限流动 关门不歇业 解封不解防 自治 隔离 核酸 居家 网课 网购 出境 入境 团购 封城 封控 解封 出入通行证 口罩 防护服 莲花清瘟 酒精 武汉病毒 SARS 非典 李文亮 方方日记 世界卫生组织 WHO 疾控 CDC 卫健委 方舱 火神山 雷神山 新闻发布会 驰援 志愿者 病例 密接 传染链 群体免疫 清零 静默 非必要不 健康码 场所码 绿码 行程码 消杀 复工复产 应收尽收 应检尽检 大白 流调 疫苗 时空伴随 转运 网格 抢菜 囤物资 张文宏 钟南山 熔断 后遗症 白肺 呼吸机 ECMO 肺部纤维化 健康宝 毒王 超级传播者 闭环 拭子 混检 单人单检 三天三检 落地检 风险地区 恶意返乡 偷渡 管控 气溶胶 人传人 无害化处理 无症状 港口 医学观察 境外输入 抑郁 焦虑 恐慌 未

	知 疲惫 医疗资源挤兑 ICU PCR 重症 变异 毒株 旅游 娱乐 餐饮 可防可控 抗 疫 躺平 物价 星号 自我防护 勤洗手 污名化 假新闻 种族歧视 报复性消费 周 边游 防疫指挥小组 科学防疫 群防群控 无接触配送 滞留 共存 居委会 孕妇 女 性 婴幼儿 青少年 青年 中年 老年 病患 慢性病 残障 残疾
排除 词	判决书 股份有限公司 控股有限公司 人才引进资格 道指跌 纳指跌 标普 500 跌 中概股普跌 资源股 加密货币 国防 汽车之家

2.数据抓取所需字段：

1)序号	11)原文内的标签
2)标题/微博内容	12)原文内容的情绪（如果很复杂可提前告知）
3)原文链接 （原文附带的链接）	13)信源地域
4)日期	14)精准地域
5)来自（微博客户端/移动端/网页版.....）	15)原创/转发
6)媒体类型	16)涉及词（抓取方案中涉及的词标亮）
7)原文作者	17)转发数
8)认证类型（政府/企业团体/个人认证/普通用户）	18)评论数
9)粉丝数	19)点赞数
10)原文	20)图片/视频/....

效果展示

C1	C2	C3	C4	C5	C6	C7	C8	C9
author	sex	official organization certification	creators manage members	Entrance	origin	fans num	ip address	Content
中国食品安全报社	男	中国食品安全报社官方微博	true	微博网页版	true	4.7万	北京	【国家游140
能开心就好	女	false	false	HUAWEIMate40E	true	585	江苏	游140
黑暗中的鲨鱼	女	中国作协会员，河北作协签约作家，新浪读书原创	false	荣耀305G	true	47.8万	河北	《您知
海淀融媒	女	北京市海淀区融媒体中心官方微博	true	微博weibo.com	true	9.1万	null	【造福
布里亚Bria	女	false	false	iPhone	true	519	美国	上个月
安徽之声	男	安徽广播电视台安徽综合广播FM95.5、AM936官方	true	微博网页版	true	15.1万	null	【泗县
哈尔滨交通广播	女	哈尔滨交通广播官方微博	true	微博weibo.com	true	83.2万	null	【关于
新浪科技	女	新浪科技官方微博	true	微博weibo.com	true	2369.2万	null	【“躲
无锡交通广播	男	无锡交通广播官方微博	true	iPhone客户端	true	75.1万	null	【无锡
无锡日报	女	无锡日报官方微博	true	微博weibo.com	true	99.7万	null	【无锡
新京报	男	新京报官方微博	true	微博weibo.com	true	4719.7万	null	【】据
领事闲谈	男	外事一员	false	微博网页版	true	38.4万	安哥拉	【伦敦
北京人应该知道的北	男	知名本地博主本地资讯博主（北京）视频自媒体	false	null	true	26.7万	北京	【】今
北京晚报	男	北京晚报官方微博	true	微博weibo.com	true	1791.3万	null	【】6月
Hermann中国零售	男	false	false	微博视频号	true	65万	上海	星期五
北京交通广播	男	北京交通广播官方微博	true	三星S20FE5G	true	352.6万	null	【北京
首都健康	男	北京市卫生健康委员会官方微博	true	iPhone客户端	true	512.6万	null	开展首
小小浣熊	男	false	false	荣耀V8脱影而出	true	494	青海	美国总
沈阳地铁第一时间	女	沈阳地铁报《地铁第一时间》官方微博	true	HarmonyOS设备	true	70.8万	null	【】据
辽宁卫视	男	辽宁卫视官方微博	true	微博weibo.com	true	324.6万	null	【丹东
西充交警	男	四川省西充县公安局交警大队	true	荣耀9青春版全屏四摄	true	2万	null	【】据
平安河西	男	天津市公安局河西分局官方微博	true	iPhone客户端	true	77万	null	·国际移
凡闻资讯	男	凡闻资讯官方微博	true	微博网页版	true	1.6万	浙江	【：为
热浪MissW	女	false	false	iPhone客户端	true	112	四川	图【庐
上海黄浦	男	上海市黄浦区人民政府新闻办官方微博	true	微博视频号	true	35.6万	null	·国际移

部分搜集数据

新冠 假新闻_None_None.csv
新冠 健身_None_None.csv
新冠 全力以复_None_None.csv
新冠 共存_None_None.csv
新冠 关门不歇业_None_None.csv
新冠 出入通行证_None_None.csv
新冠 勤洗手_None_None.csv
新冠 卫健委_None_None.csv
新冠 历史无阳_None_None.csv
新冠 压茬_None_None.csv
新冠 变异_None_None.csv
新冠 口述_None_None.csv
新冠 合围区_None_None.csv
新冠 后遗症_None_None.csv
新冠 周边游_None_None.csv
新冠 地图_None_None.csv
新冠 复工复产_None_None.csv
新冠 外卖_None_None.csv
新冠 女性_None_None.csv
新冠 娱乐_None_None.csv
新冠 婴幼儿_None_None.csv
新冠 孕妇_None_None.csv
新冠 实时动态_None_None.csv
新冠 宣布_None_None.csv
新冠 密接_None_None.csv
新冠 小区_None_None.csv
新冠 居委会_None_None.csv
新冠 居家_None_None.csv
新冠 应收尽收_None_None.csv
新冠 建议_None_None.csv
新冠 快递_None_None.csv
新冠 恐慌_None_None.csv
新冠 恶意返乡_None_None.csv
新冠 感染情况_None_None.csv

部分收集数据，None 表示是否进行高级搜索

数字化记忆灾害

首页视图

地图索引

统计与词云

AI 问答

5. 内容主题细致,从政策执行细节到企业支持措施等各层面进行上报。

6. 态度积极向上,希望引导公众遵守防疫规定共同付出,推进工作有序进行。

所以总体来说,这些微博数据来源广泛,内容细致,形式多样,态度积极,侧重各地区实际防疫工作,体现了新冠疫情期间不同部门和机构共同参与防疫工作的情况。

User:

总结内容所涉及的最关键的话题

Claude:

复制全文

这些微博数据最核心和关键的话题包括:

1. 新冠疫情防控工作。作为主线,各地政府部门和企业机构都将重点放在新冠疫情防控工作上,宣传政策实施情况。
2. 重点防控领域,如出入管理、流动人口管控、通行证制度等,以切断传播链。
3. 基层执行动态,重点关注社区和街道等一线实施情况,如核酸检测工作等。
4. 生活物资保障,重点企业支持保供物流和生活物资供应保障工作。
5. 重点群体服务,如妇女儿童和老年人等弱势群体体的关照。
6. 政策通告与调整,及时下发和响应地方在防疫工作中的具体问题。
7. 舆论引导,呼吁公众共同参与并支持政府防控工作。
8. 企业参与支持,提供物流保障或物资支持等形式参与防控工作。

以上几个点,可以说就是这些微博数据聚焦的核心主题,反映了防疫工作的各个重要层面。

必需品

请输入 (Ctrl+Enter 发送)

发送

设置

大模型问答页面

数字化记忆灾害

地图索引

数据列表

统计与词云

AI 问答

2020-12-01

- 2024-01-31

关键词: 保供



2020-12-01

- 2024-01-31

关键词: 保供

保供

健身

压茬

可乐

咖啡

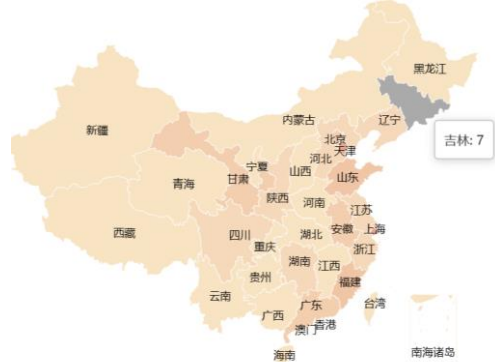
团长

封校

小区

研判

移民



选择关键词

鼠标移到地图上显示具体数值

项目总结

在数字化记忆灾害项目中，我们团队充分发挥了各自专业领域的的能力。通过使用 Selenium 进行数据采集和 etree 进行 HTML 结构解析，我们成功获取了用户信息和帖子内容，并利用 hanlp 语言模型实现了情感分析。在前端开发方面，Vue 和 T-Design 的应用使得我们能够以直观而交互式的方式展示数据，同时利用 d3 和 ECharts 等库创造了多种图表效果。引入 Claude 大模型为项目增加了深度，通过 Claude API 进行问答分析，我们成功探索了特定主题的细节。这次经验让我们深刻认识到了团队协作和数据处理在项目中的关键作用，同时为我们提供了宝贵的学习机会，为未来的数据分析和项目开发打下了坚实基础。