挖掘，分析与可视化展现

多维数据建模透视全球疫情舆情生态

南京大学 信息管理学院 信息管理与信息系统
高雨婷，董沁怡，沈姝彤，吴迪，王敬

目录》
CONTENTS

1 小组成员介绍

高雨婷

小组队长，负责推文ID和医学对话数据数据清洗，以及CBC新闻舆情数据和医学对话建模与可视化工作等，参与数据分析思路构造，和研究报告撰写。

董沁怡

负责丁香园国内新闻数据的数据分析，推特推文传播数据集数据分析和医学对话数据集英文部分数据分析，对新闻语料数据进行数据挖掘，参与文献资料搜集工作。

沈姝彤

在小组研究前期对疫情总体时间序列数据进行可视化归纳，医学对话数据集数据清洗和谣言、推特发文文本主题的数据分析。参与论文后期的标引。

王敬

负责CBC新闻为代表的海外新闻舆情的数据建模，推特ID数据集时间序列分析，新闻语料数据和医学对话数据英文部分数据分析，参与论文后期校对。

吴迪

负责丁香园国内新闻数据、推文传播数据、新闻语料数据和医学对话数据的文本主题分析，参与论文后期校对。

2 研究思路和方法

文本挖掘与主题建模

运用多种自然语言处理方法，
高频词语分析，LDA主题建模，
层次聚类，文本词向量化等

时间序列与可视化

疫情舆情量化趋势走向与时间节点探究，地理、词频分布与词语共现网络多维可视化

词典型情感分析

定量算法定性探究文本情感态度和情感关键词

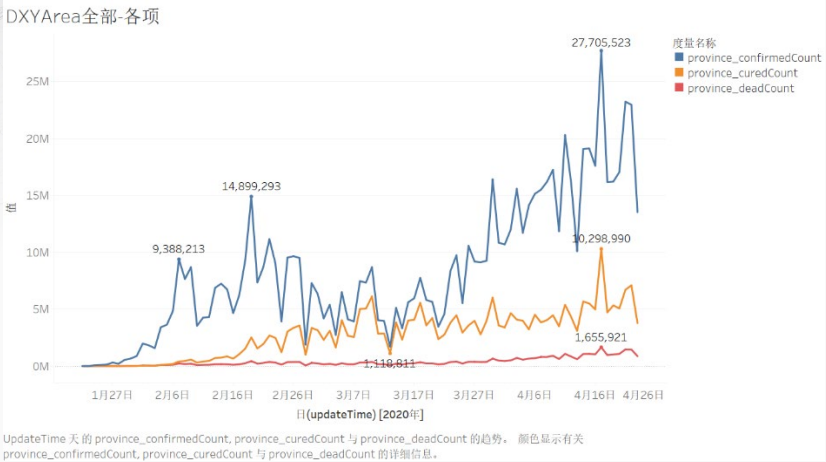
探索性数据分析

数据分布特点，数据类型，数据与主题建模的可行性挖掘

机器学习与统计计量学

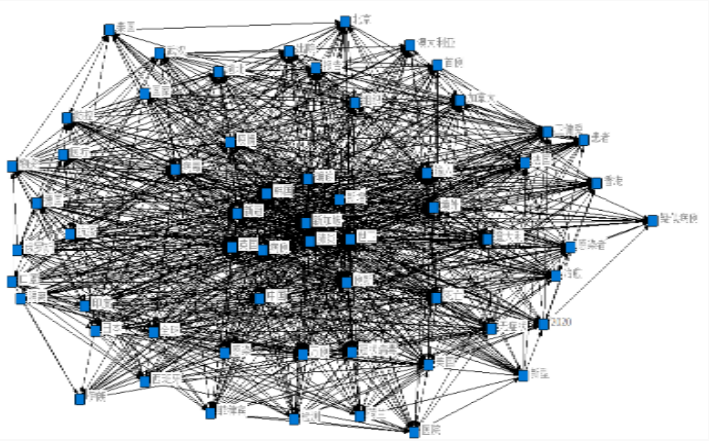
根据统计学原理，对文本进行关联规则与关联度分析，挖掘疫情与舆情，舆情之间的潜在演化规律。

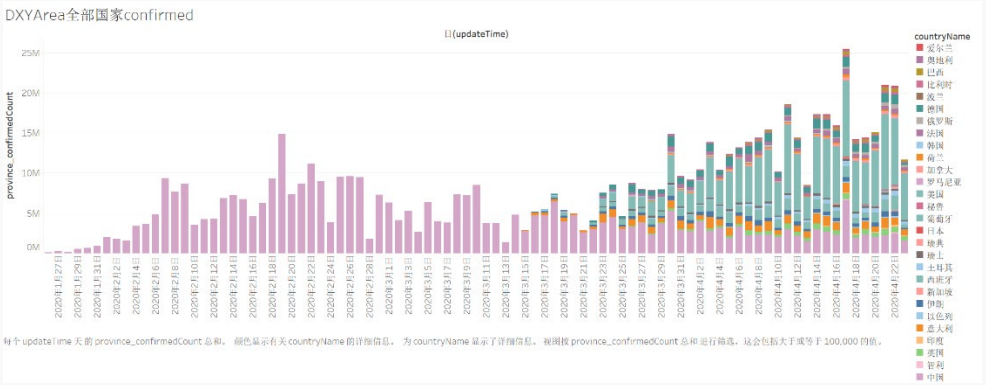
3 研究成果与分析



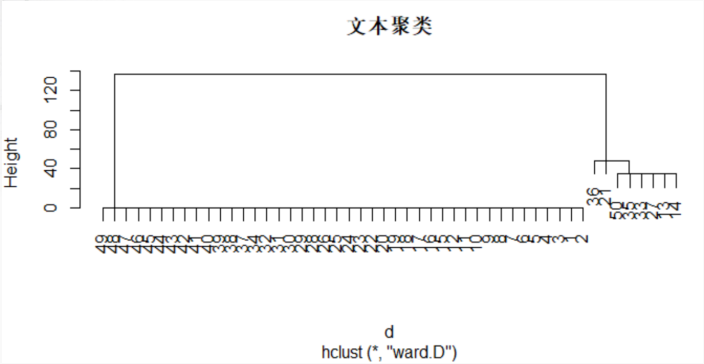
数据可视化

时间序列分布数据
时间地区分布数据
词频分布可视化
词频共现网络





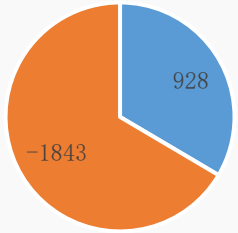
3 研究成果与分析



数据建模

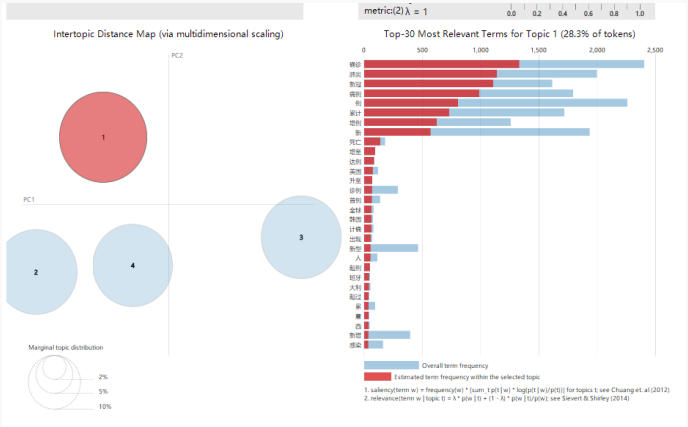
文本层次聚类
LDA主题文本建模
词典型情感分析
关联规则
词频逆文本分析
……

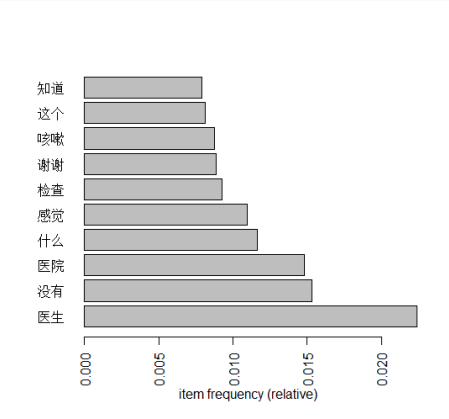
医学对话情感打分情况



医生 病人

共现词汇	共现次数
新增/确诊	782
新增/累计	660
病例/确诊	544
确诊/累计	540
病例/新增	484
确诊/肺炎	211
新增/肺炎	207
新冠/肺炎	199
新增/治愈	193
新增/出院	191
病例/累计	179
新型/肺炎	178
治愈/出院	167
病例/出院	157
病例/肺炎	155
治愈/确诊	145
确诊/出院	129
冠状病毒/新型	113
肺炎/累计	113





4 问题回答与总结

疫情舆情的重要时间结点

国内新闻舆情

平稳低增长时期
2019年12月31日至2020年1月24日

波动缓冲时期
1月25日至3月14日

焦点顶峰时期
3月15日至3月23日

冷却时期
3月22日起至4月23日期间，以4月9日为分界点

国外舆情

3月14日至3月23日期间，和疫情相关的新闻报道数量呈现波浪的形状，**3月17日至20日**，新闻报道产量达到波峰；在三月中旬和下旬两个时间结点，新闻报道数量均呈现较高的状态。**2020年3月11日**是新闻报道数量由低密度向高密度转变的时间结点；**3月12日至3月26日**期间，新闻报道数量呈现高频率的集中性特征，3月20日达到波峰。

01 全球新闻舆情

舆论在**1月29日、3月23日、4月6日**分别达到小高峰，并且在**3月23日**达到统计时间段内的最高值。

对应疫情爆发的生命周期，疫情期间的新闻舆情走向趋势和疫情爆发确诊人数走势呈现**显著正相关关系。**

从全球网络舆论数量来看，每日的全部推文数量（包含原创和转发）在**3月22日至4月6日**的这段时间内，总量一直在波动，于**3月31日**达到最低值，于**4月4日至4月5日**之间达到最高值。

02 全球网络舆情

4 问题回答与总结

疫情舆情演化的普遍规律

1. 疫情期间舆情演化趋势与疫情趋势走向呈现显著正相关关系。

2. 疫情舆情关注热点主题普遍分化在疫情认知防控，趋势走向，社会生活和政府动态等领域的一个或多个子领域中。

3. 舆情总体传递的情感极性是正向的。舆情误区和医学咨询行为的情感动机往往是负面的。

4. 政府措施和医学研究对舆论导向有较高的转化作用。