数据战“疫”，创新智“汇”

COVID-19 数据竞赛赛题一：

全球疫情舆情演化分析参赛报告

报告名称 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

参赛单位 南京大学

参赛队员 高雨婷 信息管理学院 2018 级

董沁怡 信息管理学院 2018 级

沈姝彤 信息管理学院 2018 级

王敬 信息管理学院 2018 级

吴迪 信息管理学院 2018 级

提交日期 2020.6.30

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

摘　要

COVID-19 疫情作为全球面临的严峻公共卫生安全事件，深刻地影响了世界社会的方方面面。为了从多个维度分析新冠疫情引发的社会各界对疫情认知和应对情绪的变化，掌握疫情与大众心理与信息理解的内在关联，我们利用全球新闻舆情、网络舆情和其他舆情等多种类型的舆情数据，以机器学习的原理和方法定位和分析疫情舆情产生原因、演化趋势、演化主题和影响因素，运用可视化和数学建模描述疫情舆情多个维度的信息表达，力图构建全球疫情舆情的生态框架。

研究采用的原始数据集为赛方提供的 COVID-19 疫情丁香园、CBC 新闻报道、相关推文、相关推文 ID、新闻语料库和医学对话数据集。本研究通过对全球疫情时间序列分布情况，新闻舆情、网络舆情和其他舆情进行综合文本挖掘和可视化统计学分析，定位和挖掘疫情舆情的重要时间节点，疫情爆发前期、中期、后期的舆情走向趋势，疫情中不同时期公众的关注热点，不同关注热点的受关注程度及原因，从而总结疫情期间的舆情演化的普遍规律，希望能够清晰构建全球疫情舆情的面貌，直观理性的对待疫情舆论心理和认知的趋势和走向。

关键词：新冠肺炎，舆情，全球，数据分析，自然语言处理，可视化，建模

目 录

1 绪 论

1.1 研究概述

COVID-19 疫情作为全球面临的严峻公共卫生安全事件，深刻地影响了世界社会的方方面面。为了从多个维度分析新冠疫情引发的社会各界对疫情认知和应对情绪的变化，掌握疫情与大众心理与信息理解的内在关联，我们利用全球新闻舆情、网络舆情和其他舆情等多种类型的舆情数据，以机器学习的原理和方法定位和分析疫情舆情产生原因、演化趋势、演化主题和影响因素，运用可视化和数学建模描述疫情舆情多个维度的信息表达，力图构建全球疫情舆情的生态框架。

1.2 数据来源和方法

研究采用的原始数据集为赛方提供的 COVID-19 疫情丁香园、CBC 新闻报道、相关推文、相关推文 ID、新闻语料库和医学对话数据集。我们主要使用 R 语言和 python，在对数据集进行批量获取和数据清洗后，运用机器学习的基本原理和内容展开数据分析，其中包括：探索性数据分析，时间序列，词频分布和地理可视化，信息计量学，文本词向量化，层次聚类，LDA 主题建模，词典型情感分析，关联分析，共现分析等。使用的数据可视化工具有：R，tableau，excel 等。

2 全球疫情舆情总体数据统计与情况分析

2.1 全球疫情实际状况

利用 DXYNews、DXYArea 和 DXYOverall 数据集统计疫情期间的新闻数量变化。数据集记录了从 2020 年 1 月 3 日到 2020 年 4 月 23 日的来自各大媒体的新闻数量和疫情数据。

2.1.1 全球疫情总体情况统计

全球确诊人数和死亡人数先下降后上升，在总体上波动增长，确诊人数和死亡人数都在 4 月 17 日达到最高值（分别为 27,705,523 和 10,298,990）；相对应的，治愈人数也在波动上升，同样在 4 月 17 日达到最大值 1,655,921 人。

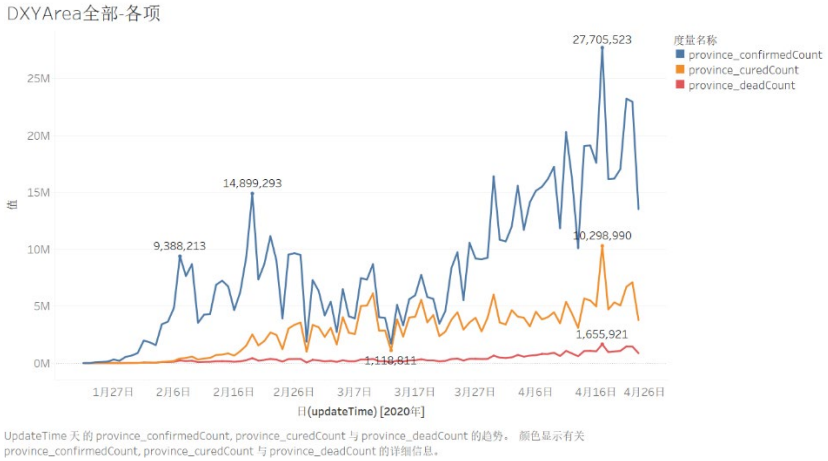


图 2.1.1 全球确诊、疑似、治愈和死亡人数统计

2.1.2 全球疫情地区分布

根据丁香园数据集，2020 年 3 月 14 日之前（含）几乎全部都是中国的确诊病例，在这些确诊病例中，湖北省的确诊数目占大部分。

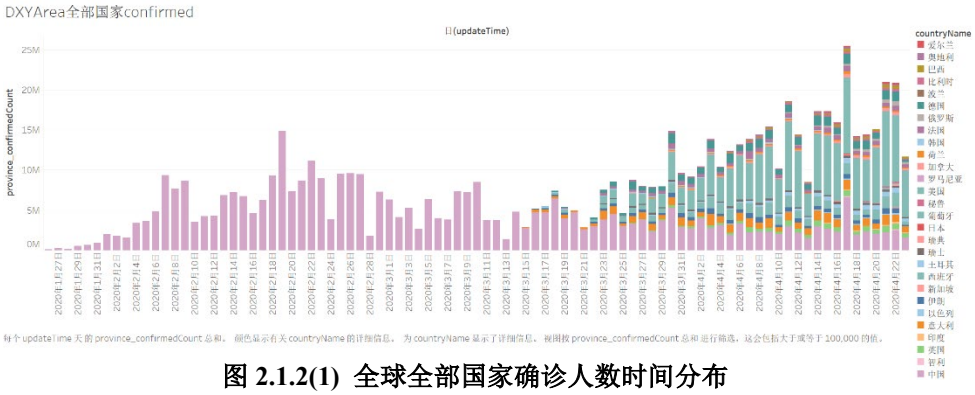


图 2.1.2(1) 全球全部国家确诊人数时间分布

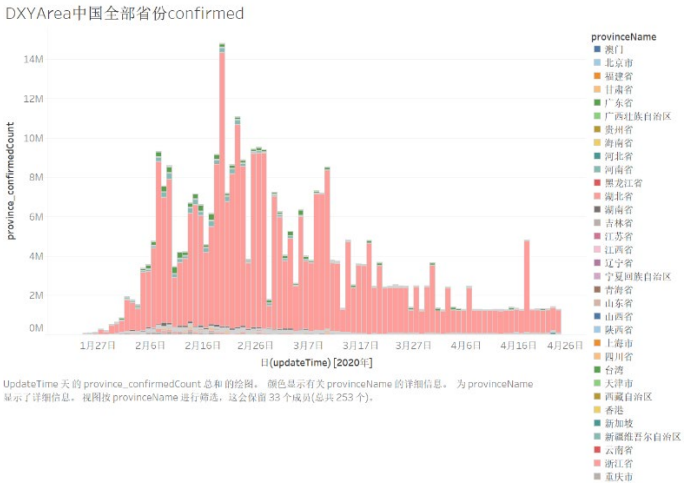


图 2.1.2(2) 中国全部省份确诊人数时间分布

而 3 月 18 日之后中国的新冠肺炎确诊人数波动下降，美国的新冠肺炎确诊人数波动上升，并迅速的超过了中国的确诊病例数，从图中可以看出，（在中国以外的国家当中）美国的确诊病例数占比高.

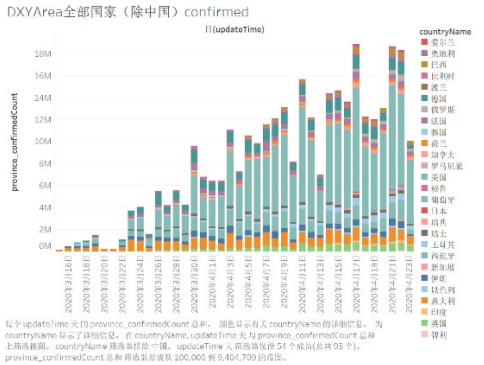


图 2.1.2(3) 海外国家确诊人数时间分布

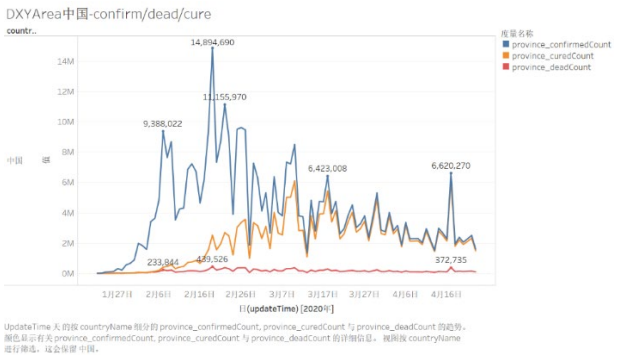


图 2.1.2(4) 中国确诊、死亡和治愈人数时间分布

根据时序图观察，前期中国确诊病例数较高，治愈率较低，2 月 17 日达到截至目前为止的最高值 14,894,690 例确诊，后期中国确诊病例数有所下降，治愈数逐渐上升 3 月 14 日之后治愈曲线几乎和确诊曲线重合；死亡曲线也较低，疫情控制状况较好。

以洲为分析单位，利用 DXYArea 数据集观察疫情数据在时间上的动态分布，并统计出疫情在各洲的蔓延情况。在 1 月 22 日至 4 月 23 日期间，北美洲和欧洲的疫情形势严峻，确诊病例最多，而北美洲的疫情形势首当其冲。亚洲和南美洲处于第三和第四。

continentName	**确诊人数**
北美洲	842629

欧洲	208389
亚洲	98674
南美洲	45757

大洋洲	6647
非洲	3659
其他	712

为了呈现海外疫情数据的地域化增长趋势，绘制北美洲和欧洲的疫情趋势曲线。综合分析可知，欧洲和北美洲疫情确诊病例开始增加的时间节点都在 2 月 1 日左右，欧洲 3 月初（约一个月）进入疫情爆发期，北美洲在 3 月上中旬进入疫情爆发期。截至 4 月中下旬，北美洲的疫情形势已经十分严峻，疫情人数也基数庞大。

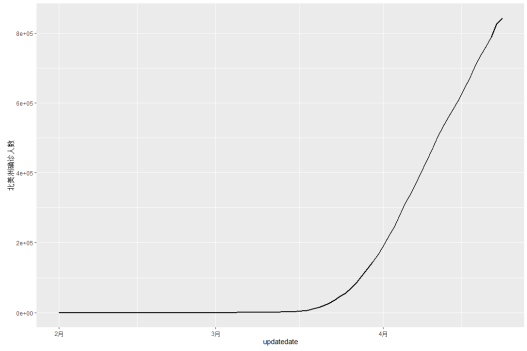


图 2.1.2(5) 北美洲确诊人数时间分布

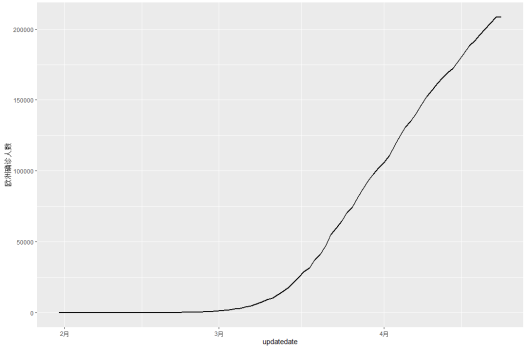


图 2.1.2(6) 欧洲确诊人数时间分布

基于 CBC 新闻报道数据集和丁香园数据集中来源于美国的多家媒体的语料挖掘目的，我们进一步调查了北美洲的美国和加拿大的疫情形势。

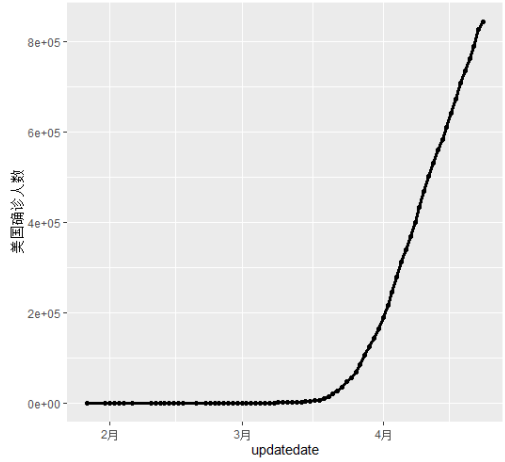


图 2.1.2(7) 美国确诊人数时间分布

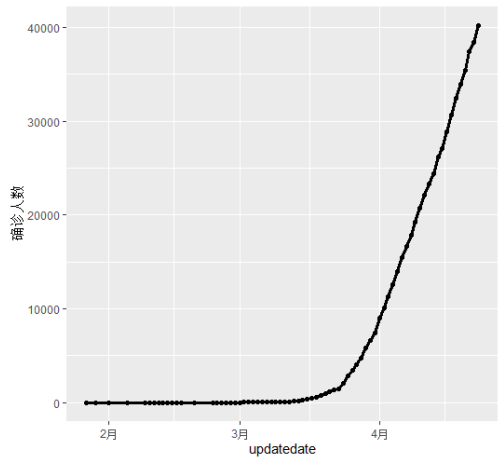
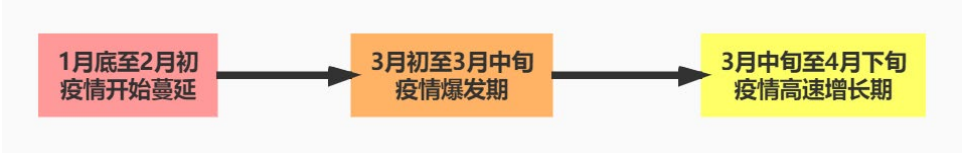


图 2.1.2(8) 加拿大确诊人数时间分布

由此，我们总结疫情从 1 月 22 日至 4 月 23 日期间在海外的疫情发展趋势为：



2.1.3 对疫情认知情况

（1）病毒

新型冠状肺炎的致病病毒，起初被命名为新型冠状病毒 2019-nCoV，最终被确定为 SARS-Cov-2，其导致疾病命名为 COVID-19。

图 2.1.3(1) 病毒认知标签时间分布图

（2）传播途径

最初仅认为存在人传人、医务人员感染、一定范围社区传播的途径；之后提出过存在粪口传播、气溶胶传播，消化道传播的可能性。最终研究认为：经呼吸道飞沫、接触传播是主要的传播途径，气溶胶传播和消化道等传播途径尚待明确。

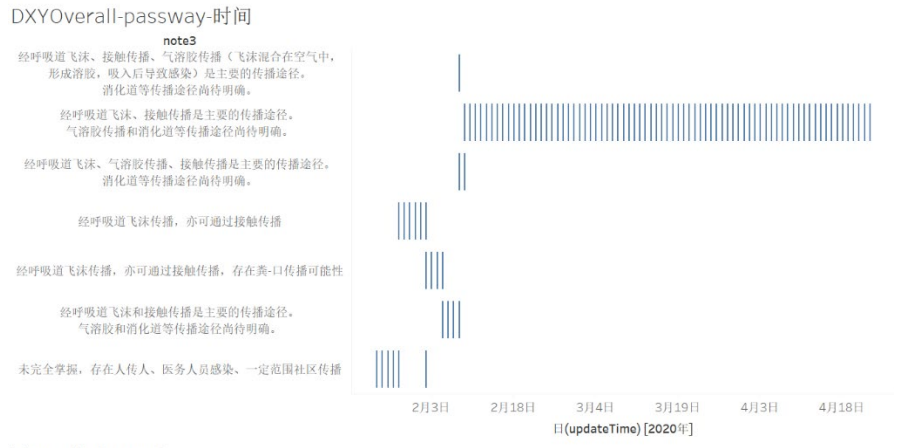


图 2.1.3(2) 传播途径标签时间分布图

（3）传染源

最初研究认为由野生动物作为传染源，此后猜测认为是中华菊头蝠或中国马蹄蝠。最终确认为：新冠肺炎的患者，无症状感染者也可能成为传染源。

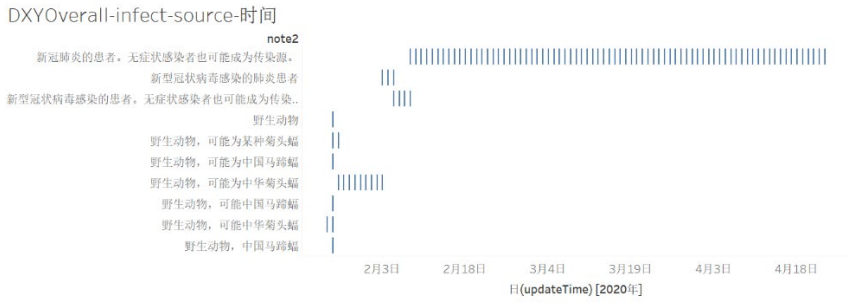


图 2.1.3(3) 传染源标签时间分布图

2.2 全球疫情舆情总体走向趋势：基于丁香园数据集

从全球的报道图来看，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。

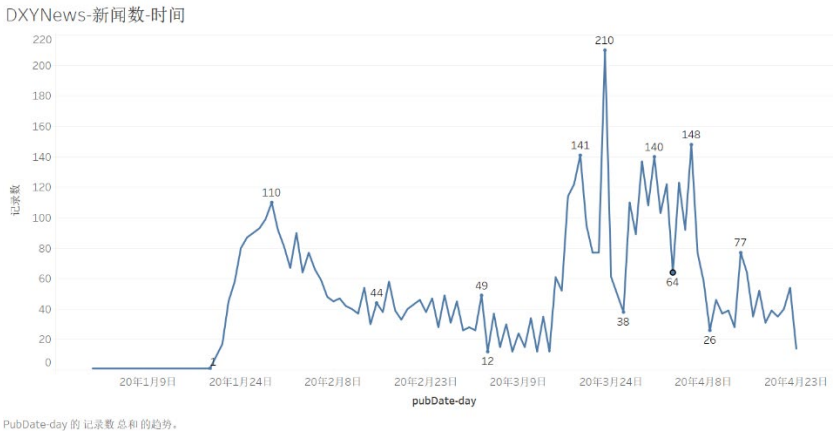


图 2.2(1) 通用新闻数量时间分布

1 月 29 日，中国还是应对新冠肺炎的重要国家，以中文报道居多新闻报道数量达到了短时间内的最高值，拉萨市卫生健康委员会发布拉萨首例新型冠状病毒感染的肺炎疑似病例的相关情况，并且西藏启动重大突发公共卫生事件一级响应；截至 1 月 29 日为止中国内地 31 省份全部启动突发公共卫生事件一级响应。

图 2.2(2) 1 月 29 日起词频分布可视化

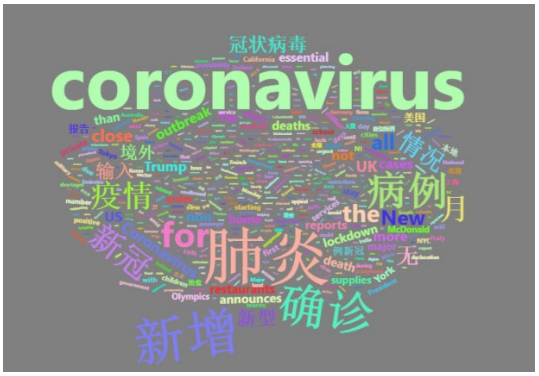


图 2.2(3) 3 月 23 日起词频分布可视化

3 月 23 日，特朗普宣布华盛顿州为新馆肺炎疫情的“重大灾区”，美国多个州关闭参观、酒吧等公共餐饮场所。并且自 3 月 23 日起，美国肺炎确诊病例数迅速上升。而中国新冠肺炎疫情控制状况良好

4 月 6 日左右，我国 31 省份新增新冠肺炎境外输入确诊病例，境内确诊病例均以输入兵力为主；西班牙新冠肺炎确诊超过 12.4 万，政府延长国家紧急状态；英国新冠肺炎死亡率又迅速超过西班牙；意大利确诊病例近 12 万。

图 2.2(4)4 月 6 日起词频分布可视化

3　全球新闻舆情：国内外新闻报道数量趋势分析与主题演化

为了准确把握新闻报道内容的准确语义信息，避免偏差和重复，对新闻报道的文本挖掘将从标题和新闻摘要两个维度展开。

3.1 国内疫情新闻舆情

3.1.1　数据可视化与时间节点划分

下图展示了全球通用新闻报道数量从 2019 年 12 月 31 日至 2020 年 4 月 23 日的数量增长趋势。从 2019 年 12 月 31 日至 2020 年 1 月 24 日，有关武汉疫情的新闻数量呈现增长趋势，并在 1 月 24 日到达了一个小高峰。从 1 月 25 日起，新闻数量增长出现波动下降趋势，然而从 3 月 15 日起，新闻数量又呈现持续激增趋势，并在 3 月 23 日左右达到顶峰，3 月 24 日起，新闻数量增长又呈现明显下降的趋势。根据时间分布曲线的增长特征，可以将疫情新闻大致分为 4 个时期：2019 年 12 月 31 日至 2020 年 1 月 24 日为平稳低增长时期，疫情新闻处于较低投放、低关注的形势；2020 年 1 月 25 日至 3 月 14 日为波动缓冲时期；3 月 15 日至 3 月 23 日为焦点顶峰时期，疫情新闻事件呈现高度饱和的状态，疫情受到全球广泛关注，有关疫情的相关报道被广泛撰写和投放，疫情成为密集型新闻热点话题；3 月 24 日起为冷却时期，武汉疫情事件的话题度降低，信息饱和并降温。

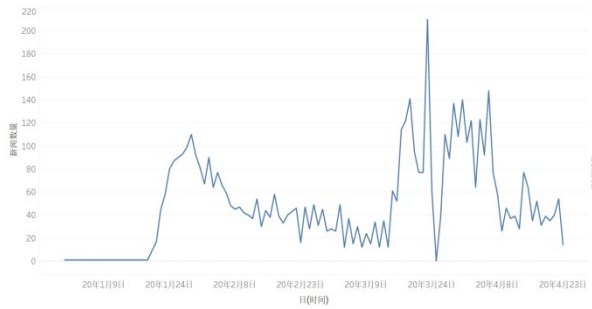


图 3.1.1(1) 通用新闻数量时间分布

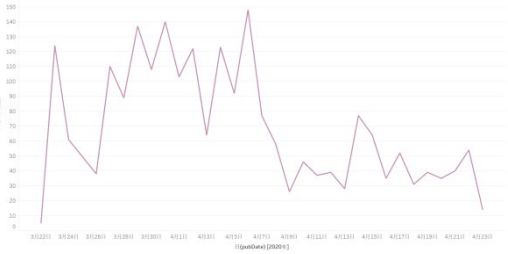


图 3.1.1(2) 新闻语料数量时间分布

针对 DXYNews 数据集提供的新闻语料信息，可以看到如上图所示的新闻数量随时间的分布情况，从 3 月 22 日起至 4 月 23 日期间新闻数量是明显呈大幅度波浪起伏分布的。可以看出以 4 月 9 日为分界点，4 月 9 日之前新闻数量随时间虽然也是呈大幅度波浪分布，但总体的新闻数量高于 4 月 9 日之后的新闻数量。

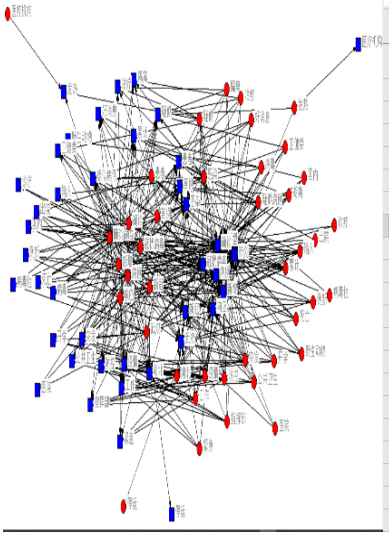
2019 年 12 月 31 日至 2020 年 1 月 24 日

词汇	词频	频率(%)
冠状病毒	43	2.9
感染	37	2.5
武汉	36	2.5
首例	24	1.6
新冠	24	1.6
发现	17	1.2
启动	17	1.2
响应	17	1.2
一级	15	1
累计	14	1
湖北	14	1

图 3.1.1(3) 2019.12.31-2020.1.24 新闻标题词频图和词云

由词云和词频统计（频率≥1）可知，2019 年 12 月 31 日至 2020 年 1 月 24 日这段时间，新闻标题的主旋律是疫情确诊人数累计和新型冠状肺炎病毒，相关关键词占据了新闻文本的前 7 位。其他的重要主题为：新型肺炎的感染和紧急措施的采取；“一级”，“感染”等关键词

——疫情的预防，诊断，治疗；从总体词语分布上来看，词汇分布在社会民生，疾病医疗、公共安全，经济发展和政治军事这多个方面。防疫工作的开展、病毒的传播情况和疫情对社会的影响在新闻中占据重要地位。



共现词汇	共现次数
新型/肺炎	97
病例/肺炎	94
确诊/肺炎	72
新型/病例	71
病例/确诊	64
新型/确诊	54
新增/肺炎	51
病例/新增	44
新型/新增	41
冠状病毒/新型	37
确诊/新增	34
冠状病毒/感染	30
新型/感染	28
冠状病毒/肺炎	26
肺炎/感染	26
新冠/肺炎	20
冠状病毒/病例	19
首例/肺炎	19
新型/首例	18
武汉/肺炎	17
病例/感染	16

图 3.1.1(4) 2019.12.31-2020.1.24 新闻标题语义网络图和词汇共现表

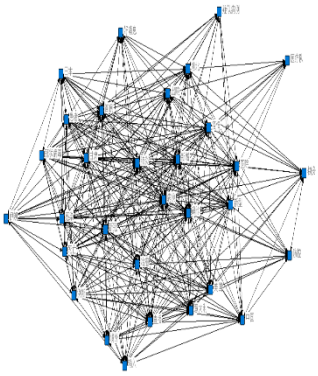
根据语义网络图和词汇共现表，词汇多与肺炎、感染等词语搭配出现，表明这个时期属于疫情初期，病毒正在全国蔓延，但并未引起社会高度重视。新闻报道标题以发现肺炎感染者，以达到警示目的。语义网络较为简单，以肺炎、感染、新型等词为中心，停运、医疗机构为边缘词语构成。

2020 年 1 月 25 日至 2020 年 3 月 14 日

词汇	词频	频率(%)
新增	1601	11
确诊	1371	9.4
累计	990	6.8
病例	856	5.9
肺炎	560	3.8
出院	345	2.4
治愈	302	2.1
新型	235	1.6
新冠	235	1.6
疫情	197	1.4

图 3.1.1(5) 2020.1.25-2020.3.14 新闻标题词频图和词云

2020 年 1 月 25 日至 2020 年 3 月 14 日，从高频词语内容可以看出处于病毒发现、感染与治疗治愈的初期阶段，主要内容是针对新型肺炎病毒确诊与治愈进行报道；这些新闻报道说明了国家和政府关注本次疫情，并采取相关措施，目的在于防控疫情，救治病患和阻击病毒。态度上是积极乐观，充满信心的。



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

图 3.1.1(6) 2020.1.25-2020.3.14 语义网络图和词汇共现表

根据语义网络图和词汇共现表，词汇多与新增、确诊等词语搭配出现，表明这个时期属于疫情中期，病毒已经在全国肆虐，但引起社会恐慌。新闻报道标题以确诊和治愈肺炎感染者，以达到安抚目的，避免民众的过度恐慌造成社会损失。语义网络较为复杂，以新增、确诊、新型等词为中心，新闻报道标题中的每个词语都息息相关。

2020 年 3 月 15 日到 2020 年 3 月 23 日

词汇	词频	频率(%)
肺炎	130	0.8
确诊	120	0.7
病例	103	0.6
新增	86	0.5
新冠	85	0.5
境外	52	0.3
输入	52	0.3
累计	50	0.3
疫情	44	0.3
情况	40	0.2
例新	29	0.2

图 3.1.1(7) 2020.3.15-2020.3.23 新闻标题词频图和词云

3 月 15 日到 3 月 23 日，新闻关键词的种类和数量都明显增多。从空间来看，包含的地点关键词种类激增，说明疫情已经扩散向全球，并从境外不断向国内输入肺炎感染者，全球肺炎蔓延趋势令人担忧；在疫情防控方面出现了更多强制性、具体化的措施，着重强调了一线封城、公安等部门的部署、严格检疫，监管等工作；全国各地的社区卫生安全防护工作被深度重视；全面全民抗击肺炎、宣传肺炎知识、驰援重度感染区被认为是刻不容缓的国家大事。在这一时期，肺炎事件的重要性和紧急性达到顶点。和肺炎疫情相关事件的主题次序为：肺炎确诊、病例新增和境外输入。新闻态度上，从上到下，从时间到空间上都展现了对疫情形势的高度关注和全民全面防控的迫切和紧急心。

共现词汇	共现次数
新冠/肺炎	71
病例/确诊	63
确诊/肺炎	55
境外/输入	45
病例/新增	40
新增/确诊	40
新冠/确诊	39
确诊/累计	38
病例/肺炎	36
情况/疫情	36
情况/肺炎	36
疫情/肺炎	36
病例/输入	33
新增/肺炎	33
病例/境外	32
新增/境外	30
新增/输入	26
肺炎/累计	25
例新/肺炎	24

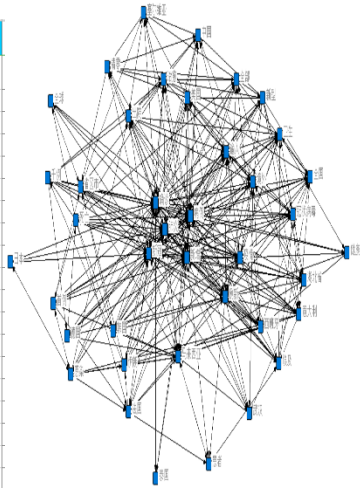


图 3.1.1(8) 2020.3.15-2020.3.23 语义网络图和词汇共现表

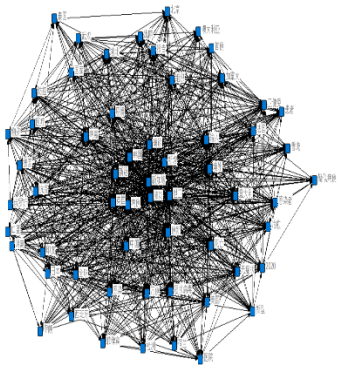
根据语义网络图和词汇共现表，词汇多与境外、输入等词语搭配出现，表明这个时期属于疫情中后期，病毒已经在向全球扩散，情况严重。新闻报道标题以境外输入肺炎感染者，表明抗击疫情已经变成了全球的责任，以达到呼吁共同合作抗疫的目的。语义网络较为简单，以境外、输入、新增等词为中心，新闻报道标题中出现了法国、泰国等国家，与肺炎、新增等词联系。

2020 年 3 月 23 日后

词汇	词频	频率(%)
确诊	1706	8.9
肺炎	1407	7.4
病例	1386	7.2
新增	1110	5.8
新冠	937	4.9
累计	816	4.3
例新	427	2.2
输入	269	1.4
境外	256	1.3
疫情	218	1.1

图 3.1.1(9) 2020.3.23-至今 新闻标题词频图和词云

从 3 月 24 日的统计关键词中，变化情况如下：“新增”与“确诊”说明新型冠状病毒肺炎的确诊病例和新增情况变为新闻关注的热点；世卫和医疗队在关键词中的高频出现说明有关肺炎的官方调查情况和防治相关研究受到广泛关注；患者的症状、治疗和疫情隔离情况，政府的病情通报和卫生安全工作受到密切追踪。“复工”说明疫情得到了极大的控制。对比前期，这段时间关于病毒的情况已经得到更为完善的确认，新闻的主题从认知病毒感染迁移到防控工作中。这段时期新闻的要点密切围绕政府和权威机构发布的动态。



共现词汇	共现次数
确诊/病例	1111
确诊/肺炎	1001
病例/肺炎	862
新冠/肺炎	780
确诊/累计	644
确诊/新增	623
确诊/新冠	585
累计/病例	552
肺炎/新增	551
病例/新增	534
新冠/病例	486
累计/肺炎	465
新冠/新增	359
例新/肺炎	321
例新/新增	306
例新/确诊	281
新冠/累计	239
例新/病例	230
境外/输入	221

图 3.1.1(10) 2020.3.23-至今 语义网络图和词汇共现表

根据语义网络图和词汇共现表，词汇多与确诊、境外、输入等词语搭配出现，表明这个时期属于疫情后期，病毒已经在全球肆虐，确诊人数不断增多，疫情严重性达到了新的高度。新闻报道标题以境外输入肺炎感染者和确诊肺炎感染，表明国内疫情已得到较好控制，但国外疫情不容乐观。语义网络非常复杂，每个新闻标题词语之间都非常多的联系。

对于 3 月 23 日以后的新闻报道，可以以 4 月 9 日为节点进行更精细的划分。

图 3.1.1(11) 2020.3.23-2020.4.9 新闻标题词云图

图 3.1.1(11) 2020.4.9-至今 新闻标题词云图

上面的图分别是整体新闻报道摘要中文 4 月 9 日之前和之后的词云图。4 月 9 日之前的高频词与整体的高频词是大体是一致的，主要都是围绕“病例”“确诊”“新增”“肺炎”“冠状病毒”等，即新闻的主要维度是新冠病毒感染确诊情况、治愈出院情况等等。在 4 月 9 日之后的高频词中可以看到出现了“当地”“输入”“境外”等词汇，结合国外疫情的发展趋势，可以知道 4 月 9 日左右，国外新冠确诊人数大幅度增长，因此与国外实时疫情相关的新闻开始出现与增多。

3.1.2 文本向量化与主题聚类

对文本分词后构建文档矩阵，筛选词频过低的词汇。使用层次聚类中的离差平方和法实现新闻标题的聚类。如图所示，模型被聚类成为二叉树模型。各级的广度分别为 40，80，120。查看分类器分类所得特征文档，这些文档反映了国内疫情严重，国家各方面都受到严重打击。政府采取一级响应措施，研制肺炎疫苗，医疗物资供给，并且全社会抗疫并采取强制性措施。而随着病毒的高传染性，新冠肺炎快速在国外蔓延开来，并不断向中国境内输入感染病例，对国家的抗疫情行动构成了极大的威胁和挑战。

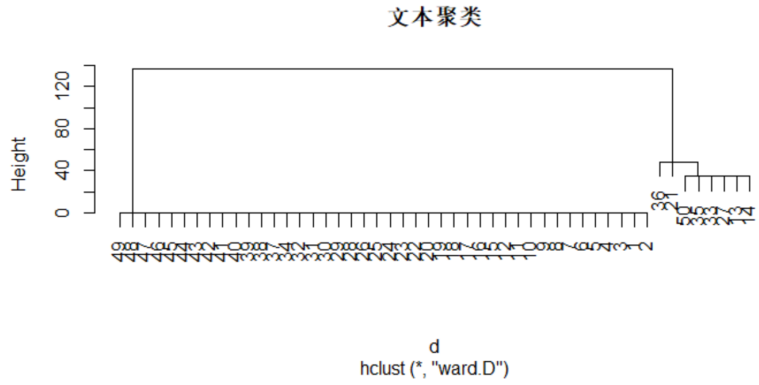


图 3.1.2(1) 新闻文本层次聚类

基于 LDA 主题建模对新闻文本进行聚类分析 k=9 为起点，对文本进行 LDA 主题建模分析（alpha = 0.10，eta =0.02），依次分析 k=5，4，3 时依据文本主题分类的效果。

K=3 时，第一类文本聚焦于疫情下的民生，如社区，职业，和公共预防措施等。第二类文本集中表现了对企业和经济事件的关注。第三类文本聚焦于疫情传播和新冠病毒研究进展。

K=5 时，分类主题分化的更加明显。第一类大致为教育和就业在疫情中受到的影响；第二类为政府决策和防治措施；第三类为各地区的社会民生；第四类为经济状况和资本变动；第五类为疫情医疗和社会联动情况。

根据分类结果进行比较，则 k=4 时分类效果最好，从而将主题分为四类。并将分类情况进行可视化（挑选出每类出现频次最高的词语）：

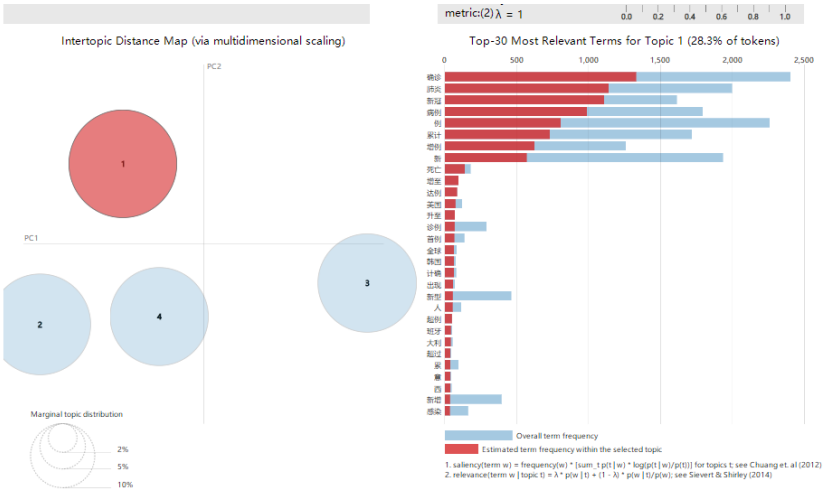


图 3.1.2(2) 新闻文本 LDA 主题建模分析

TOPIC1：肺炎患者发现和确诊统计。该类别多次出现“确诊”，“新增”，涉及的高频词大部分针对疫情中的感染者的基本统计，以便为抗疫工作更好的开展。
TOPIC2：全国疫情发展和抗疫行动开展。该类别以武汉、广东等疫情重灾区和相关事件为主要关注点，层层推进，关注全国各个疫情灾区的工作部署、疫情发展。
TOPIC3：疾控政策与政府防疫动向。该类别的行为主体是党员及医疗集体，涉及的词汇包含了政府的官方措辞和权威发布，语句较为严谨，并提及了多种会议和公共的卫生措施。
TOPIC4：国内抗疫成效和境外输入。该类别涉及国内和国外的疫情发展，在国内疫情

取得突破性进展的同时，国外疫情形势严峻，境外输入成为国内肺炎病毒不断传播的重要原因。

3.2 国外疫情新闻舆情

3.2.1 数据可视化与节点划分

针对 DXYNews 和 CBC 新闻报道数据集提供的新闻语料信息，我们得到以下的统计结果。由于丁香园数据集中仅包含了 3 月 14 日至 3 月 23 日期间来自欧洲和美国的少量主流媒体的部分新闻报道数据，故仅统计这一段时间内的报道增长对比情况。同时，分析 CBC 新闻报道数据集新闻报道的日增长情况。

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。CBC 数据集记录了 CBC 新闻报道数量在 2020 年 1 月 8 日至 3 月 27 日期间的时间分布。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，3 月 27 日得到暂时下降（有可能由于数据集统计不完全导致）。两份新闻语料数据充分反映了三月中旬和中下旬是疫情相关报道波动的主要时间结点。

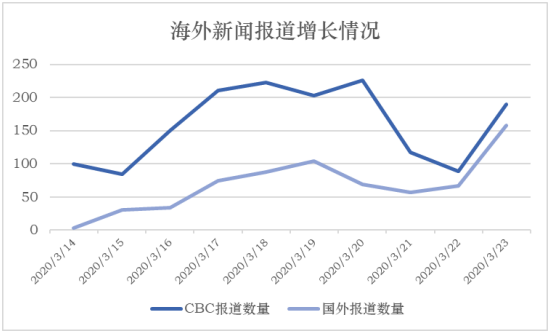


图 3.2.1(1) 海外新闻报道增长情况

图 3.2.1(2) CBC 报道数量增长情况

我们猜想新闻报道密集性特征可能是由于疫情形势导致的。进一步分析加拿大和欧美地区疫情数据和新闻报道数量数据的相关性，并检验统计显著性。加拿大疫情和海外新闻报道数量的相关系数为 0.5298489，检验 p 值为 0.0001281，通过显著性检验。这证明新闻报道与疫情有着较强的相关性。

基于 CBC 新闻数据集对加拿大新闻报道文本进行内容分析。标题词汇词云图中出现频率较高的是 covid、coronavirus、cases、pandemic、health、new、amid、outbreak、home、alberta、Canada、China 等单词，说明加拿大的居民在关注加拿大本地的新冠疫情发展态势的同时，同样也很关注中国的疫情发展状况，以及如何应对疫情、对健康的影响以及医学方面对疫情是否有突破等方面。新闻摘要词云图中，关于内容，主要以新冠病毒及症状，地区确诊情况，政府公共卫生安全措施，社会民生和医疗健康等几个维度展开；有关病毒和疫情的来源认知为来源于中国；CBC 将疫情的定性为危急，目前的疫情形势明确为爆发。诸多关键词体现了 CBC 对加拿大的疫情具有高关注度。

图 3.2.1(4) CBC 新闻摘要词云图中

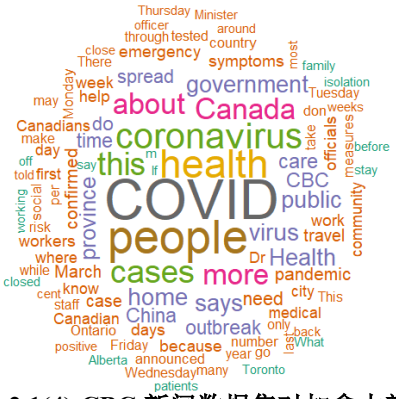


图 3.2.1(4) CBC 新闻数据集对加拿大新闻报道文本进行内容

以 3 月 11 日为分界点，对比疫情发作期和爆发期期间的 CBC 新闻报道内容，前者的新闻主体内容在于强调疫情病毒的风险，症状，来源，并探寻疫情的防控和传播来源，从具体上来说，与中国武汉相关和携带病毒的乘客相关的关键词词频较高，将此次病毒和 SARS 相对比，以疫情传播和数据的报道较多；3 月 11 日以后，与“中国”“武汉”“确诊”相关的关键词词频明显降低，新闻报道的主要内容在于北美各地区的疫情防控和疫情期间的社会运转与资源分配等相关问题。新闻开始以经济，政治，教育就业等一系列维度为角度剖析和报道政府和各地对应疫情做出的重大决策和调整，这恰说明了疫情形势严峻和人们对疫情的研究进展推进，导致国家和政府面临的问题更加焦灼和复杂。

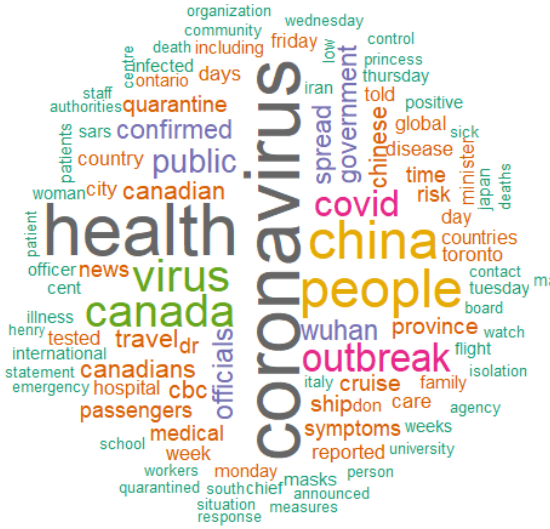


图 3.2.1(5) 2020.3.11 前 CBC 报道内容词云图

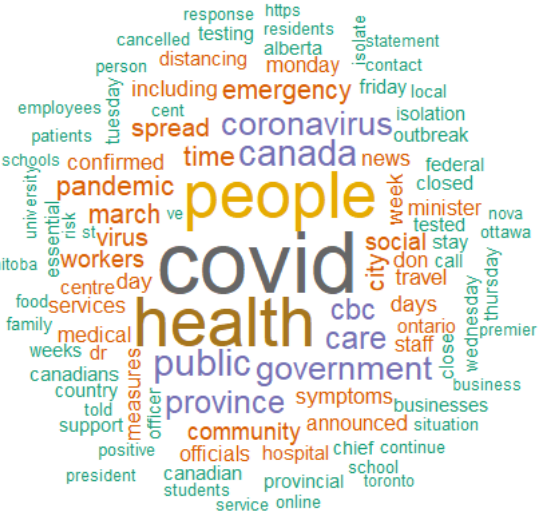


图 3.2.1(6) 2020.3.11 后 CBC 报道内容词云图

在 3 月 14 日至 23 日期间，丁香园数据集中国外主流媒体的新闻报道的主要内容以疫情聚焦与防控，政府的主要应对措施和国家生命健康财产危机为主，表现了对政府政治行为的高度关注。

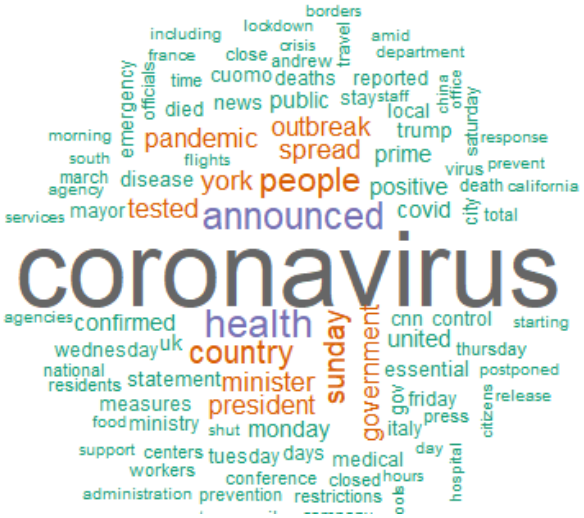


图 3.2.1(7) 2020.3.14-2020.3.23 丁香园数据集国外主流媒体新闻报道

3.2.2 文本向量化与主题聚类

对 CBC 新闻标题层次聚类。由文本的层次聚类结果可知，各级广度为 2、3、4、5、6、29，查看分类器分类所得特征文档，这些文档主要是加拿大政府对疫情采取的各种措施、在医疗上的一些突破、加拿大公众对疫情的关注程度、社会各阶层以及各领域受到的影响。考虑 k 值为 2、3、4、5、6 的主题建模。

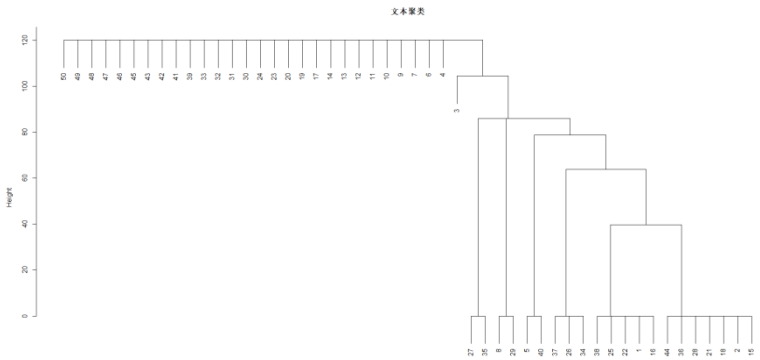


图 3.2.2(1) CBC 新闻标题层次聚类

对文本分词后构建文档矩阵，筛选词频过低的词汇，获得 624 条有效词汇。使用层次聚类中的 Complete 相似分析法实现新闻的聚类。如图所示，模型被聚类成为具有 10 个叶子结点的二叉树模型。各级的广度分别为 2，3，5，7，10.查看分类器分类所得特征文档，这些文档反映了加拿大疫情形势严峻，政治和经济局势受到冲击。政府呼吁社会规避旅行，大规模消毒和全社会抗疫并采取强制性措施，众多活动被取消，疫情控制阻碍重重，社会医疗不堪重负。北美政府正企图限制出入境，增加资金救助，并刺激产量的供应。

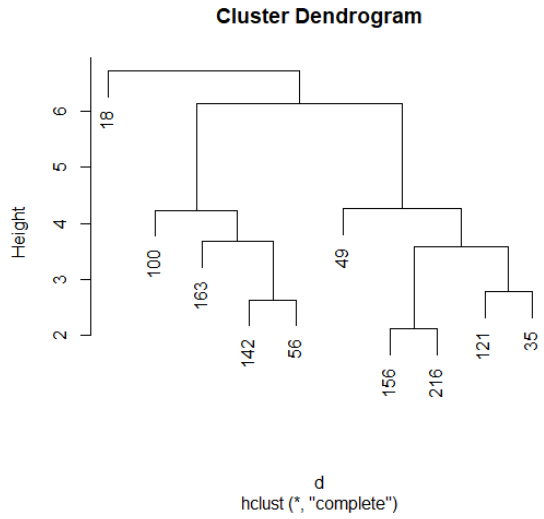


图 3.2.2(2) CBC 新闻文本层次聚类

（1）新闻标题文本聚类

对新闻标题文本聚类后的新闻数据进行基于 LDA 模型的主题建模，k 值分别为 2、3、4、5，得到以下结果：

当 k=5 时，文本被划分为五类。第一类文本是聚焦于疫情中的紧急状况出现如何应对以及疫情的传播情况；第二类文本的重心是医疗方面对疫情的一些进度和是否有突破，公众对此的看法等；第三类文本是疫情对加拿大的公众出行的影响，主要集中于飞行方面的外出等活动；第四类文本是以疫情的特点以及社区等地方的医疗设施是否完备、医生是否持续对当地的疫情进行处理等；第五类文本着重点放在了经济方面，涉及到地区的候选人、学校的资助等方面，都是牵扯到该地区经济发展的因素。

当 k=4 时，文本被划分为四类。第一类文本主要聚焦于中国的疫情情况，中国的医院的数量变化、病例增长情况等，可以看出，加拿大的新闻也对中国的疫情状况有明显地关注；第二类文本以疫情中的紧急状况出现如何应对以及疫情的传播为中心；第三类文本题主要是底层工作者在疫情中受到的影响，在底层工作者中出现的一些病例、对底层工作者的一些补助等；第四类文本的焦点是医学上对于疫情是否有突破，在疫情上的一些进展等。

当 k=3 时，文本被划分为三类。第一类文本主要聚焦于加拿大当地以疫情发展状况以及对社会各阶层的影响等；第二类文本的重点是加拿大对疫情采取的一些措施，包括医疗关怀、资金补助等；第三类文本主要是以医学上对疫情是否有突破以及相关进展、医生的一些回应等。

（2）新闻内容文本聚类

K=2 时，文本被分为两类，分析两类模型的词汇组成。第一类文本大多着眼于疫情传播形势的播报，以及在全球和全国各地区造成的公共安全事件情况；第二类文本聚焦于社会在疫情局势下受到的影响，如医疗，政府，社会媒体，公共计划方案和各职业的生存情况等。从总体上，CBC 新闻播报以新闻事件主体和社会生活两个维度展开。

K=3 时，第一类文本聚焦于疫情下的民生，如社区，职业，和公共预防措施等。第二类文本集中表现了对企业和经济事件的关注。第三类文本聚焦于疫情传播和新冠病毒研究进展。

K=5 时，分类主题分化的更加明显。第一类大致为教育和就业在疫情中受到的影响；第二类为政府决策和防治措施；第三类为各地区的社会民生；第四类为经济状况和资本变

动；第五类为疫情医疗和社会联动情况。

K=7 或 10，分类主题的界限变得模糊，根据情况进行剪枝，发现 k=5 时的分类主题结果最清晰。

综上所述，可以大致将加拿大疫情相关的新闻报道主题进行以下的分类：

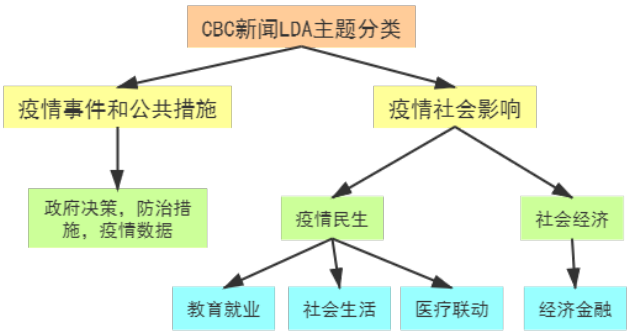


图 3.2.2(3) 加拿大疫情相关新闻报道主题分类

对丁香园数据集中提供的其他海外主流媒体相关新闻报道进行层次聚类。层次聚类所得的特征文档表明疫情形势严峻，美国的纽约，华盛顿和加利福利亚等地区疫情严重，不容乐观；政府要求学校和企业员工居家隔离；疫情传染的进一步特征并不明确。利用 LDA 主题模型进行预测，k=3 时，新闻报道可以划分为政府政治动向，疫情防控，社会民生（以各社会群体疫情隔离和医疗为主）。

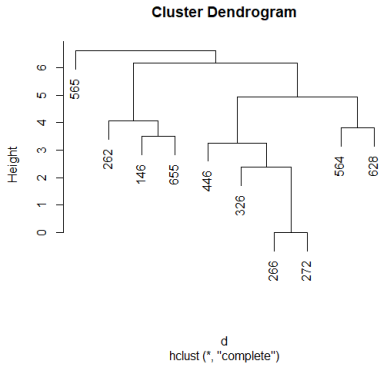
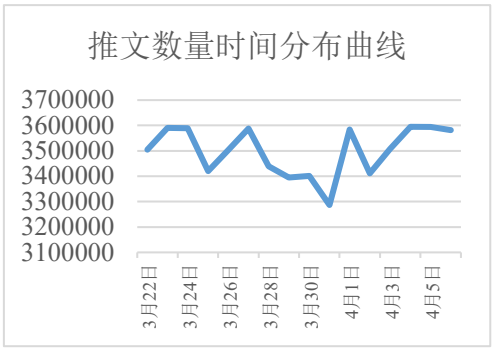


图 3.2.2(4) 其他海外主流媒体相关新闻报道层次聚类图

从总体上，海外疫情从 2 月下旬至 3 月初开始蔓延，3 月中旬达到迅速爆发期，直到 4 月中下旬，疫情仍未得到有效控制。在新闻报道主体上，政府和政治成为最为核心的决策群体，政府决策的效用在疫情期间得到迅速凸显。政府的执政水平也与疫情防控水平息息相关。在新闻报道中传达出的防控疫情的信息是社会资源的保障至关重要，维护和巩固供应链困难重重。而疫情控制是政治经济得以正常运行的必要条件。

4　全球网络舆情：社交媒体舆论随时间的传播演化规律

4.1 推特推文信息的转发行为频率考察



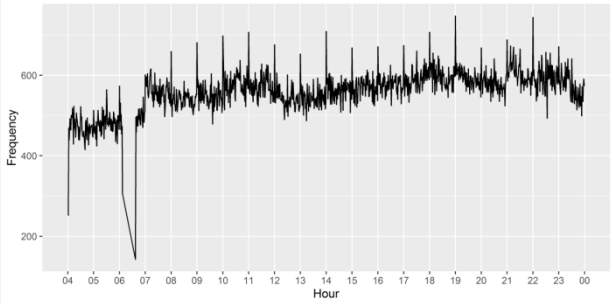


图 4.1(1)每日推文总量时间分布曲线　　**图 4.1(2) 单日原创推文发布数量时间分布曲线**

从总体上来看，每日的全部推文数量（包含原创和转发）随时间的变化如上图所示。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。

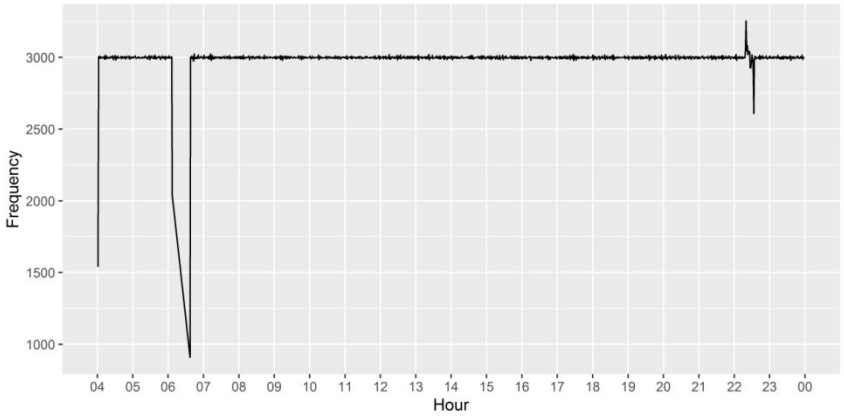


图 4.1(3) 单日全部推文发布数量时间分布曲线

每一天的发推数量变化趋势（包括原创和转发）都基本一致。以 3 月 22 日的数据为例，（不含转发量的）原创推特总数在 4:00 和 6:30 左右会有谷值；（含转发量的）发推总量在凌晨 4:00 和早晨 6:30 左右会迎来一天中的低谷时间段，而在晚 22:00 左右会有一个发推的小高峰。

3 月 22 日，原创 twitter 数量大概在 500-600 之间；3.23 之后 twitter 数量上升到 600-700 之间。包含转发量的 twitter 量，可以体现活跃情况，总体上所有时间段的数量大都维持在 3000 条左右，在每日凌晨 4:00 左右和 6：30 左右会达到每一天的两个低谷值。

3 月 22 日，twitter 上的讨论量不高，讨论重心在意大利和中国（Italy、China）的流行病疫情（pandemic）上。3 月 23 日起，讨论量有所上升，在突然爆发的疫情危机中（crisis、outbreak）导致的公共场所和机构的关闭（lockdown）也引起了大量关注；美国总统特朗普（DonaldTrump、Trump、president）和民主党（Democrats）在本次疫情中的行动也受到了大量的讨论。

3 月 24 日至 3 月 25 日，疫情引起进一步的讨论（pandemic、corona），而在此基础上，查尔斯王子（Charles、Prince）确诊新冠肺炎也引起了一定的讨论。

3 月 26 日至 3 月 28 日，在此前的讨论基础上，居家隔离、隔离期（quarantine、stayhome、stayhomeandstaysafe）、受疫情影响而关闭公共场合（coronalockdown）等概念也受到了关注。

3 月 29 日，特朗普政府因应对疫情不力广受质疑，有声音认为美未能在全球抗疫中发挥领导力（华盛顿邮报），并且美接受了中国送来的物资，这一点也引起了大量讨论。

3 月 30 日至 4 月 4 日，阿根廷、印度尼西亚等国家相继进入严备的疫情境界状态。3 月 30 日，阿根廷新增新冠肺炎确诊病例 146 例，累积 966 例，阿根廷政府（gobierno）正式宣布延长“全民隔离”。3 月 31 日，印度尼西亚总统宣布国家进入卫生紧急状态，禁止外国人入境，印尼工作人员在王宫建筑群、伊斯兰教堂等多个场合喷洒消毒剂（penyemprotan）的照片也引起了讨论。

4 月 5 日，美国总统特朗普与印度总理纳伦德拉·莫迪（Narendra·Modi）进行会面，讨论了新冠肺炎疫情和供应链问题，强调了希望摸底放行美国订购的羟氯喹（hydroxychloroquine，预期成为治疗新冠病毒的一种成功药物）。与此同时，全球新冠肺炎累计确诊超过 120 万例，西班牙成为欧洲疫情最严重的国家（pandimia）。

4 月 6 日，英国首相鲍里斯·约翰逊（Johnson）的严重病情引起了广泛的讨论，而英国首相府称约翰逊未诊断出肺炎。

4.2 推特特定用户发文时间分布规律挖掘

一月份 ID 数目：

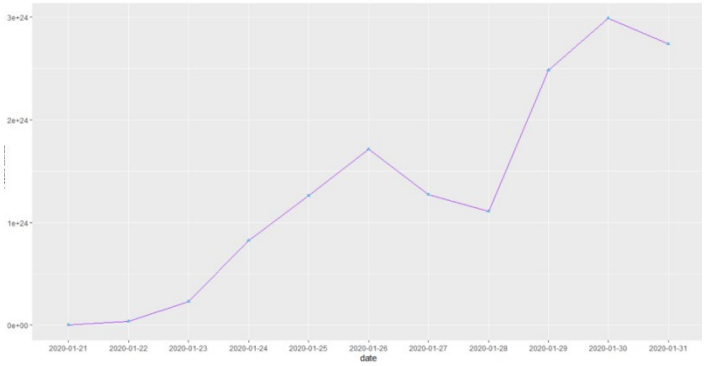


图 4.2(1) 2020.1.21-2020.1.31 每日发推 ID 数量分布曲线

从 2020 年 1 月 21 日至 2020 年 1 月 31 日，推特上推文有关疫情的 ID 数量的趋势是先上升载下降，然后在 1 月 30 日上升至一月份的顶峰。联系实际情况可知，1 月 30 日前后是疫情爆发期，推特的推文数量呈上升趋势是可以理解的。

二月份 ID 数目：

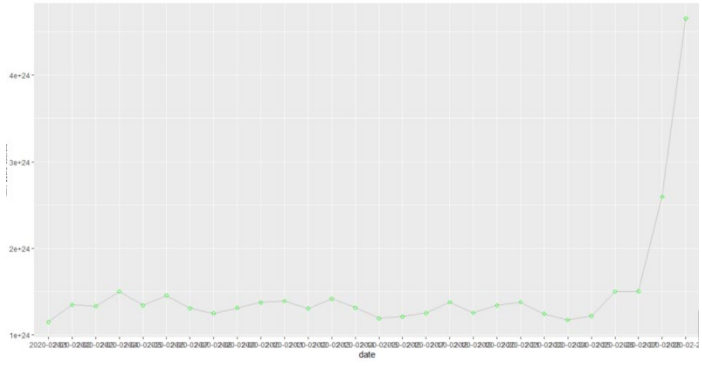


图 4.2(2) 2020 年 2 月每日发推 ID 数量分布曲线

从 2020 年 2 月 1 日至 2020 年 2 月 29 日，推特上推文有关疫情的 ID 数量前期一直是呈一个起伏但变化不大的趋势，最终在从 2 月 27 日开始快速上升，到 2 月 29 日达到峰值。实际上，二月底确实是国外疫情爆发的阶段。

三月份推文 ID 数目：

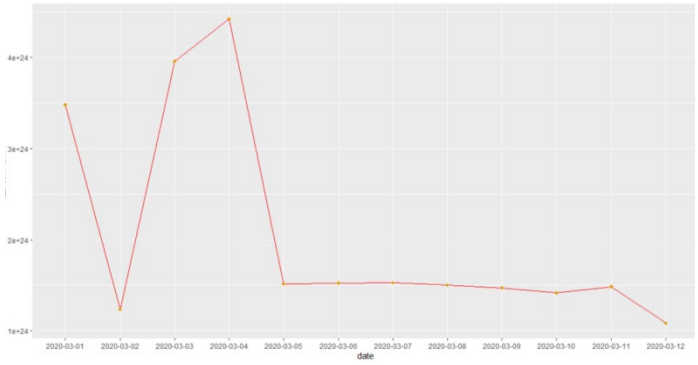


图 4.2(3) 2020.3.1-2020.3.12 每日发推 ID 数量分布曲线

从 2020 年 3 月 1 日至 2020 年 3 月 12 日，推特上推文有关疫情的 ID 数量呈先下降后上升，在 3 月 4 日到达三月份的峰值之后直线下降，后阶段就呈现一个比较低的水平。

推特 ID：

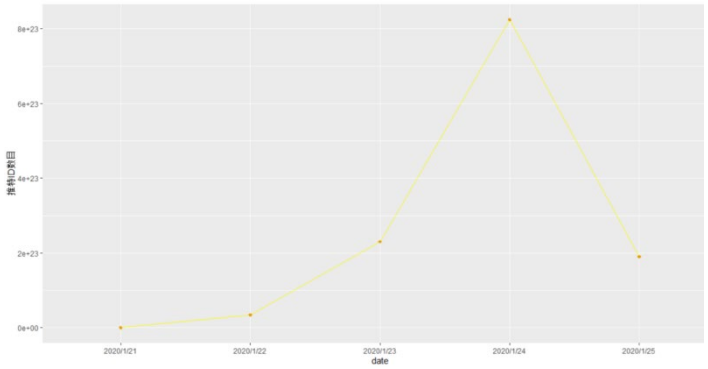


图 4.2(4) 2020.1.21-2020.1.25 每日发推 ID 数量分布曲线

从 2020 年 1 月 21 日至 2020 年 1 月 25 日，推特 ID 数量的趋势是一个从低到高，在 24 日达到顶峰，到 25 日下降。

日推文数量统计：

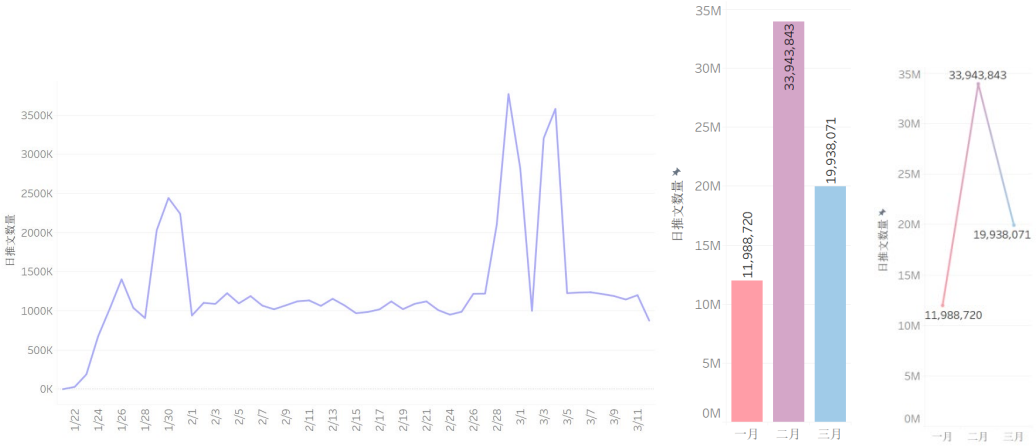


图 4.2(5) 日推文数量分布图

从 1 月 21 日到 3 月 12 日,推特 ID 发文数量顶峰主要出现在 1 月 30 日、2 月 29 日、三月 5 日左右，这三个日期也分别是疫情病例快速上升的几个节点。

对一月、二月、三月疫情的情况进行总计并绘制柱状图和折线图。从一月到三月的整个趋势是由低到高，在二月达到顶峰，在三月下降。和一月、三月相比，二月份的推文数量是最多的。与国内相比，国内的疫情状况也是在二月份达到一个全面爆发的状态，在三

月份局势稍有好转，与国外有着相似的趋势。

4.3 网络社交论坛舆论演化分析

由于提供的数据集并未显示正确或错误的标引标签，我们选择爬取论坛网站上的标题语料。

4.3.1 时间分析

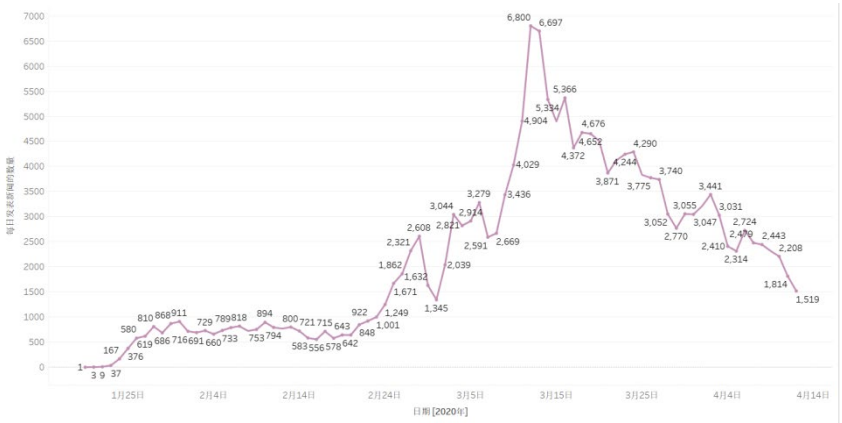


图 4.3.1(1) 论坛网站标题数量随时间分布曲线

从 1 月 20 日至 4 月 12 日，整体的舆论数量大致是呈一个先递增后递减的趋势。详细来看，从 1 月 20 日开始，舆论数量缓慢上升，到 2 月 29 日左右，达到了第一个小高峰，后又下降，在 3 月 5 日左右出现了两个小高峰，3 月 5 日之后舆论数量略微有下降，都没有大量舆论的报道，可能是由于国内的疫情虽然爆发了，但是国外的疫情状况较轻，没有引起广泛的关注；而 3 月 5 日的略微下降之后，舆论数量以一个极快的速度开始增加，在 3 月 12 日达到了这段时间范围内的日舆论数量顶峰，峰值为 6800，3 月初是国外疫情的爆发期，所以舆论数量从 3 月初到 3 月中旬有一个快速增长的阶段，积累了大量的疫情舆论，终于在 3 月 12 日这一天舆论数量达到的顶峰；在 3 月 12 日峰值之后，舆论数量以动荡的趋势逐渐下降，并且下降的速度比 3 月 12 日前上升的速度相较要慢，疫情在 3 月中旬之后，国内外均稍有缓解，但是情况还是不容小觑，所以舆论数量虽然是缓慢减少，但是还是在一个较高的水平。

4.3.2 词频统计和词云图

在对数据集中各个网页的舆论网址源码的 title，进行爬取、整理、筛选之后得到了一个较为规整的数据集，原舆论网址有 7 万多条，在爬取清洗之后得到的有用的 title 数据集有 13682 条，对该数据集进行分词和词频统计处理之后，得到了词频统计的矩阵，去除其中的停用词，得到了最终按词频倒序排列的词频矩阵，按照词频对这些词进行分词处理，绘制词云图如下：

图 4.3.2(1) 网页舆论网址源码 title 词云

图 4.3.2(2) （清理后）网页舆论网址源码 title 词云

在这些舆论中出现词频最高的是“coronavirus”、“news”、“covid”、“cases”、“wordpress”、“times”、“health”、“positive”等词汇，说明舆论对疫情的关注度主要体现在一

下五个方面：疫情的特征、疫情的案例数与发展趋势、疫情舆论的来源途径、疫情的影响、大众对疫情的态度。首先，疫情的特征可以看到，这次的疫情是具有广泛的传染性的，并且传播途径不定；其次，疫情案例越来越多，是呈一个快速上升的趋势，并且没有下降的势头；再者，疫情舆论的来源有博客、舆论刊物等途径；第四，疫情对大众的健康、生活、出行等方面都是有重重影响的；最后，关于大众对疫情的态度，大部分还都是保持一个乐观的态度。

4.3.3 文本主题分析

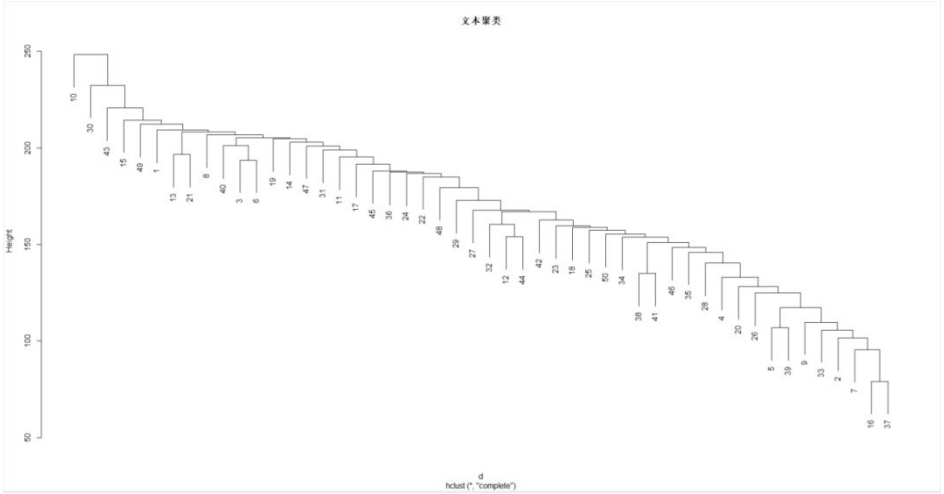


图 4.3.3(1) 网页舆论文本层次聚类分析

对文本分词后构建文档矩阵，筛选词频过低的词汇。使用层次聚类中的离差平方和法实现舆论标题的聚类。如图所示，模型被聚类成为二叉树模型，各级广度为 50、100、150、200 和 250.查看分类器所得特征文档，这些文档反映了国外疫情本身的情况，以及全球疫情的实时动态、影响疫情的一些因素、疫情给人们带来的影响等等，同时还包括了一些舆论本身的信息，例如舆论的属性（经济类、政治类等）还有发表舆论的平台信息等等。

使用 LDA 主题模型对降低稀疏度后的矩阵进行主题聚类。基于层次聚类的结果，分别以 k=3,4,5 为参数构建 LDA 主题训练模型。

K=3 时，文本被分成了三类，分析三类模型的词汇组成。第一类文本词汇主要聚焦于“coronavirus”“health”“positive”“pandemic”“outbreak”“world”，所以第一类文本着眼于病毒本身的特质、该病毒对人们带来的影响以及病毒在世界范围内的传播等；第二类文本词汇主要聚焦于“coronavirus”“republica”“sports”“weather”，可以看到第二类文本在着眼于病毒的同时，还着眼于影响病毒的一些因素，例如天气温度等，并且还会着眼于一些预防病毒的方法，例如做运动提高免疫力等等，第二类文本相比较于第一类文本，主题聚类不是特别明显；第三类文本词汇主要聚焦于“wordpress”“thehill”，第三类文本主要反映的是一些发表舆论的平台信息，例如 WordPress 博客平台、thehill 舆论平台等等。

K=4 时，文本被分成四类，分析四类模型的词汇组成。第一类文本词汇主要聚焦的词汇与 K=3 的第一类文本词汇类似，也都是集中于病毒本身的特质、该病毒对人们带来的影响以及病毒在世界范围内的传播等；第二类文本词汇出现了一些新的词汇例如“economic”“science”“business”这类词汇反映了舆论的类型，即疫情相关舆论是有关经济的报道还是有关科学的报道等等；第三类和第四类文本词汇还是聚焦于平台本身，着眼于一些平台的名称以及与平台相关的信息。

K=5 时，文本被分成五类，分析五类模型的组成。第一类和第二类文本词汇主要聚焦于疫情本身，包括全球的疫情情况、对人们身体健康的影响程度、感染者的死亡情况以及学校

学生的情况等等；第三类和第四类反映的舆论平台本身的信息；第五类是聚焦于疫情舆论的属性与特点，是关于哪一方面的舆论，例如经济、政治、体育等等。

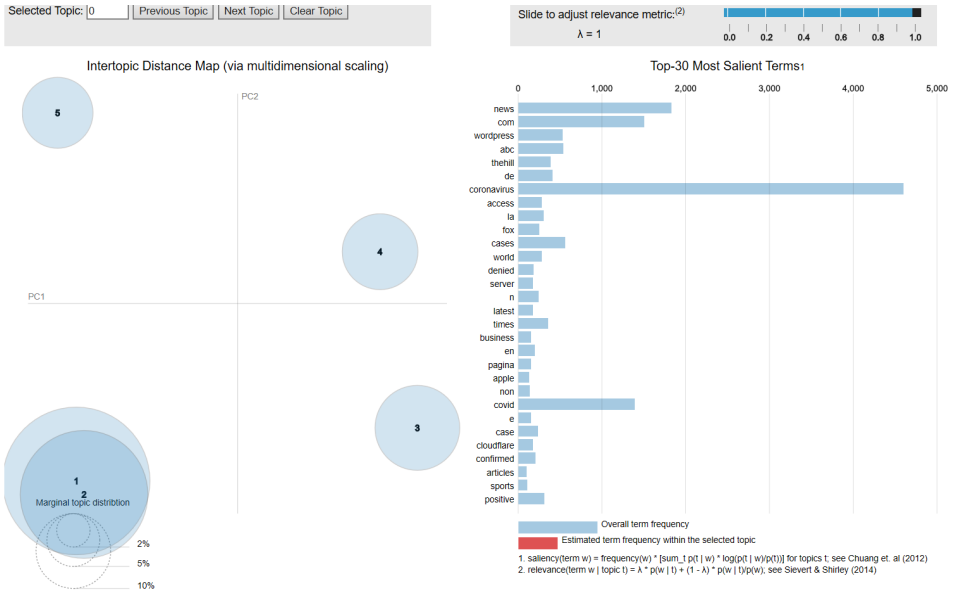


图 4.3.3(2) 网页舆论文本 LDA 主题聚类

5 特殊舆情信息文本挖掘：社会谣言与医学问答

5.1 深入舆情误区：疫情谣言形态和主题分析

利用 DXYRumor 数据集统计所有的传言/谣言文本数据，利用词频统计、词云图绘制、文本主题聚类分析等技术进行分析研究。

5.1.1 总体比例分析

从整体的谣言/传言类型上来看，谣言/传言文本中有 79.577%是错误的内容；有 14.789%是内容尚未确定的传言；在所有的内容中，只有 5.634%是真实正确的内容。

对谣言数据集进行文本挖掘，探索谣言相关消息的总体信息分布特征，首先对 rumor 数据集的 mainSummary 列进行文本分词，总体词云如下所示。利用总体词云查看谣言相关报道的高频关键词。

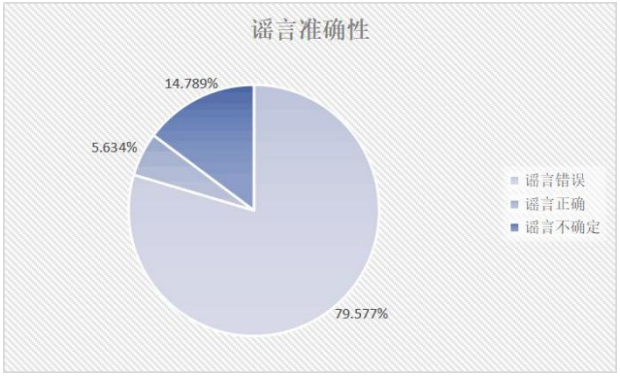


图 5.1.1(1) 谣言准确性占比

5.1.2 词频可视化分析

词云表明谣言相关的报道主要集中在以下的几个方面：关于具体谣言及辟谣的内容，涉及病毒的特性（病毒的命名、病毒的生化特征）、病毒的防控方式（包括公共空间统一防控和个人防护等）、病毒的传播途径（传播方式、传播媒介等）、病毒的临床治疗；关于谣言和辟谣的主体，涉及多个个人和团队（包括在辟谣工作中做出巨大贡献的丁香医生团队，还有

疫情防控方面做出巨大贡献的李兰娟院士、钟南山院士等）。

词汇	词频	频率(%)
辟谣	177	4.4
医生	169	4.2
团队	169	4.2
丁香	166	4.1
病毒	153	3.8
新冠	137	3.4
可以	66	1.6
预防	61	1.5
口罩	53	1.3
感染	52	1.3
肺炎	44	1.1
没有	41	1
消毒	33	0.8
冠状病毒	28	0.7
传播	26	0.6
宠物	26	0.6
可能	23	0.6
目前	22	0.5
不能	22	0.5
新型	22	0.5
作用	20	0.5
传染	20	0.5
证据	19	0.5
效果	18	0.4
通过	17	0.4

图 5.1.2(1) 谣言文本词频图和词云图

5.1.3 文本主题分析

这里的层次聚类法效果并不理想，类与类之间的分布比较不平衡，可能是文章的主干关键词出现频次不够，使得文章没有反映某种主题；或者没有准确分割某些常用词；对建模不利的干扰词没有完全剔除。查看分类器分类所得特征文档，这些文档反映了在谣言/辟谣的新闻文本中：在来源方面，来自研究员和科技部/科技司的言论起了重要的舆论导向作用；在内容方面，针对病毒的分子式的解释和科普对辟谣产生了重要作用，针对各地红十字会的新闻也在谣言/传言文本中占了一定比重。

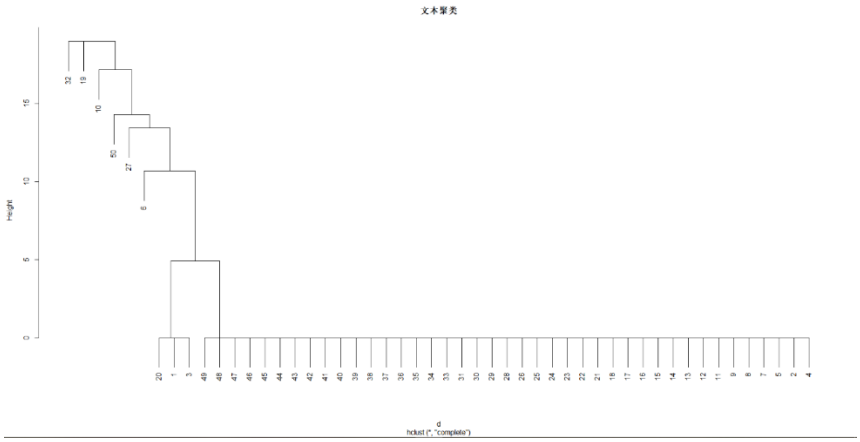


图 5.1.3(1) 谣言文本层次聚类

使用 LDA 主题模型对降低稀疏度后的矩阵进行主题聚类。基于层次聚类的结果，分别以 k=3,4,5 为参数构建 LDA 主题训练模型。

K=3 时，文本被分为三类，分析三类模型的词汇组成。第一类谣言/传言文本大多着眼于病毒本身的特性；第二类谣言/传言文本大多聚焦于疫情防控手段；第三类谣言/传言文本描述总体的疫情情况。

K=4 时，文本被分为四类，第一类谣言/传言文本聚焦于病毒的特征、传播途径、临床特性、防空手段；第二类谣言/传言文本着眼于日常防控手段；第三类文本聚焦于在疫情期间贡献突出的人员团队；第四类谣言/传言文本描述疫情期间的防控政策和规定。第一类与第二类中有交叉重复的部分。

K=5 时，文本被分为五类有的类别描述关于疫情的总体数据情况；有的类别描述我国关于疫情的防控工作，总体上分类主题的界限变得模糊不清晰。

根据以上情况综合发现 k=3 时的分类主题结果最清晰。

5.1.4 总结

由于网络信息传播具有迅速的特点，传言中谣言（或是不确定真假的信息）含量高，且谣言产生速度快，传播范围广，涉及主题多。这需要读者和观众、网络用户提高甄别能力，从正规官方渠道获取新闻和科普，不受到谣言的干扰和鼓动，不制造和传播恐慌，传播正确信息，构建和谐健康的网络环境。

5.2 舆情背后的医学情绪：疫情医学问答数据分析

5.2.1 中文对话数据文本挖掘

5.2.1.1 探索性数据分析

该数据集描述了好大夫在线网站疫情期间和肺部疾病相关的医学对话，共计对话篇幅 399 条，共记录了 3761 条医生对话和 4600 余条病人对话内容，及相关篇目的病情描述和解决情况。

5.2.1.1.1 患者病情描述和咨询情况

患者最常见的疾病咨询需求为肺炎，肺部感染相关疾病。和肺炎相关的咨询占疾病总数的 86%以上，和新冠肺炎相关的预防、诊断与治疗咨询相关并明确提到新冠肺炎相关术语的疾病咨询占 36.7%。

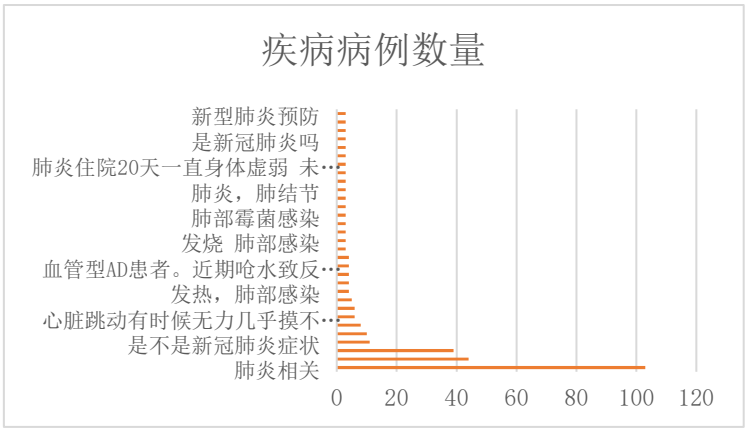


图 5.2.1.1.1(1) 疾病病例情况数量分布

从中文医学对话数据集的病情摘要和初步印象内容来看，患者进行的咨询主要有呼吸器官感染和炎症，感冒，肺部疾病咨询以及由病情引发的心理状态异常。在疫情期间，和肺炎、新冠肺炎相关的咨询数量较多，而心理状态异常（焦虑，抑郁）的病例达到 10 例以上。病人普遍具有敏感多疑情绪是疫情期间肺部疾病咨询的显著特征。

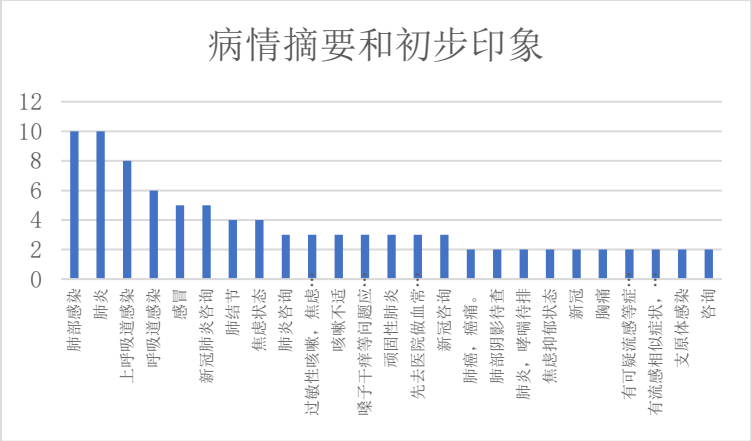


图 5.2.1.1.1(2) 病情摘要和初步印象分布情况

分析病人希望获得帮助的词频，病人的问题集中在病情诊断和检查，如何治疗，症状疑难解答等。分析病人病情描述，患者出现的主要病情症状有咳嗽，肺炎，发烧，发热，感染，胸闷等。

myfile.word	Freq
肺炎	97
需要	67
新冠	55
医生	52
什么	51
医院	46

怎么	45
治疗	42
检查	37
如何	37
是否	33
感染	30
没有	22

引起	22
症状	18
建议	17
解答	17
病情	16

5.2.1.1.2 医生建议和解决方案

医生多次给出的建议有：治疗，复查，CT 检查或其他检测，抗炎，对症下药，观察，休息，多喝水，注意饮食等。

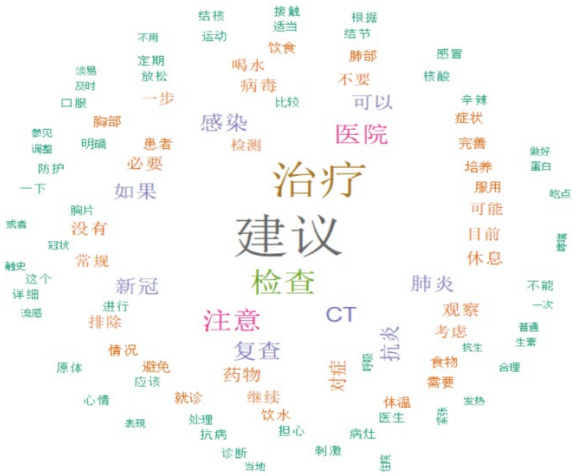


图 5.2.1.1.2(1) 医生建议和解决方案词云图

5.2.1.1.3 医学对话内容文本分析

基于患者对话数据构造词频为前 100 的词云。患者对话中，“医生”，“没有”和“医院”出

现的频率最高。在高频词词云中，“新冠”，“肺炎”等词汇出现频率位列前 50，显示了患者对新冠肺炎的焦虑和自我怀疑程度较高。在疾病描述上，患者们出现的症状包含以下几个共同特征：胸闷，心慌，发热，喉咙不适，咳嗽；患者预先采用的常用手段有医院门诊，测量体温，吃药和 CT 检查等。

图 5.2.1.1.3(1) 病人对话内容词云分布图

医生的对话内容大多是基于患者的问题进行的针对性答复。医生对话中，“没有”，“可以”，“如果”出现的频率最高，是医生诊断中的判断和假设。医生的对话主题侧重于核实病情症状和诊断与治疗方案建议。医生对话中体现的高频动词是阅读，检查和治疗，这说明医生对患者病情的实际状况考察不局限于患者病情的描述，还基于实际的医学数据。对于患者的症状和病情，医生对话中也出现了和新冠肺炎病毒相关的关键词。从整体上来讲，医生对话数据揭示了医生诊断的常见建议或行为为：胸部检查（如 CT 等），核酸等筛查，医学检查数据研读，病人身体体征观察，消炎等。

5.2.1.2 词典型情感分析

为了提高情感分析的准确性,在引入知网 HotNet 和 ntusd 情感词典的基础上，根据医学对话数据集的对话词语情感含义，适当的增加一些褒贬义训练词汇。由于缺少对话的时间信息，我们以分离后的医生和病人对话文本作为分析对象。

医生的词典型情感得分为 928 分，褒义词在种类上比贬义词更占优势。从情感词的分布上可以看出，医生对话中词频较高的正面词汇偏向于安慰，舒缓病人的情绪和对病人病情的积极反馈。医生对话中词频较高的负面词汇偏向于对病人负面情绪的描述和症状的解读。从整体来看，医生对话中展现了明显的积极情感极性。

word	weight	label
感谢	621	1
可以	424	1
详细	412	1
信任	402	1
可能	213	1
客气	201	1
治疗	174	1
正常	156	1
明显	124	1
考虑	121	1
不用	110	1

最好	86	1
注意	82	1
常规	75	1
健康	72	1
明确	66	1
一定	65	1
主要	65	1
肯定	58	1
必要	54	1
确实	54	1
容易	54	1
放心	47	1

清楚	46	1
安全	42	1
舒服	42	1
适当	40	1
放松	39	1
休息	37	1
流行	36	1
确定	30	1
帮助	29	1
特别	29	1
合理	28	1
希望	28	1

有效	26	1
完全	24	1
知道	24	1

方便	22	1
改善	22	1
决定	22	1

一样	22	1
专家	20	1

正向情感高频词汇

病人对话文本的词典型情感模型的得分为-1843，负向情感词的种类明显多于正向情感词汇。这说明病人对话中表现了明显的负向情感情绪。综合观察正向和负向高频情感词汇，病人对话中表达了病人对病情的焦虑，恐惧。在以肺部疾病为主题的咨询中，病人对症状的叙述偏向于使用消极、夸张的词汇；对情绪的描述也偏向于焦虑、恐惧和痛苦。

压力	-17	-1
空洞	-18	-1
复发	-23	-1
不知	-26	-1
打扰	-26	-1
反复	-27	-1
难受	-28	-1
危险	-28	-1
头痛	-30	-1
突然	-31	-1
无力	-33	-1
东西	-36	-1
异常	-38	-1
可怕	-39	-1

不要	-40	-1
厉害	-40	-1
老是	-44	-1
慢性	-48	-1
失眠	-50	-1
疾病	-51	-1
传染	-53	-1
困难	-54	-1
怀疑	-60	-1
隔离	-64	-1
害怕	-68	-1
心慌	-68	-1
不能	-69	-1
发热	-69	-1

不到	-90	-1
发烧	-96	-1
麻烦	-144	-1
不好	-156	-1
感冒	-159	-1
不会	-165	-1
感染	-186	-1
紧张	-186	-1
焦虑	-252	-1
担心	-322	-1
问题	-370	-1
严重	-429	-1

负向情感高频词汇

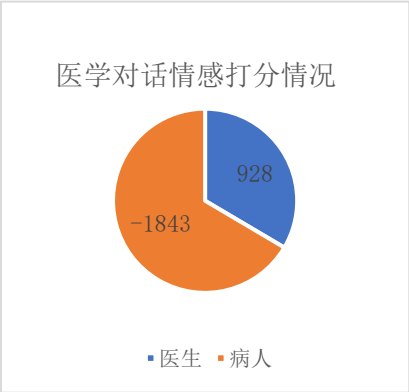


图 5.2.1.2(1) 医学对话情感打分情况

从理论上来讲，词典型情感分析的精确度不高。但医生和病人情感极性差异说明，在COVID-19 疫情期间，病人的负向情感倾向显著，肺部疾病患者呈现出较多对病情的担忧和恐惧情绪；医生在对话中呈现较为积极和克制的情绪，会对患者进行定向的情绪疏导。

5.2.1.3 文本聚类

1. 层次聚类

对医生对话文本分词并将 dtm 稀疏系数设定为 0,9993，得到 254 个词汇，据此建立层次聚类模型。模型根据词向量距离将词汇聚为 3 类，查看词汇分布，第一类是医生对病人问题的判断，大多回答为“没有”，“有”和是否；第二类与肺炎相关；第三类是医生对话

中的其他文本内容。对病人对话文本分词，dtm 的稀疏系数设为 0.9994.得到 256 个词汇，模型将词汇聚为 2 类，第一类描述了病人的感觉；第二类是病人口述的其他信息。

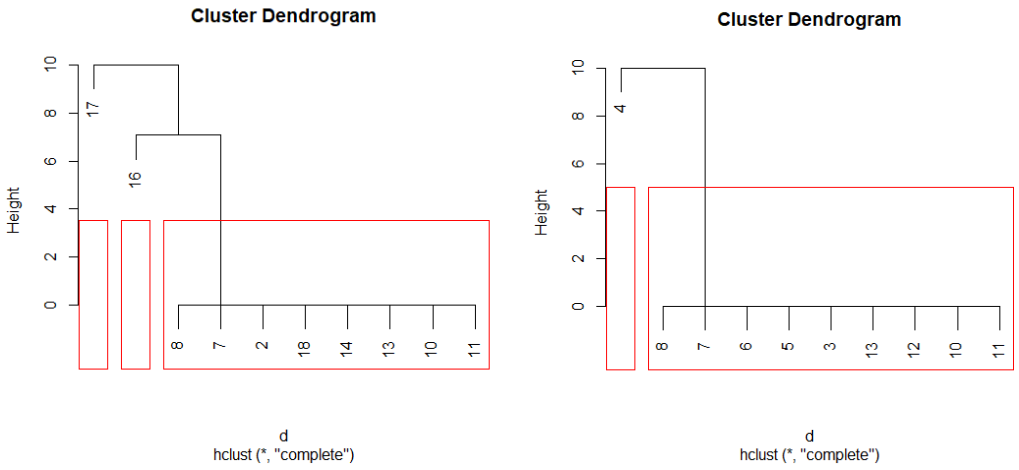


图 5.2.1.3(1) 医生对话内容文本层次聚类

2. kmeans 聚类

由于 kmeans 崽文本聚类中的效果较差，这里略去不做讨论。

3. 关联规则

关联规则算法发现，在医生对话中，“感染”，“可以”，“没有”和“如果”出现的频率最高，而“资料”，“肺炎”，“病情”，“医院”，“问题”，“感染”等的词语相关度较高，说明病人的医学资料，病情和所在诊医院在医生对话中具有较高的相关度。暂时没有挖掘到可用的关联词语规则。

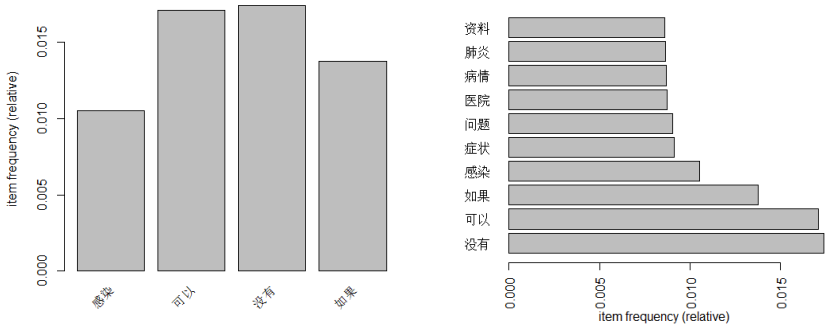


图 5.2.1.3(2) 医生对话内容关联规则挖掘

在病人对话中，“知道”，“这个”，“咳嗽”，“谢谢”，“感觉”，“什么”等词汇的相关度较高，说明病人在对话中缺乏医学知识，对自身症状的描述大多出于主观性描述，对医生的信任，感谢和依赖程度较高，在对话过程中，病人的疑问，症状，检查和医院联系紧密。暂时没有挖掘到可用的关联词语规则。

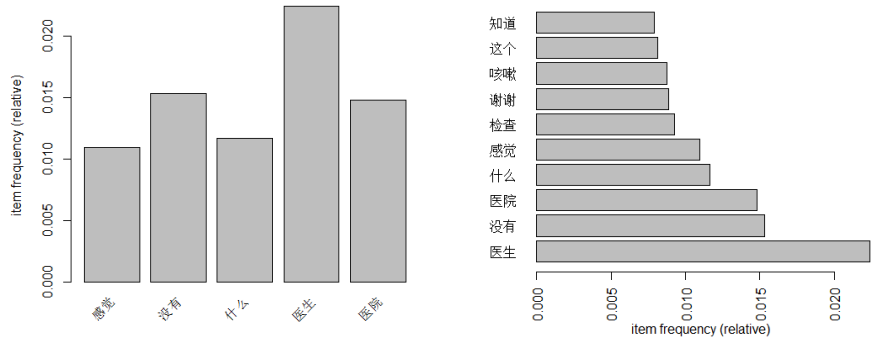


图 5.2.1.3(3) 病人对话内容关联规则挖掘

4. 主题分析

对医生建议的文本进行 LDA 主题分类，将主题分为三类。并将分类情况进行可视化（挑选出每类出现频次最高的词语）：

TOPIC1：针对出现肺炎感染病症，比如咳嗽、发烧等症状的病人提出的建议。该类别多次出现“检查”、“注意”等关键词，表明在出现咳嗽、发烧等感染病症时，要及时就医检查，确定是否为新型冠状病毒，避免个人和家人的恐慌，同时也是为社会负责。

TOPIC2：是针对疫情期间的恐慌心理的医生建议。该类别多次出现“焦虑”、“担心”等词语。，这些情绪都是疫情下的正常反应，经过对症药物治疗和心理疏导，并且得到很好的休息，可以减轻这种担忧。

TOPIC3：是针对已经确诊感染肺炎患者的医生建议。该类别多次出现“胸部”、“核酸”等确诊肺炎的专业词语。确诊新型冠状肺炎一般是取咽部或鼻部以及血液标本检测新型冠状病毒核酸阳性。还有取呼吸道或血液标本病毒基因测序，与已知的新型冠状病毒高度同源，此时也可以确诊。目前在医院用的主要是检测病毒核酸的方法，在没有确诊之前属于疑似病例仍然需要隔离，隔离期间会及时根据症状来使用药物治疗。

5. 共现分析

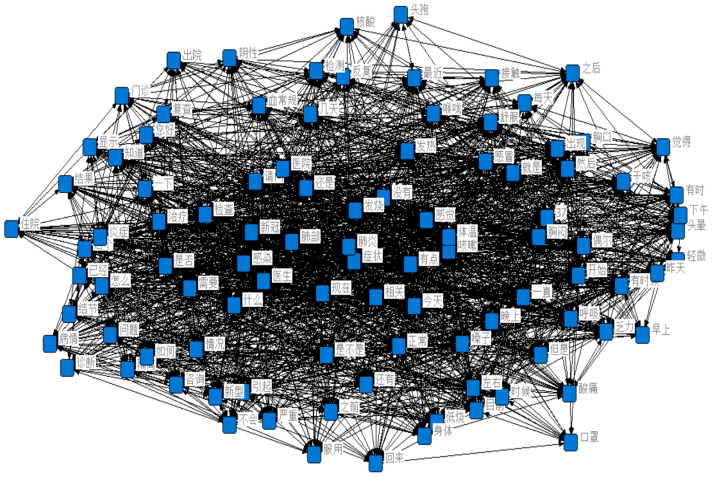


图 5.2.1.3(4) 对话数据集病情描述共现词矩阵

在病情描述中，根据共现词矩阵构成的词频共现图显示，“肺炎”、“肺部”、“医生”、“发烧”、“体温”等词语位于图的中心位置，与其他词语息息相关。测量体温是目前防治新型冠状病毒感染非常重要的手段。感染新型冠状病毒以后，最主要的表现可能是

发热，所以测量体温是发现病人非常简单的一种方式。所以疫情期间，要及时测量体温，遇到发烧、咳嗽等症状要及时就医检查肺部，进行核酸测试。

5.2.2 英文对话数据文本挖掘

对数据集中的病人、医生、description 三个数据集进行分词和词频汇总：

词频比较高的词语有：covid、pneumonia、symptoms、throat、cough、video、virus、brief、chat、text、fever 等，是和新冠的主要症状与特点的词语。

将描述文本分为四类。第一类文本是患上新冠之后的明显的症状，例如头疼、发烧、咳嗽等。第二类文本是疫情在发生过程中的一些发展的特点与趋势。第三类文本是疫情可能会带来的一些风险与可能陷入的险境。第三类文本是新冠的病患会出现的一些情况，主要体现在行为方面，而不是身体出现的问题。第四类文本是面对疫情可以采取的一些措施与方案。

将病人对话文本分为三类。第一类文本词汇主要聚焦于“pneumonia”，“cough”，“lung”，“years”，“day”，“antibiotics”，“diagnosed”，“hospital”等词汇，其中“pneumonia”出现的次数最多，可以看出第一类文本主要反映的是疾病的性质是肺炎以及是与人肺有关，还反映了感染者会出现的症状有咳嗽等，此外还有关于病毒的潜伏期、患病时间等与时间有关的信息；第二类文本词汇主要聚焦于“cough”，“throat”，“symptoms”，“fever”，“headache”，“breathing”这些词汇，这类词汇明显主要反映的是新冠肺炎感染者会出现的种种症状；第三类文本词汇主要聚焦于人们感染新冠病毒的风险，以及一些应对新冠肺炎的措施，应该待在家中不要出去。

将医生对话文本分成三类。第一类文本词汇主要聚焦于“chat”，“symptoms”，“stay home”，“contact”，“infection”等，这些词汇主要反映了该文本的性质是医生与病人之间的对话，以及感染新冠肺炎后会出现的一些相关症状，此外还反映了预防新冠肺炎的一些措施，例如待在家中，避免与他人的直接接触等等；第二类文本词汇主要聚焦于感染新冠肺炎后会出现的一些症状以及该疾病的性质是肺炎一类的，此外还反映了医生建议病人要做好健康护理并且做好疾病的治疗。第三类文本主要反映了文本的性质是两者之间的对话，高频词中包含了许多打招呼用语，例如“hello”，“welcome”，“hi”等，以及医生给病人提出的一些关于健康的建议。

6 总结与问题回答

6.1 疫情舆情的重要时间节点

6.1.1 全球疫情舆情总体时间节点分布

从全球的新闻报道数量来看，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。

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。跟踪与疫情舆论传播密切相关的推特 ID,从 1 月 21 日到 3 月 12 日,推特 ID 发文数量顶峰主要出现在 1 月 30 日、2 月 29 日、三月 5 日左右，这三个日期也分别是疫情病例快速上升的几个节点。观察网络社交论坛的舆论走势，从 1 月 20 日至 4 月 12 日，整体的舆论数量大致是呈一个先递增后递减的趋势。从 1 月 20 日开始，舆论数量缓慢上升，到 2 月 29 日左右，达到了第一个小高峰，后又下降，在 3 月 5 日左右出现了两个小高峰，3 月 5 日之后舆论数量略微有下降；而 3 月 5 日的略微下降之后，舆论数量以一个极快的速度开始增加，在 3 月 12 日达到了这段时间范围内的日舆论数量顶峰，舆论数量从 3 月初到 3 月中旬有一个快速增长的阶段，积累了大量的疫情舆论，在 3 月 12 日达到峰值；舆论数量以动荡的趋势逐渐下降，并且下降的速度比 3 月 12 日前上升的速度相较要慢，舆论数量虽然缓慢减少，但是还是在一个较高的水平。

6.1.2 国内疫情新闻舆情时间节点分布

根据时间分布曲线的增长特征，可以将疫情国内新闻舆情大致分为 4 个时期：2019 年 12 月 31 日至 2020 年 1 月 24 日为平稳低增长时期，疫情新闻处于较低投放、低关注的形势；2020 年 1 月 25 日至 3 月 14 日为波动缓冲时期；3 月 15 日至 3 月 23 日为焦点顶峰时期，疫情新闻事件呈现高度饱和的状态，疫情受到全球广泛关注，有关疫情的相关报道被广泛撰写和投放，疫情成为密集型新闻热点话题；3 月 24 日起为冷却时期，武汉疫情事件的话题度降低，信息饱和并降温。可以将 3 月 23 日之后的新闻舆情进行进一步细分：3 月 22 日起至 4 月 23 日期间，以 4 月 9 日为分界点，4 月 9 日之前新闻数量随时间虽然也是呈大幅度波浪分布，但总体的新闻数量高于 4 月 9 日之后的新闻数量。

6.1.3 国外疫情新闻舆情时间节点分布

在新闻报道上，根据丁香园的通用新闻数据，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。CBC 数据集记录了 CBC 新闻报道数量在 2020 年 1 月 8 日至 3 月 27 日期间的时间分布。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，3 月 27 日得到暂时下降（有可能由于数据集统计不完全导致）。

6.2 疫情爆发前期、中期、后期的舆情走向趋势

全球确诊人数和死亡人数先下降后上升，在总体上波动增长，确诊人数和死亡人数都在 4 月 17 日达到最高值（分别为 27,705,523 和 10,298,990）；相对应的，治愈人数也在波动上升，同样在 4 月 17 日达到最大值 1,655,921 人。

1 月底至 2 月上旬是国内疫情爆发前期，确诊数量不断增长；2 月中旬是国内疫情爆发的中期，疫情确诊数量达到峰值；2 月中下旬至 3 月中旬为国内疫情爆发后期，疫情确诊人数下降，疫情得到控制；3 月上中旬是海外疫情爆发的前期，疫情确诊人数逐渐增多；3 月中下旬是海外疫情爆发的中期，在 4 月上中旬攀升到疫情增长的顶点；4 月中旬以后，海外疫情从总体上向后期发展。

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。在国内疫情爆发的前期，在 2020 年 1 月 25 日至 3 月 14 日期间，国内新闻舆情波动缓冲增长；在国内疫情爆发的中后期，3 月 15 日至 3 月 23 日为焦点顶峰时期，疫情新闻事件呈现高度饱和的状态；3 月 24 日起为冷却时期，武汉疫情事件的话题度降低，信息饱和并降温。国内新闻舆情也呈现对全球疫情爆发时期的高度相关关系：3 月 22 日起至 4 月 23 日期间，以 4 月 9 日为分界点，4 月 9 日之前新闻数量随时间虽然也是呈大幅度波浪分布，但总体的新闻数量高于 4 月 9 日之后的新闻数量。而海外新闻，以 CBC 为例，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。

在疫情期间的网络舆情数量分布上，推特 ID 发文数量顶峰主要出现在 1 月 30 日、2 月 29 日、三月 5 日左右，这三个日期也分别是疫情病例快速上升的几个节点。在网络社交论坛讨论话题上，2 月下旬，3 月中上旬的拐点，正好对应了国内外疫情爆发的分期节点。

6.3 疫情中不同时期公众的关注热点和程度

6.3.1 总体疫情舆情关注热点事件

1 月 29 日，中国还是应对新冠肺炎的重要国家，以中文报道居多新闻报道数量达到了短时间内的最高值，拉萨市卫生健康委员会发布拉萨首例新型冠状病毒感染的肺炎疑似病例的相关情况，并且西藏启动重大突发公共卫生事件一级响应；截至 1 月 29 日为止中国内地 31 省份全部启动突发公共卫生事件一级响应。

3 月 23 日，特朗普宣布华盛顿州为新馆肺炎疫情的“重大灾区”，美国多个州关闭参观、酒吧等公共餐饮场所。并且自 3 月 23 日起，美国肺炎确诊病例数迅速上升。而中国新冠肺炎疫情控制状况良好。

4 月 6 日左右，我国 31 省份新增新冠肺炎境外输入确诊病例，境内确诊病例均以输入兵力为主；西班牙新冠肺炎确诊超过 12.4 万，政府延长国家紧急状态；英国新冠肺炎死亡率又迅速超过西班牙；意大利确诊病例近 12 万。

6.3.2 国内疫情新闻舆情关注热点及程度

2019 年 12 月 31 日至 2020 年 1 月 24 日

2019 年 12 月 31 日至 2020 年 1 月 24 日这段时间，新闻标题的主旋律是疫情确诊人数累计和新型冠状肺炎病毒。其他的重要主题为：新型肺炎的感染和紧急措施的采取；“一级”，“感染”等关键词——疫情的预防，诊断，治疗；从总体词语分布上来看，词汇分布在社会民生，疾病医疗、公共安全，经济发展和政治军事这多个方面。防疫工作的开展、病毒的传播情况和疫情对社会的影响在新闻中占据重要地位。

2020 年 1 月 25 日至 2020 年 3 月 14 日

2020 年 1 月 25 日至 2020 年 3 月 14 日，从高频词语内容可以看出处于病毒发现、感染与治疗治愈的初期阶段，主要内容是针对新型肺炎病毒确诊与治愈进行报道。

2020 年 3 月 15 日到 2020 年 3 月 23 日

3 月 15 日到 3 月 23 日，在这一时期，肺炎事件的重要性和紧急性达到顶点。和肺炎疫情相关事件的主题次序为：肺炎确诊、病例新增和境外输入。新闻态度上，从上到下，从时间到空间上都展现了对疫情形势的高度关注和全民全面防控的迫切和紧急心。

2020 年 3 月 23 日后

从 3 月 24 日的统计关键词中，变化情况如下：“新增”与“确诊”说明新型冠状病毒肺炎的确诊病例和新增情况变为新闻关注的热点；世卫和医疗队高频出现；患者的症状、治疗和疫情隔离情况，政府的病情通报和卫生安全工作受到密切追踪。

对于 3 月 23 日以后的新闻报道，可以以 4 月 9 日为节点进行更精细的划分。

4 月 9 日之前的高频词与整体的高频词是大体是一致的，主要都是围绕“病例”“确诊”“新增”“肺炎”“冠状病毒”等，即新闻的主要维度是新冠病毒感染确诊情况、治愈出院情况等等。在 4 月 9 日之后，与国外实时疫情相关的新闻开始出现与增多。

新闻舆情的主题分类如下：

TOPIC1：肺炎患者发现和确诊统计。该类别多次出现“确诊”，“新增”，涉及的高频词大部分针对疫情中的感染者的基本统计，以便为抗疫工作更好的开展。
TOPIC2：全国疫情发展和抗疫行动开展。该类别以武汉、广东等疫情重灾区和相关事件为主要关注点，层层推进，关注全国各个疫情灾区的工作部署、疫情发展。
TOPIC3：疾控政策与政府防疫动向。该类别的行为主体是党员及医疗集体，涉及的词汇包含了政府的官方措辞和权威发布，语句较为严谨，并提及了多种会议和公共的卫生措施。
TOPIC4：国内抗疫成效和境外输入。该类别涉及国内和国外的疫情发展，在国内疫情取得突破性进展的同时，国外疫情形势严峻，境外输入成为国内肺炎病毒不断传播的重要原因。

6.3.3 国外疫情新闻舆情关注热点及程度

加拿大在关注加拿大本地的新冠疫情发展态势的同时，同样也很关注中国的疫情发展状况，以及如何应对疫情、对健康的影响以及医学方面对疫情是否有突破等方面。新闻热点主要以新冠病毒及症状，地区确诊情况，政府公共卫生安全措施，社会民生和医疗健康等几个维度展开。

以 3 月 11 日为分界点，对比疫情发作期和爆发期期间的 CBC 新闻报道内容，前者的新闻主体内容在于强调疫情病毒的风险，症状，来源，并探寻疫情的防控和传播来源，从具体上来说，与中国武汉相关和携带病毒的乘客相关的关键词词频较高，将此次病毒和 SARS 相对比，以疫情传播和数据的报道较多；3 月 11 日以后，与“中国”“武汉”“确诊”相关的关键词词频明显降低，新闻报道的主要内容在于北美各地区的疫情防控和疫情期间的社会运转与资源分配等相关问题。

在 3 月 14 日至 23 日期间，丁香园数据集中国外主流媒体的新闻报道的主要内容以疫情聚焦与防控，政府的主要应对措施和国家生命健康财产危机为主，表现了对政府政治行为的高度关注。

国外新闻舆情的主题分类如下：

第一类大致为教育和就业在疫情中受到的影响；第二类为政府决策和防治措施；第三类为各地区的社会民生；第四类为经济状况和资本变动；第五类为疫情医疗和社会联动情况。

丁香园数据集中提供的其他海外主流媒体相关新闻报道可以划分为政府政治动向，疫情防控，社会民生（以各社会群体疫情隔离和医疗为主）。

6.3.4 疫情网络舆情关注热点及程度

推特推文相关数据展现了在推特平台上广为流传的网络舆论。

3 月 22 日，twitter 上的讨论量不高，讨论重心在意大利和中国（Italy、China）的流行病疫情（pandemic）上。3 月 23 日起，讨论量有所上升，在突然爆发的疫情危机中（crisis、outbreak）导致的公共场所和机构的关闭（lockdown）也引起了大量关注；美国总统特朗普（DonaldTrump、Trump、president）和民主党（Democrats）在本次疫情中的行动也受到了大量的讨论。

3 月 24 日至 3 月 25 日，疫情引起进一步的讨论（pandemic、corona），而在此基础上，查尔斯王子（Charles、Prince）确诊新冠肺炎也引起了一定的讨论。

3 月 26 日至 3 月 28 日，在此前的讨论基础上，居家隔离、隔离期（quarantine、stayhome、stayhomeandstaysafe）、受疫情影响而关闭公共场合（coronalockdown）等概念也受到了关注。

3 月 29 日，特朗普政府因应对疫情不力广受质疑，有声音认为美未能在全球抗疫中发挥领导力（华盛顿邮报），并且美接受了中国送来的物资，这一点也引起了大量讨论。

3 月 30 日至 4 月 4 日，阿根廷、印度尼西亚等国家相继进入严备的疫情境界状态。3 月 30 日，阿根廷新增新冠肺炎确诊病例 146 例，累积 966 例，阿根廷政府（gobierno）正式宣布延长“全民隔离”。3 月 31 日，印度尼西亚总统宣布国家进入卫生紧急状态，禁止外国人入境，印尼工作人员在王宫建筑群、伊斯兰教堂等多个场合喷洒消毒剂（penyemprotan）的照片也引起了讨论。

4 月 5 日，美国总统特朗普与印度总理纳伦德拉·莫迪（Narendra·Modi）进行会面，讨论了新冠肺炎疫情和供应链问题，强调了希望摸底放行美国订购的羟氯喹（hydroxychloroquine，预期成为治疗新冠病毒的一种成功药物）。与此同时，全球新冠肺炎累计确诊超过 120 万例，西班牙成为欧洲疫情最严重的国家（pandimia）。

4 月 6 日，英国首相鲍里斯·约翰逊（Johnson）的严重病情引起了广泛的讨论，而英国首相府称约翰逊未诊断出肺炎。

网络社交论坛数据则从另一维度揭示了疫情舆情的关注热点。舆论对疫情的关注度主要体现在一下五个方面：疫情的特征、疫情的案例数与发展趋势、疫情舆论的来源途径、疫情的影响、大众对疫情的态度。对社交论坛舆情主题分析显示，被分成五类时，第一类和第二类文本词汇主要聚焦于疫情本身，包括全球的疫情情况、对人们身体健康的影响程度、感染

者的死亡情况以及学校学生的情况等等；第三类和第四类反映的舆论平台本身的信息；第五类是聚焦于疫情舆论的属性与特点，是关于哪一方面的舆论，例如经济、政治、体育等等。

6.3.5 舆情误区（谣言）关注热点及程度

谣言相关的报道主要集中在以下的几个方面：关于具体谣言及辟谣的内容，涉及病毒的特性（病毒的命名、病毒的生化特征）、病毒的防控方式（包括公共空间统一防控和个人防护等）、病毒的传播途径（传播方式、传播媒介等）、病毒的临床治疗；关于谣言和辟谣的主体，涉及多个个人和团队（包括在辟谣工作中做出巨大贡献的丁香医生团队，还有疫情防控方面做出巨大贡献的李兰娟院士、钟南山院士等）。

谣言可以大致分为三类。第一类谣言/传言文本大多着眼于病毒本身的特性；第二类谣言/传言文本大多聚焦于疫情防控手段；第三类谣言/传言文本描述总体的疫情情况。

6.3.6 疫情医学咨询关注热点及程度

患者最常见的疾病咨询需求为肺炎，肺部感染相关疾病。

患者进行的咨询主要有呼吸器官感染和炎症，感冒，肺部疾病咨询以及由病情引发的心理状态异常。病人普遍具有敏感多疑情绪是疫情期间肺部疾病咨询的显著特征。

病人的问题集中在病情诊断和检查，如何治疗，症状疑难解答等。患者出现的主要病情症状有咳嗽，肺炎，发烧，发热，感染，胸闷等。

患者对话中，“医生”，“没有”和“医院”出现的频率最高。在疾病描述上，患者们出现的症状包含以下几个共同特征：胸闷，心慌，发热，喉咙不适，咳嗽；患者预先采用的常用手段有医院门诊，测量体温，吃药和 CT 检查等。

医生的对话内容大多是基于患者的问题进行的针对性答复。从整体上来讲，医生对话数据揭示了医生诊断的常见建议或行为为：胸部检查（如 CT 等），核酸等筛查，医学检查数据研读，病人身体体征观察，消炎等。

6.4 不同关注热点的受关注原因

6.4.1 国内疫情关注热点受关注原因

2019 年 12 月 31 日至 2020 年 1 月 24 日

这个时期属于疫情初期，病毒正在全国蔓延，但并未引起社会高度重视。新闻报道标题以发现肺炎感染者，以达到警示目的。

2020 年 1 月 25 日至 2020 年 3 月 14 日

这些新闻报道说明了国家和政府关注本次疫情，并采取相关措施，目的在于防控疫情，救治病患和阻击病毒。态度上是积极乐观，充满信心的。

这个时期属于疫情中期，病毒已经在全国肆虐，但引起社会恐慌。新闻报道标题以确诊和治愈肺炎感染者，以达到安抚目的，避免民众的过度恐慌造成社会损失。

2020 年 3 月 15 日到 2020 年 3 月 23 日

从空间来看，包含的地点关键词种类激增，说明疫情已经扩散向全球，并从境外不断向国内输入肺炎感染者，全球肺炎蔓延趋势令人担忧；在疫情防控方面出现了更多强制性、具体化的措施，着重强调了一线封城、公安等部门的部署、严格检疫，监管等工作；全国各地的社区卫生安全防护工作被深度重视；全面全民抗击肺炎、宣传肺炎知识、驰援重度感染区被认为是刻不容缓的国家大事。

这个时期属于疫情中后期，病毒已经在向全球扩散，情况严重。新闻报道标题以境外输入肺炎感染者，表明抗击疫情已经变成了全球的责任，以达到呼吁共同合作抗疫的目的。

2020 年 3 月 23 日后

世卫和医疗队在关键词中的高频出现说明有关肺炎的官方调查情况和防治相关研究受到广泛关注。对比前期，这段时间关于病毒的情况已经得到更为完善的确认，新闻的主题从认知病毒感染迁移到防控工作中。

这个时期属于疫情后期，病毒已经在全球肆虐，确诊人数不断增多，疫情严重性达到了新的高度。国内疫情已得到较好控制，但国外疫情不容乐观。4 月 9 日左右，国外新冠确诊人数大幅度增长，因此与国外实时疫情相关的新闻开始出现与增多。

6.4.2 国外疫情关注热点受关注原因

新闻开始以经济，政治，教育就业等一系列维度为角度剖析和报道政府和各地对应疫情做出的重大决策和调整，这恰说明了疫情形势严峻和人们对疫情的研究进展推进，导致国家和政府面临的问题更加焦灼和复杂。加拿大对有关病毒和疫情的来源认知为来源于中国；CBC 将疫情的定性为危急，目前的疫情形势明确为爆发。诸多关键词体现了 CBC 对加拿大的疫情具有高关注度。

在 3 月 14 日至 23 日期间，政府政治行为的高度关注。

在网络舆情关注热点上，有以下几个受关注原因：

首先，疫情的特征可以看到，这次的疫情是具有广泛的传染性的，并且传播途径不定；其次，疫情案例越来越多，是呈一个快速上升的趋势，并且没有下降的势头；再者，疫情舆论的来源有博客、舆论刊物等途径；第四，疫情对大众的健康、生活、出行等方面都是有重重影响的；最后，关于大众对疫情的态度，大部分还都是保持一个乐观的态度，防疫宣传中心理宣传引导是全民抗疫中重要的环节。

6.4.3 医学咨询关注热点受关注原因

在高频词词云中，“新冠”，“肺炎”等词汇出现频率位列前 50，显示了患者对新冠肺炎的焦虑和自我怀疑程度较高。

病人对话中表现了明显的负向情感情绪，病人对话中表达了病人对病情的焦虑，恐惧。在以肺部疾病为主题的咨询中，病人对症状的叙述偏向于使用消极、夸张的词汇；对情绪的描述也偏向于焦虑、恐惧和痛苦。

6.5 疫情期间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。

在国内疫情爆发的前期，国内新闻舆情波动增长；在国内疫情爆发的中后期，疫情新闻事件呈现高度饱和的状态；冷却时期，武汉疫情事件的话题度降低，国内新闻舆情也呈现对全球疫情爆发时期的高度相关关系。海外新闻，以 CBC 为例，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。加拿大疫情和海外新闻报道数量的相关系数为 0.5298489，检验 p 值为 0.0001281，通过显著性检验。这证明新闻报道与疫情有着较强的相关性。在疫情期间的网络舆情数量分布上，推特 ID 发文数量顶峰的三个日期也分别是疫情病例快速上升的几个节点。在网络社交论坛讨论话题上，2 月下旬，3 月中上旬的拐点，正好对应了国内外疫情爆发的分期节点。

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。

国内新闻舆情的主题分类有：肺炎患者发现和确诊统计、全国疫情发展和抗疫行动开展、疾控政策与政府防疫动向和国内抗疫成效与境外输入。国外新闻舆情的主题分类，第一类大致为教育和就业在疫情中受到的影响；第二类为政府决策和防治措施；第三类为各地区的社会民生；第四类为经济状况和资本变动；第五类为疫情医疗和社会联动情况。对社交论坛舆情主题分析显示，被分成五类时，第一类和第二类文本词汇主要聚焦于疫情本身，包括全球的疫情情况、对人们身体健康的影响程度、感染者的死亡情况以及学校学生的情况等等；第三类和第四类反映的舆论平台本身的信息；第五类是聚焦于疫情舆论的属性与特点，是关于哪一方面的舆论，例如经济、政治、体育等等。谣言可以大致分为三类，第一类谣言/传言文

本大多着眼于病毒本身的特性；第二类谣言/传言文本大多聚焦于疫情防控手段；第三类谣言/传言文本描述总体的疫情情况。

3. 舆情总体传递的情感极性是正向的。

2020 年 1 月 25 日至 2020 年 3 月 14 日，处于疫情爆发的早中期阶段，新闻报道说明了国家和政府关注本次疫情，并采取相关措施，目的在于防控疫情，救治病患和阻击病毒。态度上是积极乐观，充满信心的。在网络社交论坛舆论中，大众对疫情大部分还都是保持一个乐观的态度。

4. 舆情误区和医学咨询行为的情感动机往往是负面的。

由于网络信息传播具有迅速的特点，传言中谣言（或是不确定真假的信息）含量高，且谣言产生速度快，传播范围广，涉及主题多，易制造和传播恐慌。从医学对话数据集的病情摘要和初步印象内容来看，患者进行的咨询主要有呼吸器官感染和炎症，感冒，肺部疾病咨询以及由病情引发的心理状态异常。病人普遍具有敏感多疑情绪是疫情期间肺部疾病咨询的显著特征。病人对话中表达了病人对病情的焦虑，恐惧。在以肺部疾病为主题的咨询中，病人对症状的叙述偏向于使用消极、夸张的词汇；对情绪的描述也偏向于焦虑、恐惧和痛苦。

5. 政府措施和医学研究对舆论导向有较高的转化作用。

在医学咨询过程中，医生对话中词频较高的正面词汇偏向于安慰，舒缓病人的情绪和对病人病情的积极反馈。在谣言披露中，在来源方面，来自研究员和科技部/科技司的言论起了重要的舆论导向作用；在内容方面，针对病毒的分子式的解释和科普对辟谣产生了重要作用。在疫情舆情的时间分布趋势走向中，政府的公共防控措施强有力的占据和引导了舆论内容，积极影响了舆论或积极，或严肃的情感态度。

参考文献

[1]方兰婷. 基于机器学习的自然语言处理和传输技术的研究[D].东南大学,2018.
[2]杨宝强.基于语义分析的用户兴趣演化模型[J].江苏商论,2020(04):22-24.
[3]张天翊. 基于主题模型的科技新闻分析系统的设计与实现[D].北京邮电大学,2019.
[4]何博. 基于自定义词典的网络文本情感分析方法[D].电子科技大学,2019.
[5]胡晓辉,朱志祥.基于深度学习的中文分词方法研究[J].计算机与数字工程,2020,48(03):627-632.
[6]张旭,孙玉伟,成颖.不同特征对文本聚类效果的比较研究——以新闻文本为例[J].情报理论与实践,2020,43(01):169-176.
[7]李天赐. 基于情感的中文新闻分类与推荐研究[D].合肥工业大学,2019.
[8]陶洁. 基于新闻文本的关键词提取[D].华中师范大学,2019.
[9]李菁雯. 基于深度学习的新闻文本分类系统研究与实现[D].北京邮电大学,2019.
[10]阮光册,夏磊.基于共现分析的文本主题词聚类研究[J].图书馆杂志,2018,37(11):99-104+119.
[11]张航.基于语料库的财经新闻英汉文本特征分析[J].安阳师范学院学报,2018(04):98-103.

附录

详细附录文件请见提交数据包中的 dataset 文件夹。

1. 丁香园-海外地区（以洲为单位）确诊人数表
2. CBC 新闻数据集-国外新闻报道词频
3. 推文 ID 数据-日推文数量统计
4. COVID-19 新闻语料库-社交论坛讨论主题
5. 医学对话数据-病人情感分析建模
6. 医学对话数据-描述性文本
7. 医学对话数据-需要获得帮助内容词频
8. 医学对话数据-医生对话词频分布表
9. 医学对话数据-英文病人对话
10. 医学对话数据-英文医生对话
11. 医学对话数据-中文对话