天津大学
Tianjin University

疫情全回顾——

疫情期间舆情演化分析与热点事件模式挖掘

李泽宇（队长）　　**陈诺**

李思思　　**王腾**

贾世超　　**张加万（指导老师）**

ChinaVis 2020 Data Challenge

利用大数据分析技术和可视分析方法协助发现病毒传染源、监测疫情发展、调配救援物资，为疫情防控中的分析、指挥和决策提供有效依据和指南。

- 疫情时空态势分析
- 疫情传播模式分析
- 疫情预测与舆情监测
- 新冠病毒病理研究
- 疫情潜在影响与次生灾害分析

ChinaVis 2020 Data Challenge

利用大数据分析技术和可视分析方法协助发现病毒传染源、监测疫情发展、调配救援物资，为疫情防控中的分析、指挥和决策提供有效依据和指南。

- **疫情预测与舆情监测**：利用可视分析技术，预测疫情发展趋势与关键节点、分析社交媒体话题与情感的动态演变、对社会舆情进行态势感知。

疫情全回顾——

疫情期间舆情演化分析与热点事件模式挖掘

数据、分析任务、分析流程

数据 2019-12-23 ~ 2020-05-10

	热文数据	热点事件数据	疫情新闻列表
数量	300 * 140 天 = 42000	1781	690
来源	榜单：清博大数据 文章：今日头条	清博大数据	知微数据
核心字段	文章标题、内容 发布日期 阅读数	事件标题、概述、热词 发生日期 14天情感趋势 14天热度趋势	新闻标题 发生日期
类型	时序文本	时序文本，时间序列	时序文本

分析任务

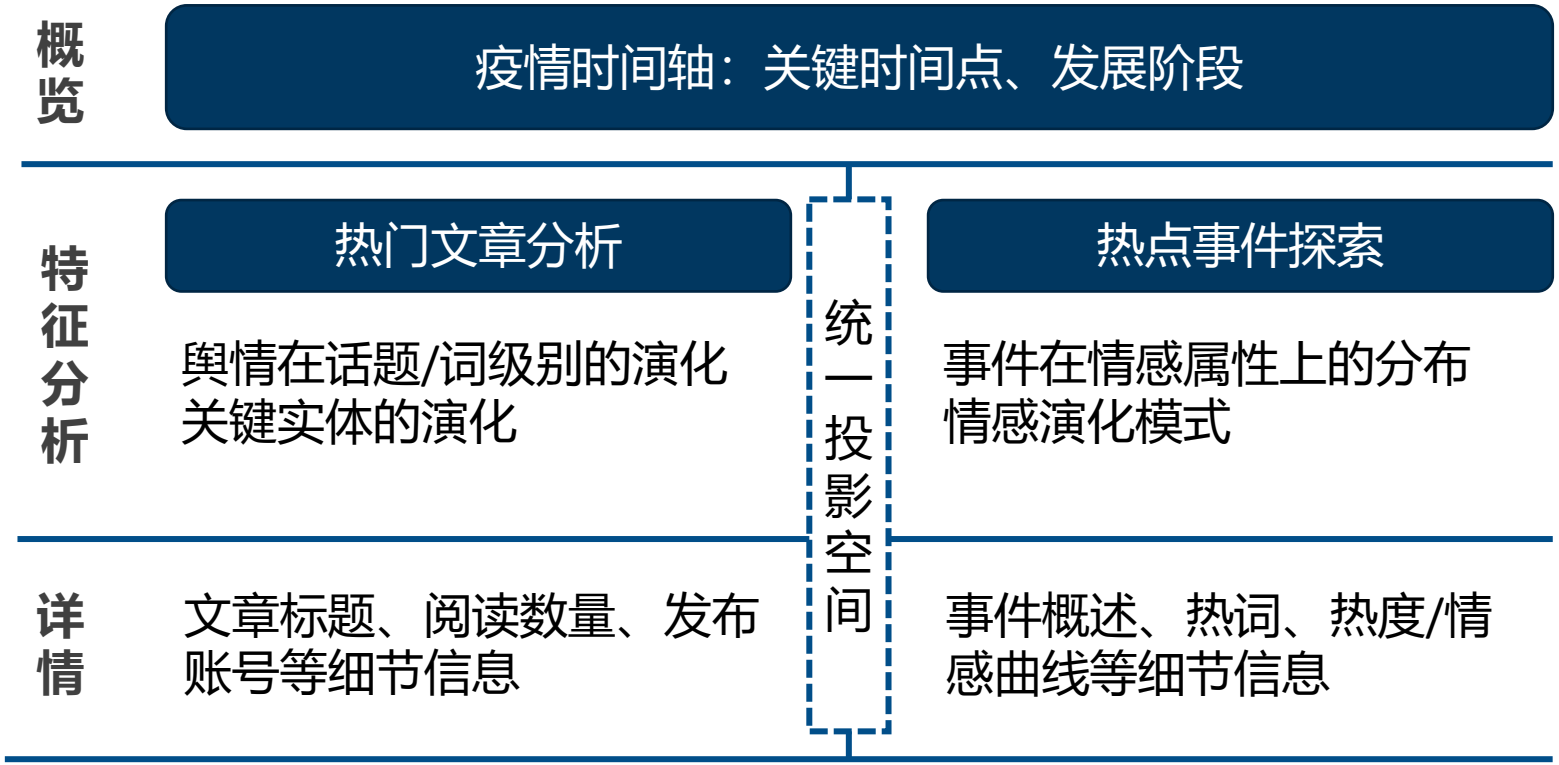
1. 明确疫情发展阶段，识别重大热点事件

2. 从多个角度（文本内容/实体）和多个尺度（话题级/词级）探究舆论关注点的转变

3. 找出在情感相关维度上特殊的热点事件

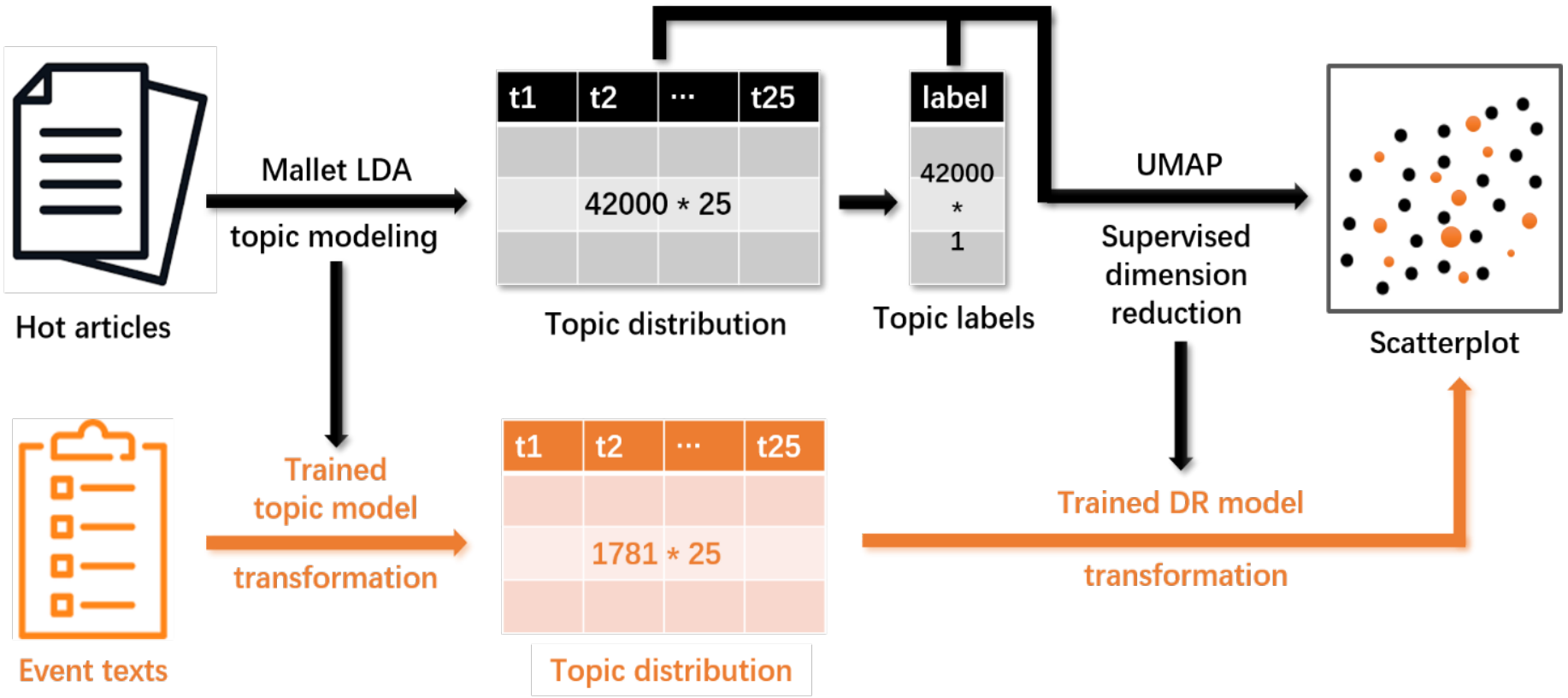
4. 挖掘热点事件的热度演化模式

分析流程


概览
疫情时间轴：关键时间点、发展阶段
特征分析
热门文章分析
舆情在话题/词级别的演化
关键实体的演化
统一投影空间
热点事件探索
事件在情感属性上的分布
情感演化模式
详情
文章标题、阅读数量、发布账号等细节信息
事件概述、热词、热度/情感曲线等细节信息

数据模型与计算

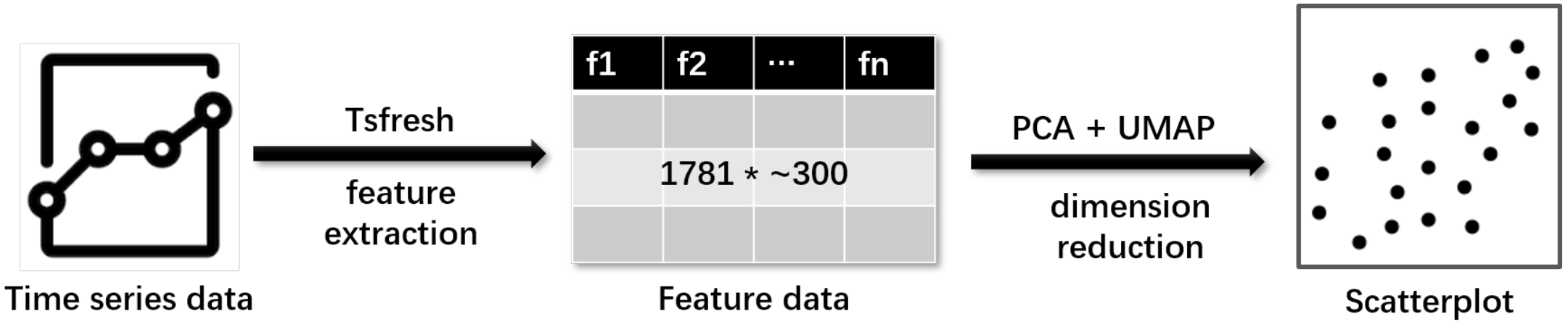
统一投影空间（主题地图）的构造



与事件情感相关的度量准则的计算

度量准则	计算公式	描述
总体情感	$\frac{\sum_i (p_i + m_i/2)}{\sum_i (p_i + m_i + n_i)}$	以“修正”后的正面情感比例来度量，值越大，总体情感越偏正面。“修正”是为了避免将“中性占比很大，正面占比很小”这种事件识别为负面事件
争议程度	$\frac{(P+N) - abs(P-N)}{P+M+N}$	正负情感值占总情感值的比例越大，且两者差异越小的事件认为争议程度越大
波动程度	$[\sigma(\tilde{p}) + \sigma(\tilde{m}) + \sigma(\tilde{n})]/3$	以三种情感比例序列的标准差的平均值来度量。值越大，情感波动程度越大
反转情况	$\begin{cases} s = sign(indexOf(\max(d)) - indexOf(\min(d)) \\ s * abs(\max(d) - \min(d)), & \text{if } \max(d) * \min(d) < 0; \\ undefined, & else \end{cases}$	对某事件，若情感比例差异序列d即有正值也有负值，则认为该事件存在情感反转，此时最大值和最小值之间的差异越大，意味着反转强度越高。最大值后于最小值出现，认为是负面转正面，反之认为是正面转负面

事件热度序列投影空间的构造



案例分析

案例1：探究舆论热点演化

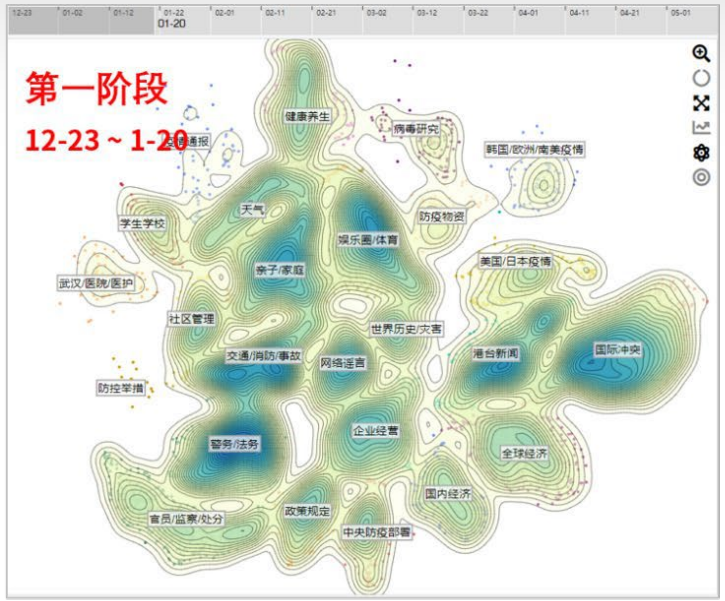
疫情发展的四个阶段

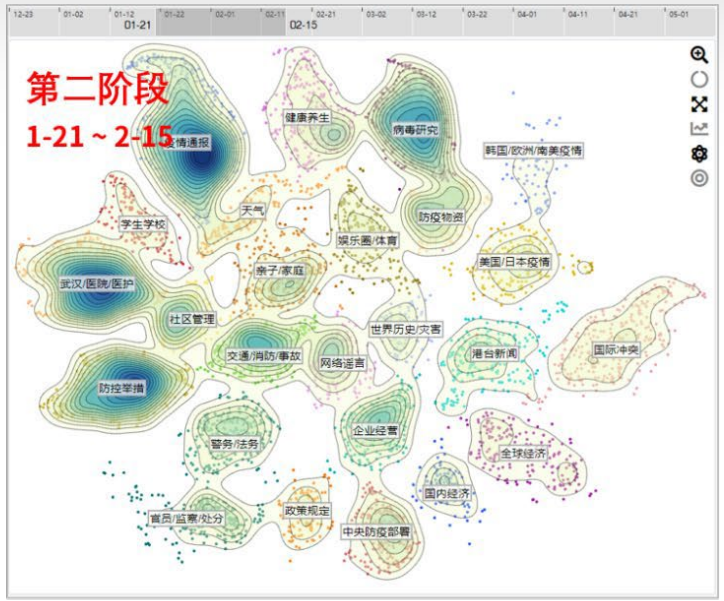

第一阶段
第二阶段
第三阶段
第四阶段
标题列表
武汉封城
山东任城监狱
以及浙江十里丰监狱
爆发了大规模聚集性感染
1.23
a
武汉"封城"：全市城市公交、地铁、轮渡、长途客运暂停运营；无特殊原因，市民不要离开武汉，机场、火车站离汉通道暂时关闭。
浙江、湖南、广东相继启动重大突发公共卫生事件一级响应。
b
2.18
2.19
2.20
山东任城监狱发生新冠肺炎疫情，截至2月20日24时，确诊病例207例，其中干警7例，服刑人员200例。
2月20日0-24时，浙江新增确诊28例，其中27例来自十里丰监狱。
3.10
c
10日起，意大利将在全国范围内实行封城禁令，学校停课时间延长至4月3日。
习近平抵武汉考察新冠肺炎疫情防控工作。
武汉/医院/医护
韩国/欧洲/南美疫情
美国/日本疫情
01-27
d
03-16
e
02-20
f

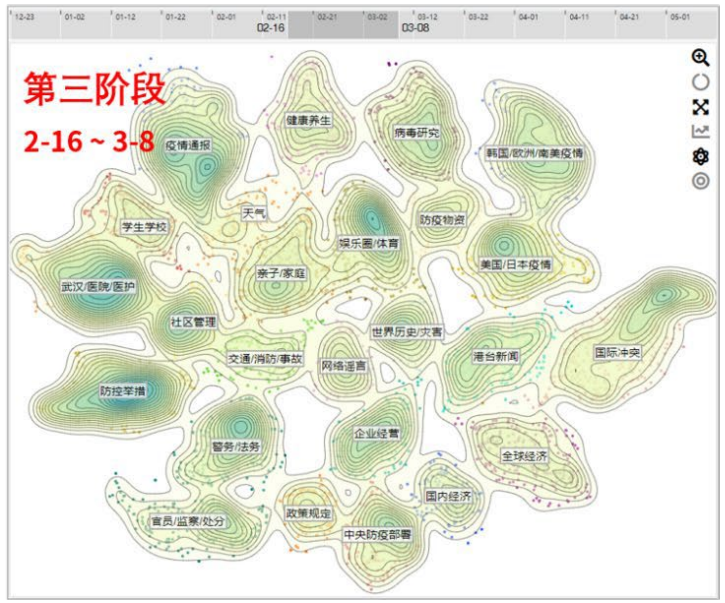
疫情发展的四个阶段

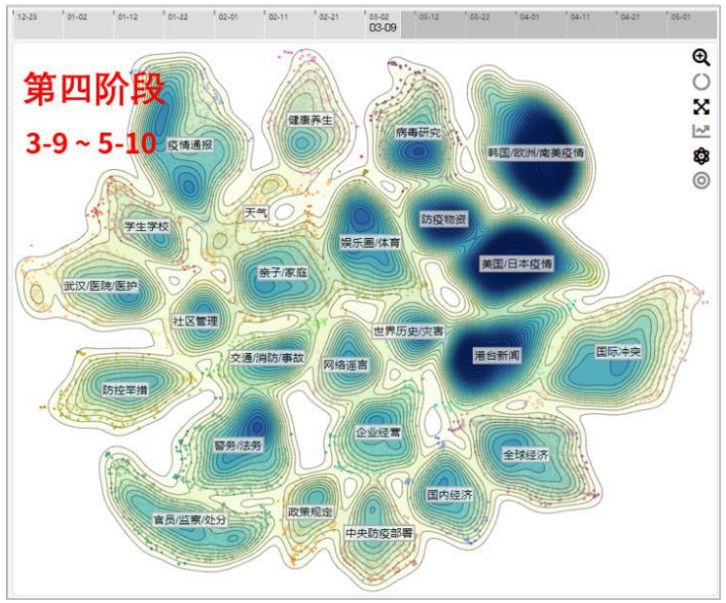

第一阶段
第二阶段
第三阶段
第四阶段
标题列表
武汉封城
山东任城监狱
以及浙江十里丰监狱
爆发了大规模聚集性感染
武汉/医院/医护
韩国/欧洲/南美疫情
美国/日本疫情
01-27
02-20
03-16
d
e
f

疫情发展的四个阶段

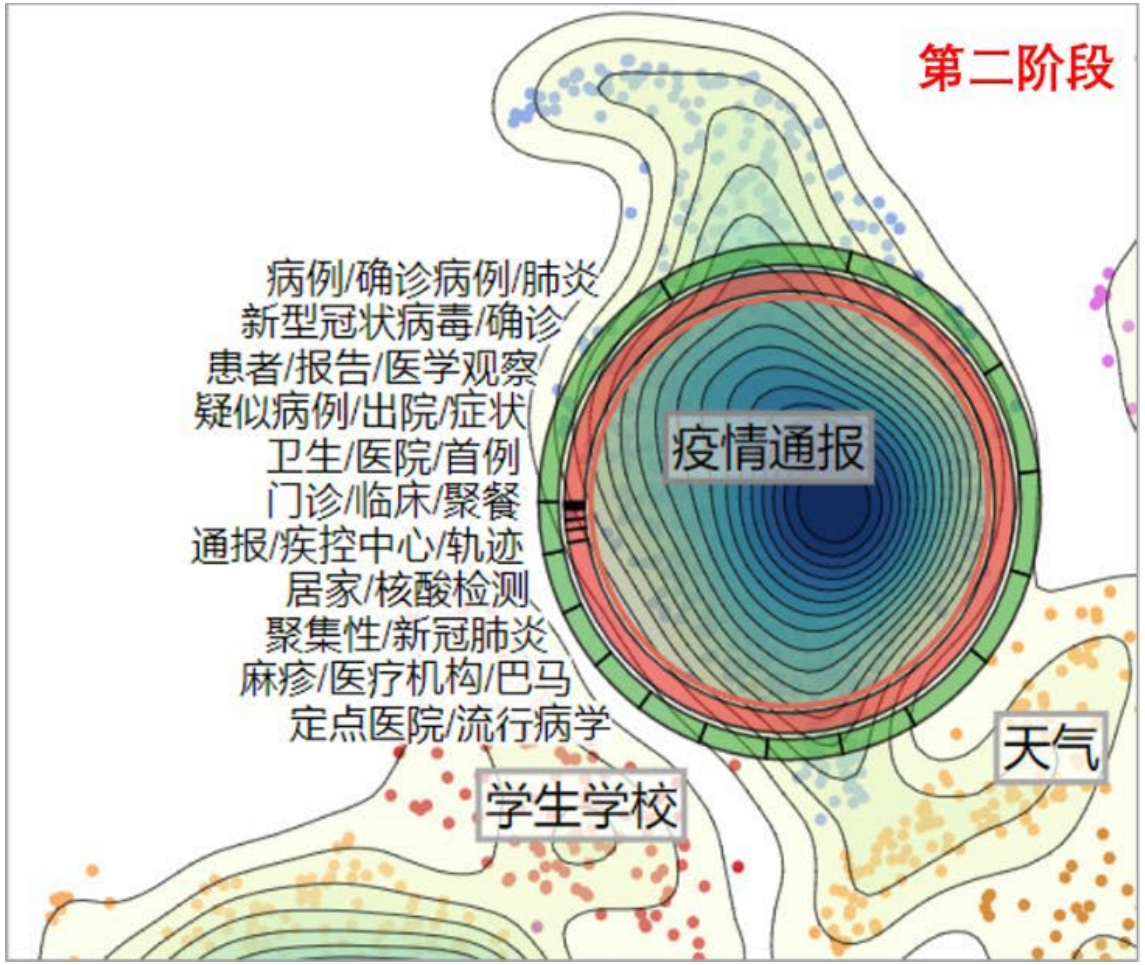

第一阶段
12-23 ~ 1-20
健康养生
病毒研究
韩国/欧洲/南美疫情
天气
学生学校
娱乐圈/体育
防疫物资
亲子/家庭
美国/日本疫情
武汉/医院/医护
社区管理
世界历史/灾害
交通/消防/事故
网络谣言
港台新闻
国际冲突
防控举措
企业经营
警务/法务
全球经济
国内经济
官员/监察/处分
政策规定
中央防疫部署

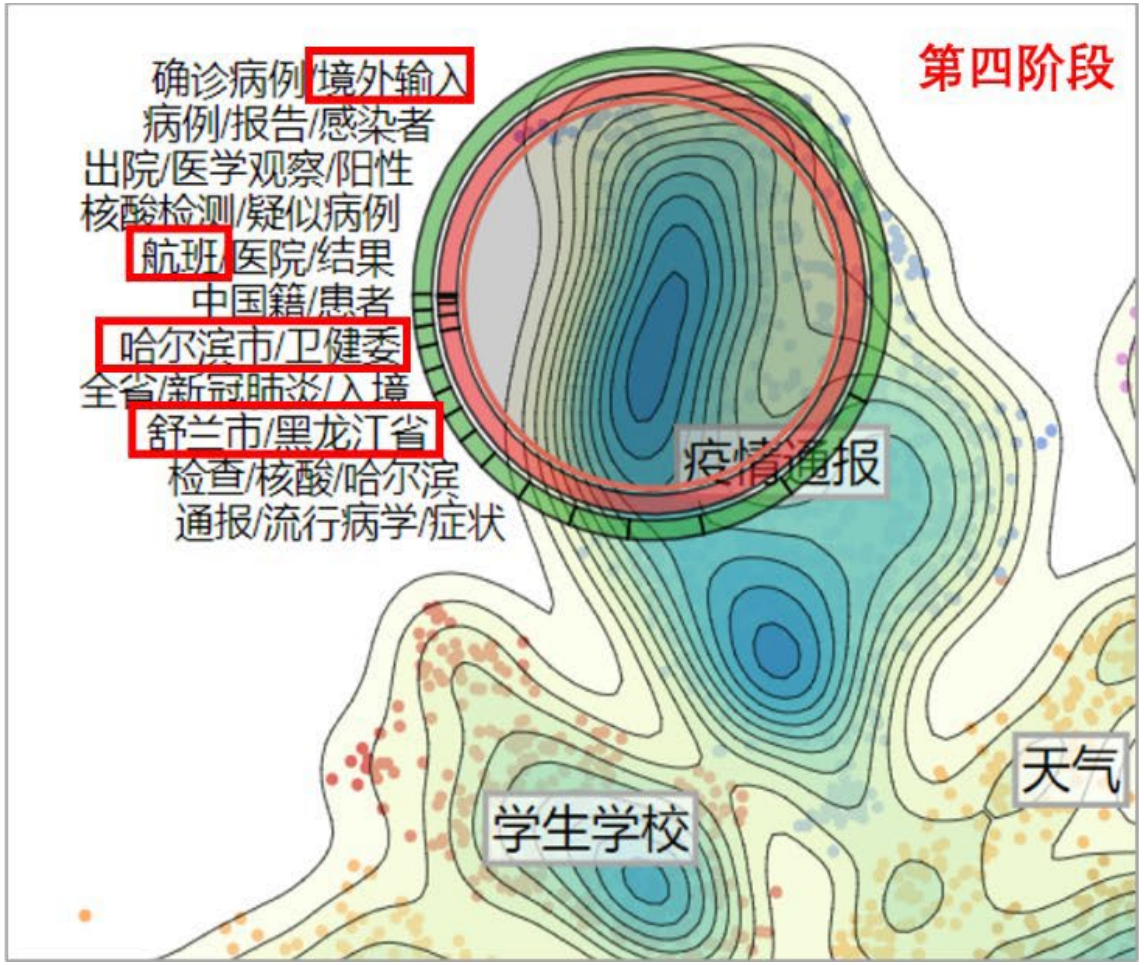

第二阶段
1-21 ~ 2-15
健康养生
病毒研究
韩国/欧洲/南美疫情
天气
学生学校
娱乐圈/体育
防疫物资
亲子/家庭
美国/日本疫情
武汉/医院/医护
社区管理
世界历史/灾害
交通/消防/事故
网络谣言
港台新闻
国际冲突
防控举措
企业经营
警务/法务
全球经济
国内经济
官员/监察/处分
政策规定
中央防疫部署


第三阶段
2-16 ~ 3-8
健康养生
病毒研究
韩国/欧洲/南美疫情
天气
学生学校
娱乐圈/体育
防疫物资
亲子/家庭
美国/日本疫情
武汉/医院/医护
社区管理
世界历史/灾害
交通/消防/事故
网络谣言
港台新闻
国际冲突
防控举措
企业经营
警务/法务
全球经济
国内经济
官员/监察/处分
政策规定
中央防疫部署

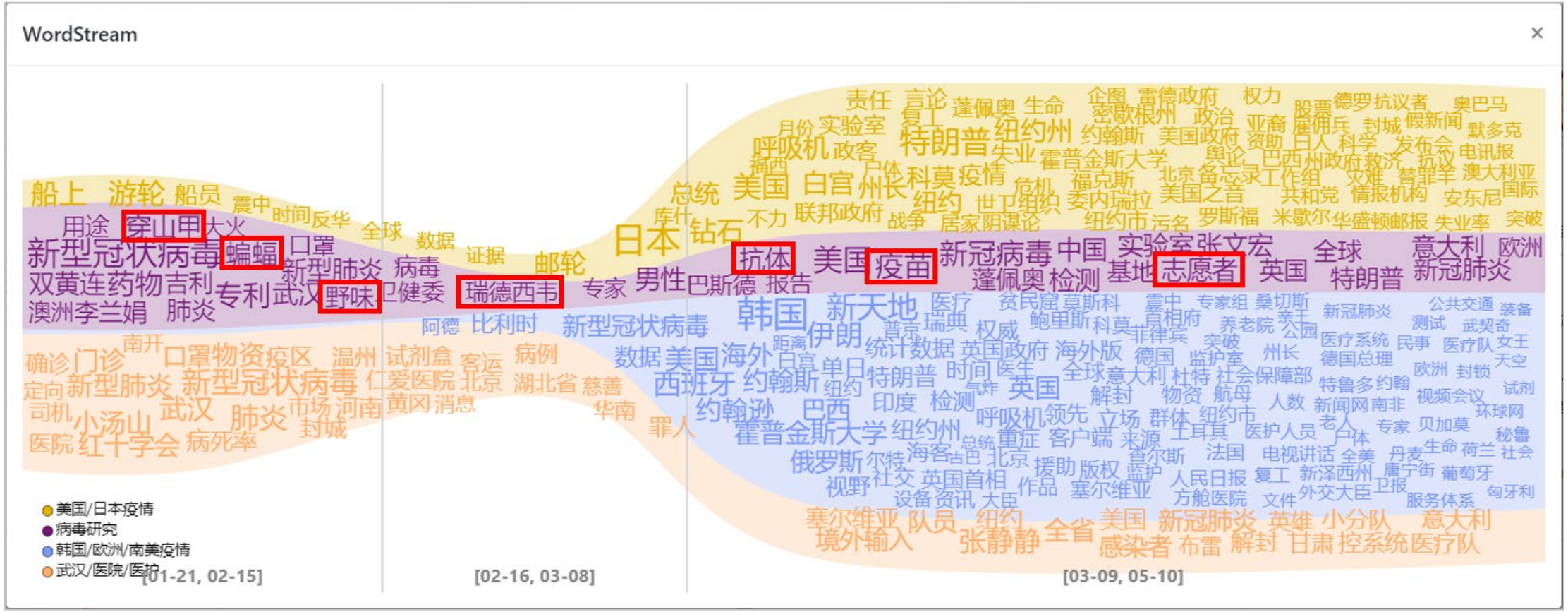

第四阶段
3-9 ~ 5-10
健康养生
病毒研究
韩国/欧洲/南美疫情
天气
学生学校
娱乐圈/体育
防疫物资
亲子/家庭
美国/日本疫情
武汉/医院/医护
社区管理
世界历史/灾害
交通/消防/事故
网络谣言
港台新闻
国际冲突
防控举措
企业经营
警务/法务
全球经济
国内经济
官员/监察/处分
政策规定
中央防疫部署

理解“疫情通报”话题在第二和第四阶段的差异

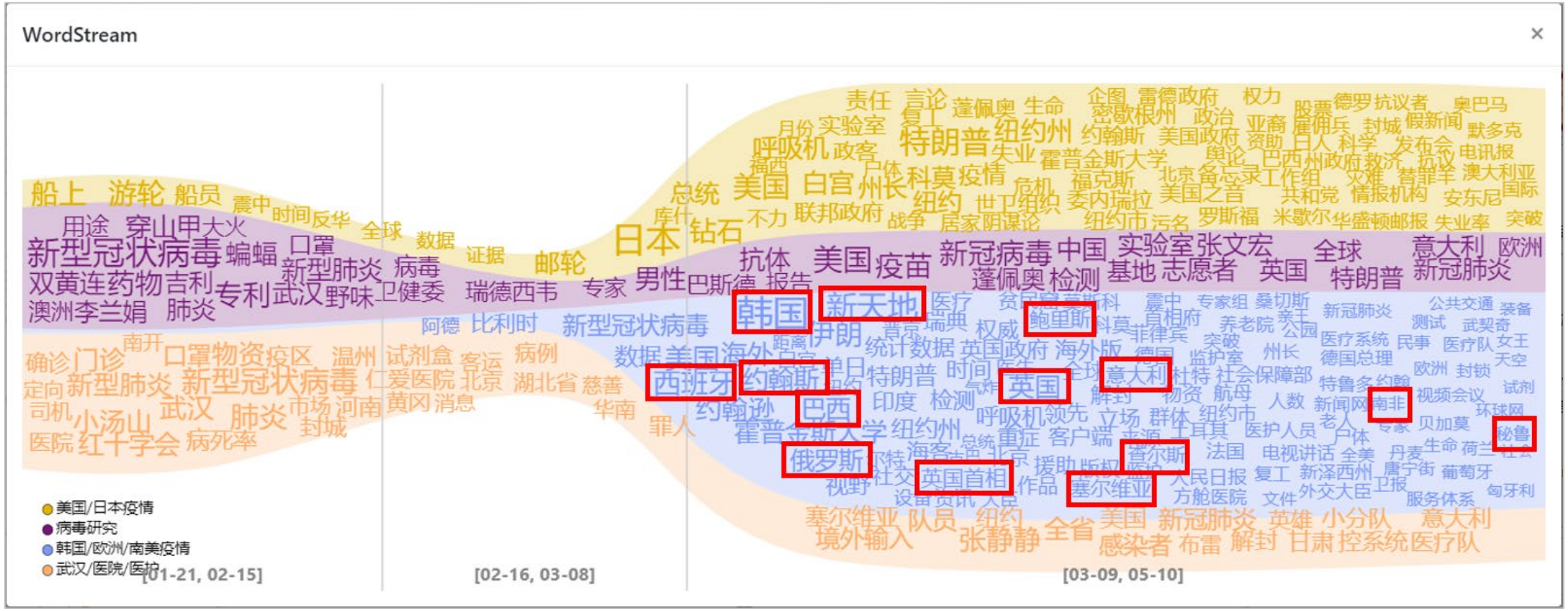




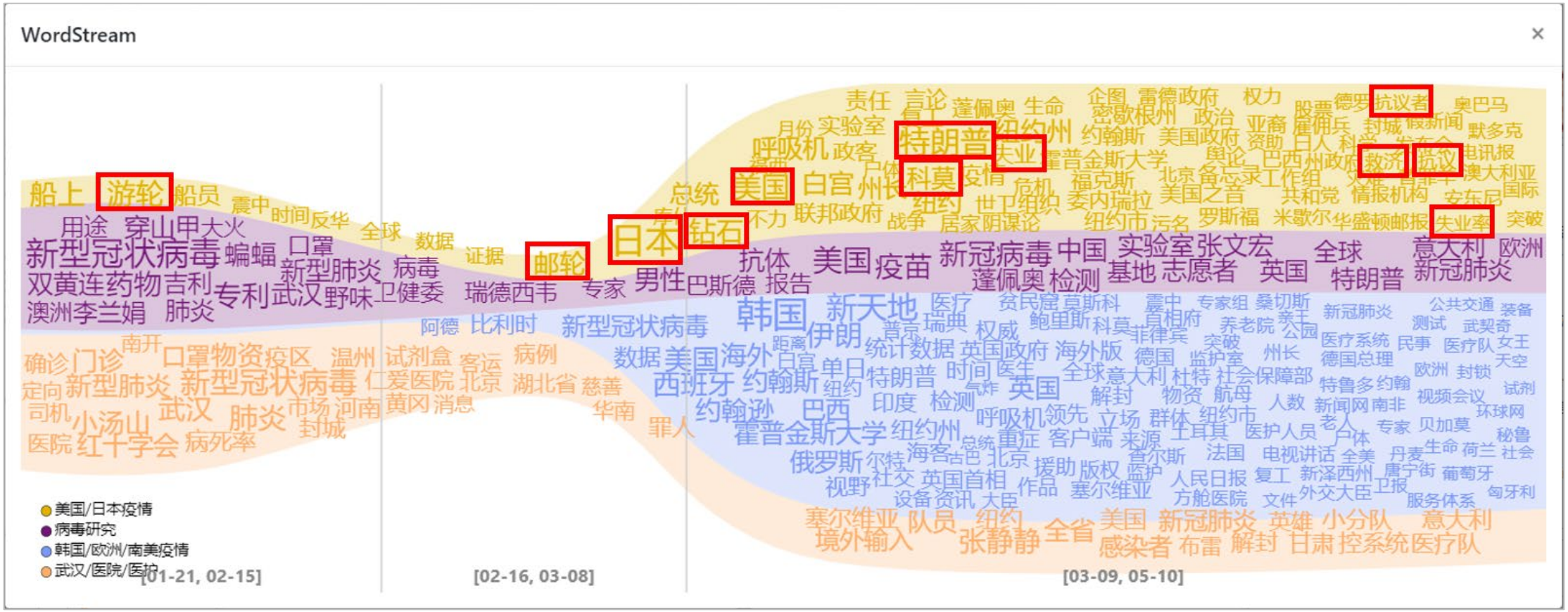
“病毒研究”话题的演化


WordStream
船上 游轮 船员 震中 时间 反华 全球 数据 证据 邮轮 日本 钻石 抗体 美国 疫苗 新冠病毒 中国 实验室 张文宏 全球 意大利 欧洲
用途 穿山甲 大火 口罩 新型肺炎 病毒 瑞德西韦 专家 男性 巴斯德 报告 蓬佩奥 检测 基地 志愿者 英国 特朗普 新冠肺炎
新型冠状病毒 蝙蝠 双黄连药物 吉利 专利 武汉 野味 卫健委 澳洲李兰娟 肺炎
韩国 新天地 伊朗 新型冠状病毒 数据 美国 海外 西班牙 约翰斯 约翰逊 巴西 霍普金斯大学 俄罗斯
确诊 门诊 口罩物资 疫区 温州 试剂盒 客运 病例 新型肺炎 新型冠状病毒 仁爱医院 北京 湖北省 慈善 华南 罪人
定向 司机 小汤山 武汉 肺炎 市场 河南 黄冈 消息 封城 医院 红十字会 病死率
美国/日本疫情
病毒研究
韩国/欧洲/南美疫情
武汉/医院/医护
[01-21, 02-15]
[02-16, 03-08]
[03-09, 05-10]

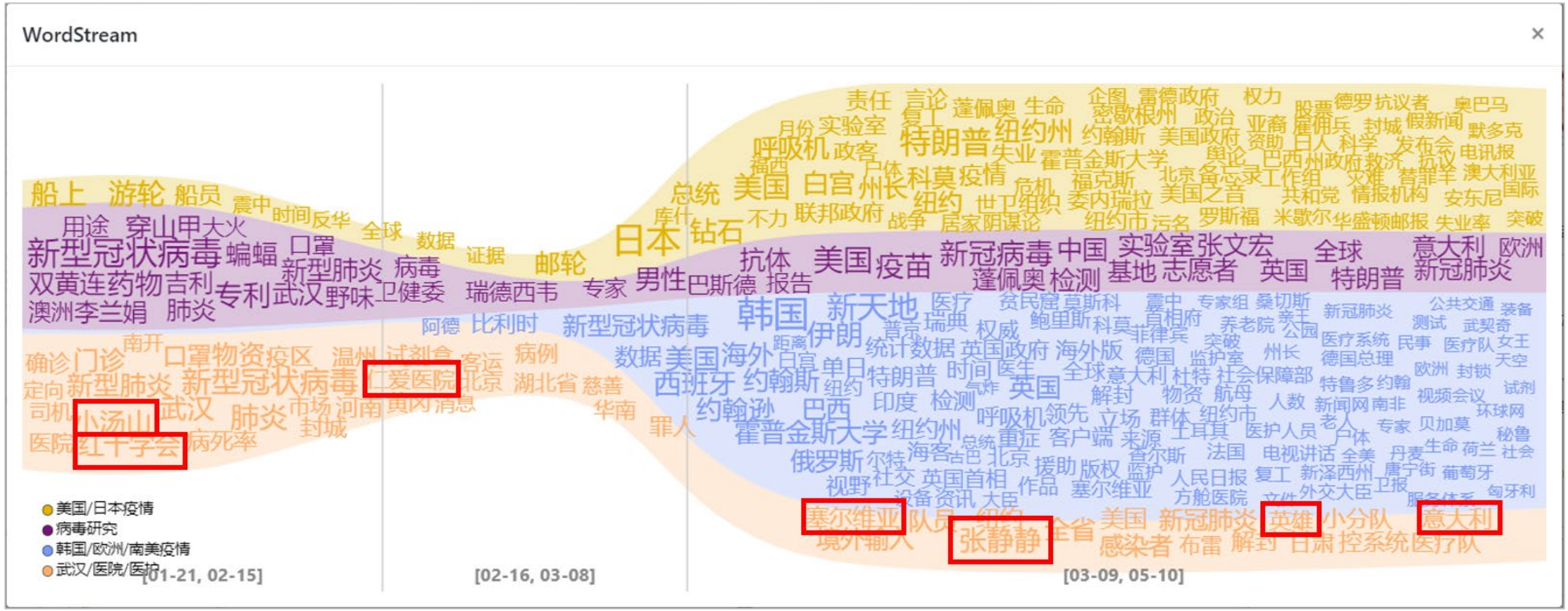
“韩国/欧洲/南美疫情”话题的演化



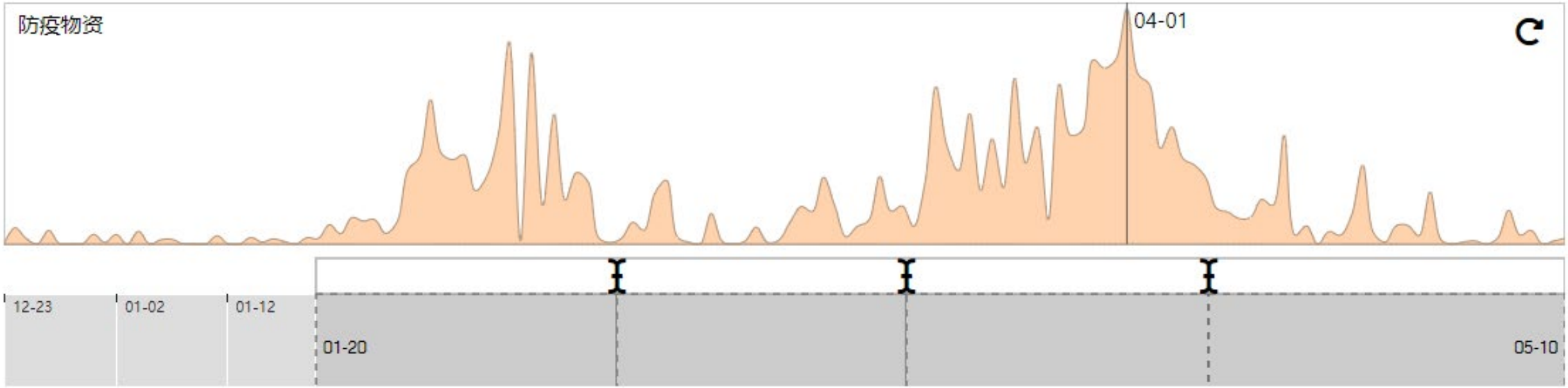
“美国/日本疫情”话题的演化



“武汉/医院/医护”话题的演化



“防疫物资”话题的演化



“防疫物资”话题的演化

描述词

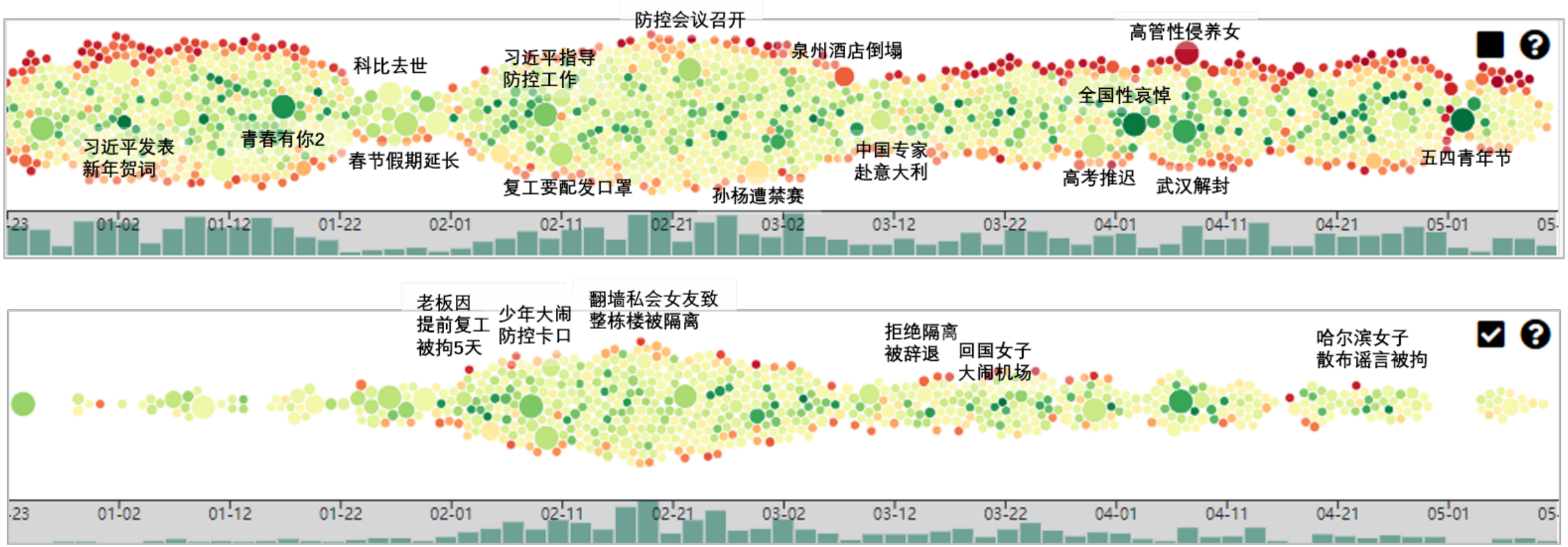
01-20 - 02-15	02-16 - 03-12	03-13 - 04-08	04-09 - 05-10
武汉	韩国	美国	渔船
口罩	青田	留学生	货机
悉尼	伊朗	呼吸机	越南
澳洲	华联	英国	出口
车站	巴基斯坦	法国	尼泊尔
变态	青田县	回国	海域
全淹	意大利	王女士	扬中
泳池	蝗灾	入境	古巴
风暴	巴林	北京	索马里
新州	空运	乘客	启迪
肺炎	瑞士	国际航班	顺水
警报	齐齐哈尔	天津	阿根廷
医用	橘子	特朗普	巴厘岛
湖北	文文	目的地	格力
药店	动物	祖国	上海

机构

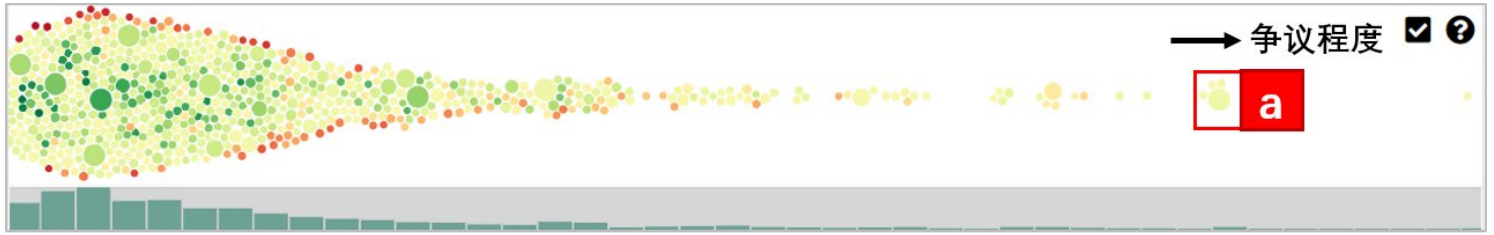
01-20 - 02-15	02-16 - 03-12	03-13 - 04-08	04-09 - 05-10
卫健委	外交部	外交部	外交部
红十字会	中国政府	国务院	商务部
国务院	国务院	民航局	海关总署
外交部	卫生部	中国政府	卫生部
京东	世卫组织	卫生部	环球网
央视	环球网	海关总署	中国政府
波音	韩国政府	环球网	民航局
世界卫生组织	商务部	新华社	新华社
环球网	中国外交部	世卫组织	路透社
淘宝	世界卫生组织	卫健委	波音
武汉大学	民航局	白宫	卫健委
中国政府	央视	霍普金斯大学	国务院
世卫组织	中新社	央视	世界卫生组织
阿里	新华社	商务部	上海浦东国际机场
人民政府	卫健委	美国政府	中国外交部

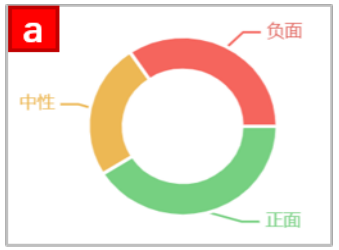
案例2：探究事件在话题和情感上的分布以及事件的演化模式

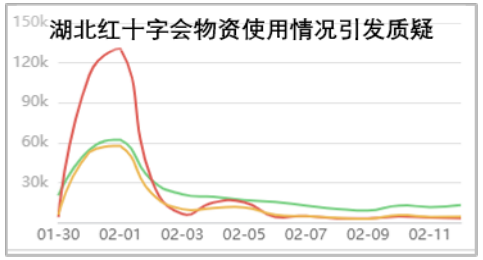
全部事件和疫情相关事件在时间维度上的分布

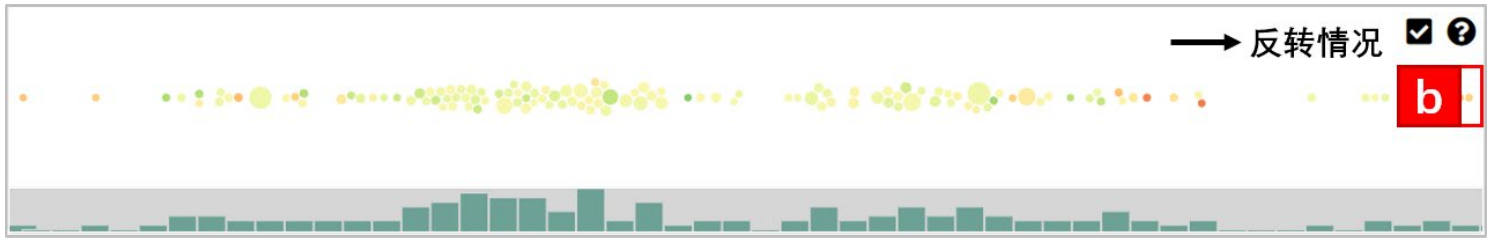


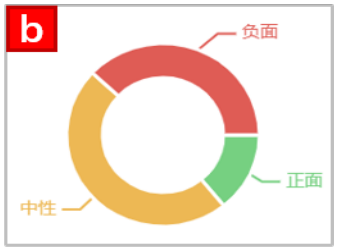
事件在三个情感属性下的分布

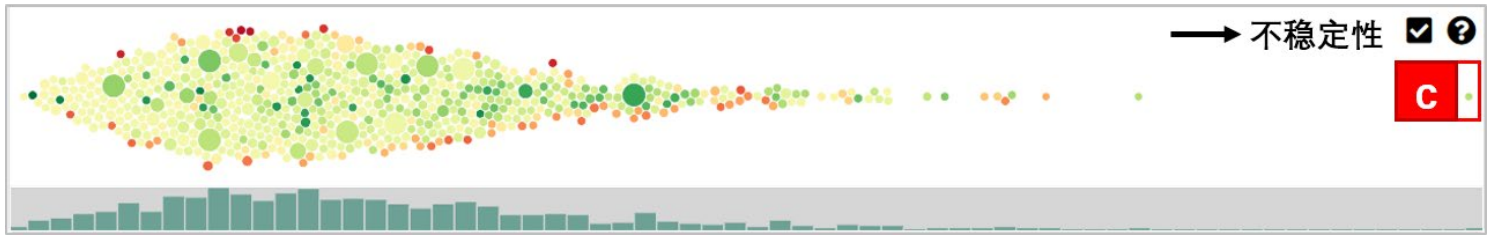


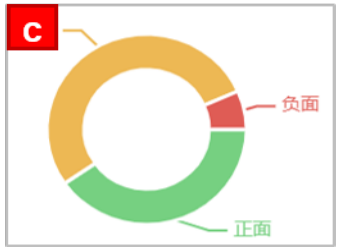


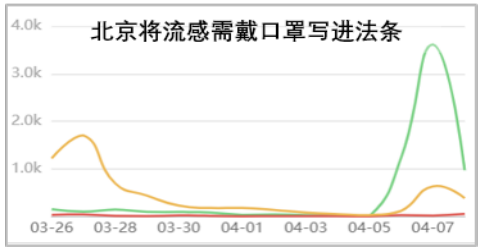




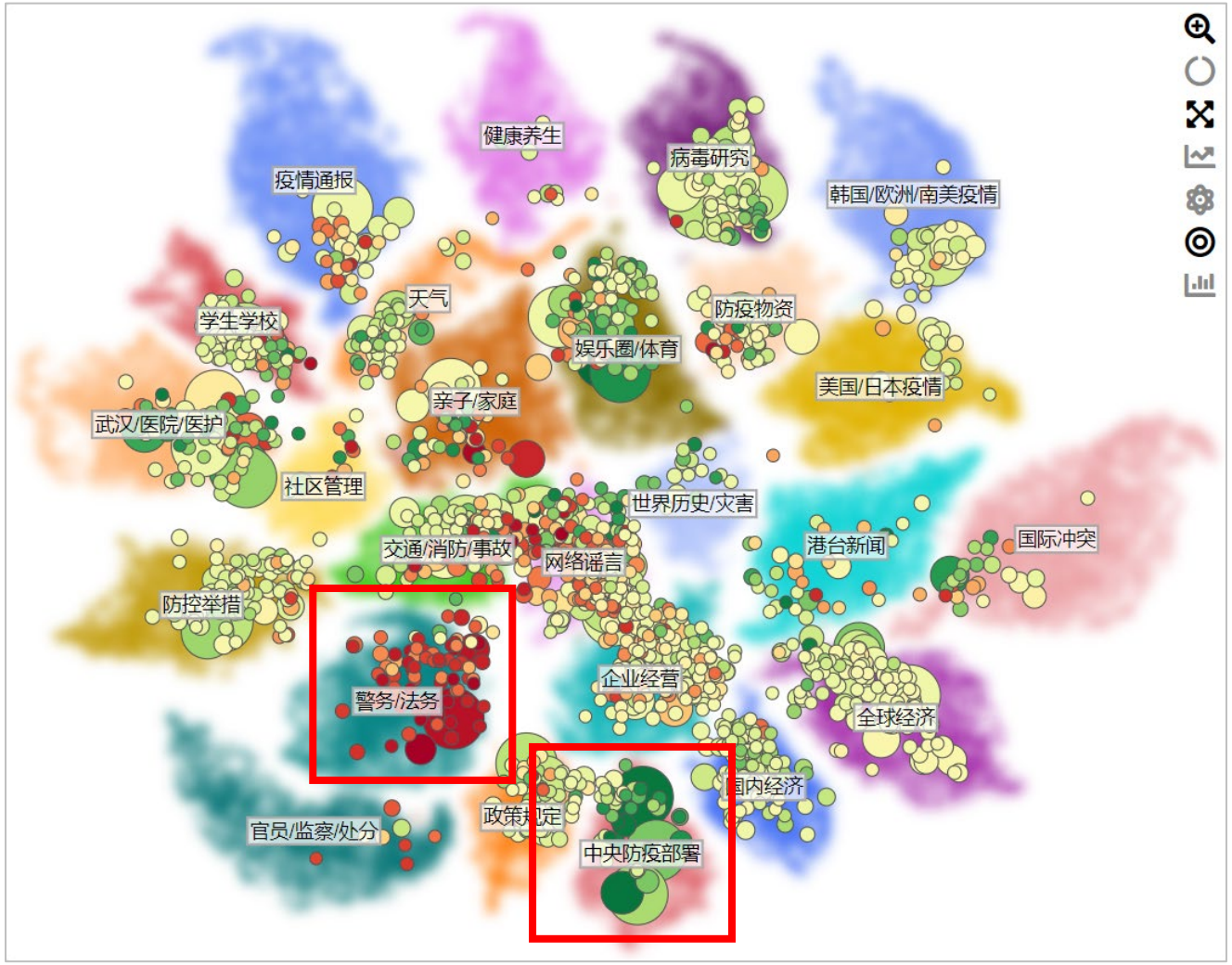




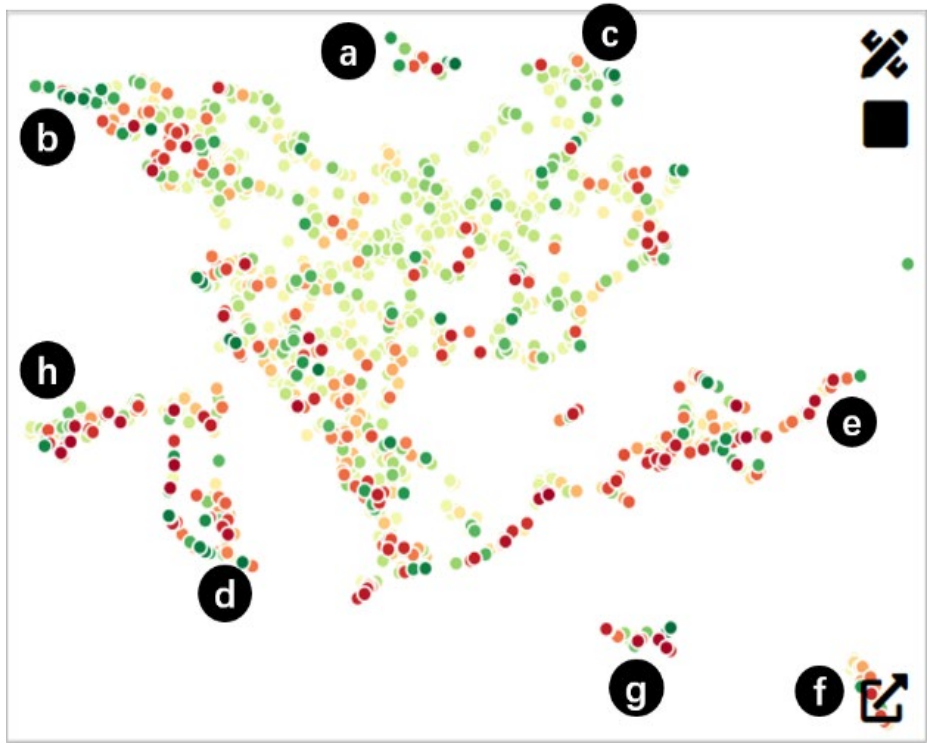


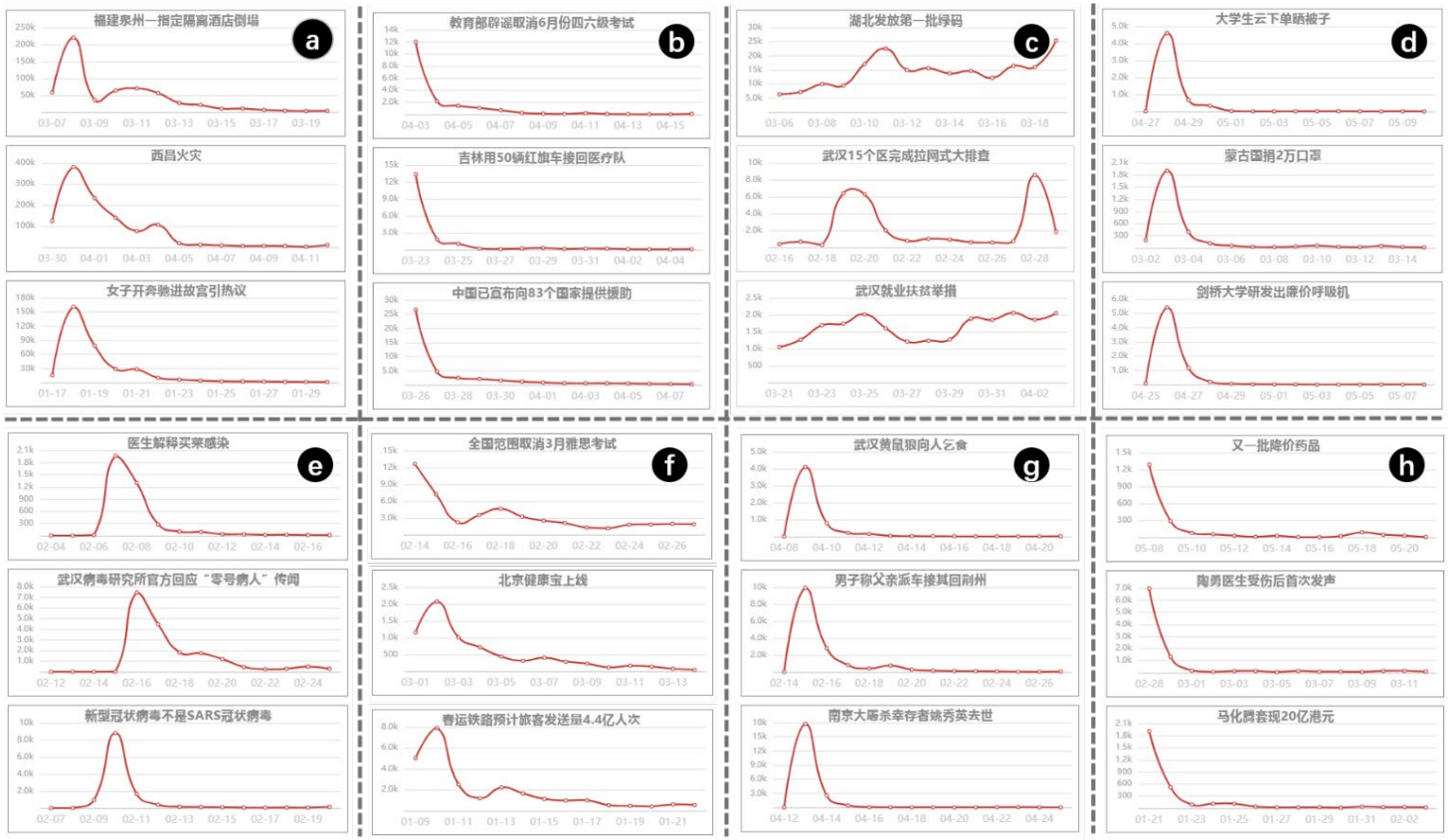


事件在各个主题的分布情况



理解事件投影空间


a
c
b
h
e
d
g
f


福建泉州一指定隔离酒店倒塌
a
教育部辟谣取消6月份四六级考试
b
湖北发放第一批绿码
c
大学生云下单晒被子
d
西昌火灾
吉林用50辆红旗车接回医疗队
武汉15个区完成拉网式大排查
蒙古国捐2万口罩
女子开奔驰进故宫引热议
中国已宣布向83个国家提供援助
武汉就业扶贫举措
剑桥大学研发出廉价呼吸机
医生解释买菜感染
e
全国范围取消3月雅思考试
f
武汉黄鼠狼向人乞食
g
又一批降价药品
h
武汉病毒研究所官方回应“零号病人”传闻
北京健康宝上线
男子称父亲派车接其回荆州
陶勇医生受伤后首次发声
新型冠状病毒不是SARS冠状病毒
春运铁路预计旅客发送量4.4亿人次
南京大屠杀幸存者姚秀英去世
马化腾套现20亿港元

讨论与总结

伸缩性

计算的伸缩性： 只有描述词提取需要在线计算，对于 G^2 statistics 方法来说，采样是一个既能大幅减少计算量又不显著降低描述词质量的方法。

可视化的伸缩性：Canvas 或者 WebGL 绘制技术。当然，采样也是一种缓解办法。

流数据

基于模型的transform，新近的文章和事件可以实时、无缝地融入已有图中。两点需要注意：

- 新发生的事件和新闻由于尚未积累起足够的影响力从而可能会被忽略
- 新数据可能不会对已训练的模型产生实质影响

本质上并没有解决流数据的问题，仍然需要定期手动更新。

总结

我们的系统可以被广泛应用在舆情状态感知、关键节点识别、话题演变分析、事件演化范式探索上。

天津大学
Tianjin University

谢谢!

李泽宇（队长）　　陈诺

李思思　　王腾

贾世超　　张加万（指导老师）

lzytianda@tju.edu.cn